Quantitative Metrics for Comparison of Hyper-dimensional LSA Spaces for Semantic Differences

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

John Christopher Martin
August 2016

ACKNOWLEDGEMENTS

I thank my research committee members, Dr. Michael Berry, Dr. Jack Dongarra, Dr. Qing Cao, and Dr. Louis Gross for supporting my doctoral candidacy and serving on my committee. Particular thanks are in order for Dr. Michael Berry for serving as my advisor and for his guidance during the dissertation process. I additionally thank Dr. Louis Gross and the National Institute for Mathematics and Biological Synthesis for supporting my graduate studies through the assistantship I have held there since returning to work on this degree.

I especially thank my wife and colleague, Dian Martin, for all of her help and support during this research, and for all of her careful proofreading and suggestions. This work would not have been possible without her. To my children, Julia, John Everett, and Wyatt, I give my thanks for their patience and understanding while Daddy was so busy.

Finally, I thank the late Dr. Tom Landauer for his pioneering work in LSA and his insights into applying LSA as a theory of meaning. While it was not my privilege to work with him formally, his ideas and influence have certainly shaped the direction of my research.

Small Bear Technologies, Inc. supported this work supplying the LSA_Toolkit™ software library for use in this research and providing the all the computational resources that were used. Portions of the research referenced in this dissertation were supported by the National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health & Human Development under Award Number R41HD074300. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

ABSTRACT

Latent Semantic Analysis (LSA) is a mathematically based machine learning technology that has demonstrated success in numerous applications in text analytics and natural language processing. The construction of a large hyper-dimensional space, a LSA space, is central to the functioning of this technique, serving to define the relationships between the information items being processed. This hyper-dimensional space serves as a semantic mapping system that represents learned meaning derived from the input content. The meaning represented in an LSA space, and therefore the mappings that are generated and the quality of the results obtained from using the space, is completely dependent on the content used to construct the space. It can be easily observed that modifying the content used to build a LSA space changes the meaning represented by the space, but in current practice the impact of these changes upon the overall body of meaning represented by the space is not understood. The research described here seeks to identify the impact of changes in the content of a LSA space on the meaning represented by that space through the development of quantitative measures. These measures will facilitate the comparison of different LSA spaces to assess their degree of semantic similarity. This insight will in turn provide reasoning leverage for answering questions about the characteristics of LSA spaces related to the overall body of meaning that they represent.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Automated systems for analyzing text and processing natural language input have become an important part of many fields in the modern information-rich culture. Several technologies exist for addressing the issues of processing the words, sentences, and paragraphs that humans use to communicate and record thoughts, ideas, and meaning. Key to the functioning of many of these machine learning systems is the ability to represent and relate informational items based on their semantics or on the understood meaning of those items by the people that use them. One such technology that has proven to be very effective in the area of the acquisition and representation of meaning is Latent Semantic Analysis (LSA) which is the focus of this research.

LSA is a machine learning system for deriving and representing the semantic relationships between items in a body of content (Landauer, 2002; Landauer, 2007). It mimics the representation of meaning that is formed by a human reader who learns language by exposure to content over time, forming a system of word associations and meaning (Landauer, 1998; Landauer et al., 1998b). This representation of meaning is embodied by LSA as a hyper-dimensional vector space known as a LSA space (Landauer et al., 1998a). While there are other vector space models or vector encoding systems used with text that have been developed for information retrieval, LSA provides a rich underlying theory of meaning rooted in cognitive science that is absent from these other statistical methods. LSA has been demonstrated to serve as an analog for human cognition in its representation of semantic properties for natural language (Landauer and Dumais, 1997; Landauer, 1998; Landauer et al., 1998b; Landauer, 2002; Landauer, 2007; Martin et al., 2016).

Just as human readers and learners form different interpretations of meaning based on the content to which they have been exposed, the system of meaning represented in the LSA space is completely dependent on the content used to construct the space (Landauer and Dumais, 1997; Landauer, 2002). It is apparent from the examination of individual document comparisons, where a document is some unit of content that expresses a collective thought such as a sentence or paragraph, that modifying the content used to build a space changes the meaning represented by the space (see results in Section 5.2). In current practice, however, the impact of these changes to the overall body of meaning represented by the LSA space is not understood. Given a different set of input content used to form a space, it can be shown that the vector projection of a particular document into one space will differ from the vector projection of the same document into another space. The projection of two such documents can have similarity measures between them that are not the same when calculated in different LSA spaces. The relationships between items represented within these two spaces therefore change, and where one LSA space may place a passage that discusses “going to the bank” with other passages about a financial institution, another space could represent it as being closer in meaning to passages that concern boating on a river.

The addition (or removal) of particular documents may produce a significant change in the meaning represented by a space, while others may have little or no effect. Similarly, the differences in the semantic representation contained within two spaces built from partially overlapping or non-overlapping content are not easily identified. Whole-space measures for examining the meaning represented in an LSA space or the difference in the meaning represented between two spaces do not currently exist.

The research described here seeks to identify the impact of changes in the content of a LSA space on the meaning representing by that space. This is done in order to facilitate the comparison of different LSA spaces for some degree of similarity and to provide reasoning leverage for answering questions about the characteristics of a LSA space related to the overall body of meaning that it represents. The techniques presented in this research might be applicable to other vector space methods used for information analysis, but are considered here only in the scope of the LSA model.

1.1 Problem Description

LSA is currently used in several high profile/high value applications ranging from essay evaluation in standardized testing (Foltz et al., 1999; Hearst, 2000; Foltz et al., 2013) to job placement and training (Laham et al., 2000; Foltz et al., 2012), as well as a model for understanding human cognition and learning (Landauer, 2002; Landauer et al., 2007; Biemiller et al., 2014). Each of these applications depends on the semantic representation encapsulated in an LSA space for their operation. The original motivation for this research was precipitated by observations made during a larger project that made use of LSA for instructional user feedback generation in a scenario-based automated driver training system (Martin et al., 2016). The system used the concept of a Base Interpretive Space (BIS) as described in Section 2.3 for mapping user responses, evaluating their semantic content, and selecting appropriate feedback information. In the process of evaluating content for inclusion in the background space, several observations were made that raised questions about the differences between LSA spaces and how they represented meaning.

1.1.1 The Driver Training Project

Knowing the specific subject domain of the scenarios in question (driver training), construction of the candidate BIS was begun by obtaining 6,629 domain specific documents from various public domain sources chosen based on subject matter (driving knowledge) to provide the basis of the domain specific content for this particular application. Since this system was intended to evaluate responses specifically from a younger age demographic (new drivers ages 15 - 18), a second set of 5,056 documents containing samples of language obtained from high school newspapers was selected to supply content representing this particular age demographic. Two different collections were examined for augmenting this initial base content set of 11,685 documents labeled S1. The first was a general linguistic space containing just over 44 thousand paragraph sized documents obtained from various public domain literature sources added to set S1 to form set S2. The second was a randomly selected set of 100,000 documents from the RTRC collection (Lewis et al., 2004) added to set S1 to form set S3. These sets were evaluated as possible candidates for use as a BIS in the project.

For evaluation purposes it was desired that the space to be broadly distributed, covering a wide range of linguistic meaning. Natural language responses to any given evaluation prompt could vary widely in the terms and linguistic constructs used to express the same meaning. It was necessary for the background space to accurately map all these varied responses. The initial content set, S1, and the two augmented content sets, S2 and S3, were analyzed by performing a full vector-to-vector analysis, a centroid analysis, and then specific subset centroid analysis based on the domain specific content groups (all of which are described in Section 3.1.2) to determine if one of the two candidate sets would be suitable. The results of these analyses are shown in Figure 1.1. These results revealed a degree of difference between the spaces that was not anticipated. The S2 and S3 spaces differed somewhat in the volume of content from which they were constructed, but the initial view of the spaces based on the vector-to-vector analysis and both the document and term centroid analyses showed little difference in the distribution of the content within the spaces.

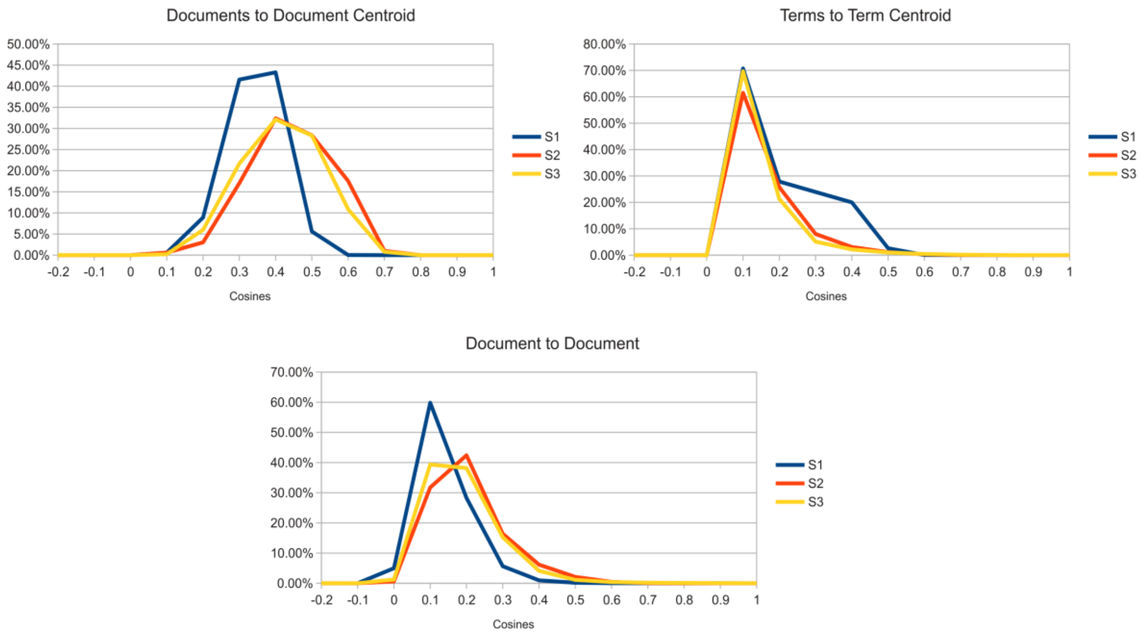

Figure 1.1 - Space analysis results comparing the candidate LSA spaces, S1, S2, and S3, for use as a BIS in the driver training system. The top left graph shows the comparative distribution between three candidate spaces based on the cosines of document items to their respective centroid vector in that space. The top right graph shows the similar comparison of the term items to their centroids, and the bottom graph shows the distribution of all pairwise document-to-document cosines for the three spaces.

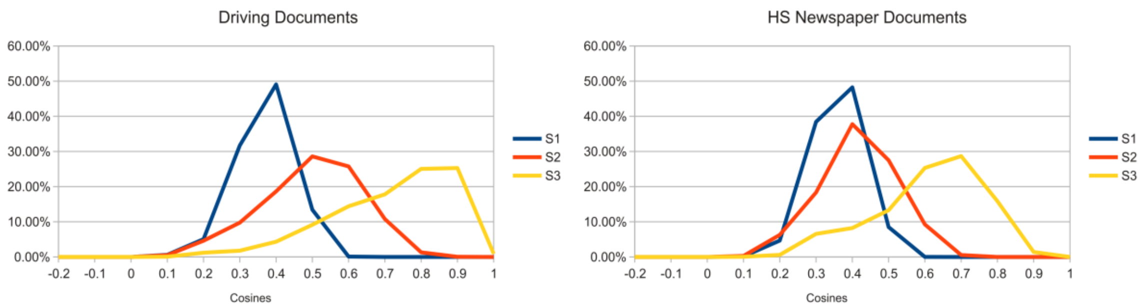

Figure 1.2 – The targeted subset centroid analysis for two identified content sets in the context of the candidate base interpretive spaces. The left graph shows the distribution of the driving document subset cosines to their respective centroid in each of the three spaces. The right graph shows the distribution of the high-school newspaper documents to their centroid in each space.

When specific content items within the spaces were targeted for analysis however, significant differences were observed in the way in which the content was related within S2 and S3. It was already known that the content of the driving knowledge subset was all related due to the selection process used to assemble it. In the whole-space analyses based on the document and term centroids or the document to document cosine distribution, there seemed to be little difference between the effect of augmenting the S1 content with either the additional items used in S2 or S3. Evaluating the distribution of targeted subsets of content items however, clearly showed the difference between the three spaces in their ability to group the driving knowledge subset (see Figure 1.2). Similarly, with the subset centroid analysis of the high school newspaper articles, the same general improvement in the cohesion of the subset content was observed in the S3 space over the S1 and S2 spaces.

1.1.2 Implications

Eventually the S3 space was selected for use in the Driver Training project based on the use of human raters and test scoring to evaluate the quality of the BIS being used in the application, and ultimately the automated tutorial system yielded favorable results (Martin et al., 2016). Questions still remained, however, about the best method to use in determining the quality of a LSA space and how modifying the content from which the space was built was affecting the meaning that was being represented by that LSA space.

The differences observed between the candidate spaces in this project illustrated the shortcomings of simply processing content with similar surface characteristics such as terms used or number of documents. The easily accessible surface features did not serve to indicate the degree of difference that was evident in the mapping of items using the different spaces. Adjustments to the input content based on the desire to include additional vocabulary or cover different topic material had an unknown effect on the overall collection of meaning embodied in the LSA space.

1.2 Broader Impacts

LSA is used frequently in evaluation applications where it is necessary to construct an LSA space from background information to serve as a framework for mapping the meaning of the information to be processed. These applications include automatically scoring essays (Landauer et al., 2003; Foltz et al., 2013), evaluating the content of medical patient notes (Swygert et al., 2003), analyzing individual and team communications (Foltz et al., 2003; Foltz and Martin, 2008), automated performance evaluations for military teams (LaVoie et al., 2008), and other education and training applications (Kintsch et al., 2007; Streeter et al. 2007; Landauer et al., 2009; Martin et al., 2016). In each of these cases, the techniques employed in the application rely on the mappings for projected items to correctly identify items that are semantically alike and simultaneously to adequately differentiate between items that are semantically similar. The construction of a LSA space for use as a BIS in these applications has largely relied on empirical testing and human evaluation of results to determine the acceptability of a space for use.

Because the use of a different LSA space as a BIS produces a different set of semantic relationships between the items mapped by those spaces, essentially a different interpretation of the meaning of those items, it is important to understand the nature of the differences between the meaning represented by spaces being considered for use. Changing the space being used or altering the content of an existing BIS to include new vocabulary or content domains alters the results produced by evaluating items using that space, possibly having an adverse effect on the overall application. In other cases it may be desirable to produce a large change or alter the interpretation of meaning that is housed in a space by changing the input content. In any case, while it is observable that changing the input content has an impact on the meaning represented by a space, the extent of such impact is currently unknown since there is no means for assessing or describing such differences based on the space in its entirety. The ability to assess and quantify the difference between the semantic mapping systems encapsulated in different LSA spaces will help answer questions about the usability of various spaces for representing knowledge in different fields or application domains.

1.3 Current Literature and Practice

There is little current literature discussing the evaluation and comparison of LSA spaces as entities in and of themselves, though the importance of such future research was noted as early as 1998 (Rehder et al., 1998). What little coverage there is primarily revolves around examination of the performance of specific queries and their correlation to human scoring judgments (Kurby et al., 2003). There is no question that the meaning represented in a space changes significantly with its content, and this fact is actually exploited by research in vocabulary acquisition that makes use of LSA to model different levels of human word understanding with carefully constructed spaces (Biemiller et al., 2014). Initial work defining a set of qualities for describing a LSA space has been done

(Martin et al., 2016), but actual direct comparison of different spaces and quantification of their differences remains unexplored.

Current practice in assessing the quality of a space for a given use has mainly been limited to precision and recall tests. To perform these tests specifically selected content is included in the input content used to construct a space and then some number of queries is performed on the space for the subjects of interest from that content set. A measurement is made of the ability to retrieve the expected items from the space using those queries. This process is usually combined with the use of human readers to provide subjective judgments on the relevancy of the items returned (Kurby et al., 2003). Precision and recall tests of this nature are commonly used to judge the information retrieval (IR) quality of a space (Berry and Browne, 1999; Dumais, 2007). While they do meet the need of assessing specific IR quality goals, they essentially provide only anecdotal evidence on the meanings represented within a given space as they provide only small slices of insight into the characteristics of a space. They also require a significant amount of a priori knowledge about the document collection and the queries that will be issued against the space. It is true that poor precision and recall results might give an indication that the LSA space is of low quality, but they provide no clues as to what problems with a space might be responsible for those results. Acceptable precision and recall results could just as easily lead to a false sense of confidence in the quality of the space as they do not capture a big picture view of the relationships within the space. They provide no means of evaluating the whole difference between two spaces and the semantic relationships that they represent. While they serve well enough to describe the performance of basic IR applications, they do not serve well for reasoning about the meaning represented in different LSA spaces or for addressing whole-space comparisons between spaces.

1.4 Innovation

The research described here seeks to identify the impact of changes in the content of a LSA space on the meaning represented by that space through the development of new whole-space quantitative measures. These measures must extend beyond the common surface view of simply counting the number of document and term instances recorded in a space, the incomplete and often subjective human view of precision and recall measures, or even the purely mathematical view of term and document vectors, dimensions, and singular values, to somehow reflect the semantic representations embodied in an entire space and reflect the differences in those mappings between two spaces of interest. To this end, several candidate measures are defined and examined in search of a viable metric or set of metrics for quantifying the degree of difference between two spaces, and different influences on the body of meaning represented in a space are explored. A novel approach toward effectively describing the degree of semantic difference between two LSA spaces using a repeatable, mathematically-based, and computationally affordable method is developed and presented as the result of this research.

1.5 Overview

To facilitate understanding and appreciation of the issues involved in this research, a brief description of LSA and its mathematical foundation is given in Chapter 2, along with a discussion of its development as a theory of meaning. The use of LSA as a semantic mapping system for the purpose of analyzing natural language text is also presented. This is followed in Chapter 3 by a survey of the various measures that were explored in this research as possible candidates for quantifying the semantic difference between two spaces. These measures are described formally and an analysis of their computational complexity is given. Chapter 4 provides details of the experiments that were conducted to test these measures and to answer questions about the factors that influence the different meanings represented by an LSA space. The methods for constructing the spaces used in the experiments as well as the data items that were collected are detailed. A general discussion of the experimental results as well as specific observations from each of the experiments and the different measures that were collected follows in Chapter 5. Finally, Chapter 6 concludes with a summary of the results and identification of future research areas to be pursued.

CHAPTER 2

BACKGROUND

All of the research described in this work is based on the use of Latent Semantic Analysis (LSA) as an operational system for representing meaning learned from text. LSA is a proven technology that has been developed and refined over the past three decades. This chapter presents a brief overview of the established theory behind LSA and its practical use.

2.1 Latent Semantic Analysis

Latent Semantic Analysis has evolved into a theory of learning, a computational model of human thought, and a powerful text analytics tool. It takes its name from the fact that it presumes the existence of an underlying or “latent” structure relating the meanings (semantic value) of words within a body of text (Dumais et al., 1988).

LSA was developed initially in the late 1980s as an approach to dealing with the synonymy problem in retrieval systems. The pioneer research presenting LSA as an unsupervised learning system was presented in the 1997 paper by Landauer and Dumais describing LSA as “a solution to Plato's problem.” Plato's problem is basically the question of how humans develop the knowledge they have based on the relatively limited amount of information they have received. In their paper, Landauer and Dumais put forward LSA as a model of human learning as an answer to this question, marking a change from viewing LSA as retrieval method to that of a powerful learning system.

2.1.1 Mathematical Background

LSA is a mathematical model of the meaning represented in human language and the acquisition of knowledge. The assumption is that people learn by associating words and contexts that are experienced together over time. This accretion of past experience provides a dynamic framework for predicting new experiences. Human cognition takes in all these experiences, word and context co-occurrences, fits them into a semantic map that represents how each word and context is related to each other. The mathematical model of LSA does much the same thing. A semantic space is constructed by first digesting a large body of textual information which is processed to form a common mapping system. LSA represents words and contexts as vectors within the semantic space mapping the meaning of each item, modeling how people acquire meaning from experience. The meanings are learned from the text itself in a manner similar to the way a human learns language. While the input to LSA is limited to electronic text where a human has multiple learning sources and additional perceptual information, and though it lacks certain capabilities present in our human cognitive model such as word order, syntax, morphology, etc., the representation of meaning produced by the LSA process is highly similar to that of humans. It is based on concept and semantics not keywords or syntactic constructs (Foltz et al., 1998; Landauer and Dumais, 1997). LSA has been shown to produce a linguistic and cognitively effective representation of word meanings even without specific, or any, prior knowledge (Landauer, 2002).

Construction and use of a LSA semantic space can be described as three phases of work: parsing, singular value decomposition, and query and analysis post-processing (Martin et al., 2016).

Parsing and Weighting

The formation of a LSA space begins with the conversion of the input text into a numeric representation that can be processed by a computer. Initially a simple tabular matrix is constructed from a collection of input text to start the parsing process. This tabular matrix is typically formed with the rows representing unique terms in the collection and the columns representing the documents in which the terms are used. Terms are usually taken to be the single word items in the text, but may also be formed from multiple juxtaposed words, sometimes called *n*-grams. Documents, which are more generally referred to as contexts or passages, can be designated as any size unit of textual content. Possible document units can be short phrases, sentences, paragraphs, multiple paragraphs, or articles, but are usually selected to be single paragraphs since they generally represent an expression of meaning based on a unified topic (Landauer, 2007).

Upon completion of the initial parsing work every cell in the term-by-document matrix will have been assigned a value indicating the number of times each individual term appears in each individual document, with the bulk of these entries being zero. The result is a large sparse matrix, often referred to as the term-by-document matrix. A weighting function is then applied to this term-by-document matrix which serves to normalize the occurrence of terms within documents and across the collection of documents. In the weighting process the value of each element is adjusted by both a global weight, which reflects the importance of the term across the entire collection, and local weight, the importance of a term within each single document. The local weighting function is typically chosen to reduce the influence of terms that occur frequently within a single document while a global weighting function is selected to reduce the influence of terms that occur more frequently across the entire collection (Martin et al., 2016). A combination of local log and global entropy is the typical weighting scheme used with LSA as this scheme was demonstrated to produce the best performance out of several different weighting schemes in precision recall tests (Dumais, 1991).

Singular Value Decomposition

Once parsing has been completed and the weighting functions have been applied, the data represented in the term-by-document matrix still describes the input content simply as a co-occurrence of terms and documents. It can be observed that this matrix reflects a coefficient matrix for a system of simultaneous linear equations with each equation representing a document in the collection. The basis of LSA is the computation of a solution to this system in order to infer the meaning for each term as a vector which will reflect the "meaning" or mapping for that term based on the documents in which it does and does not appear. Singular Value Decomposition (SVD) is used to process this system of equations, yielding as the factored output a set of vectors in a high dimensional "semantic space" and the corresponding dimensional singular values. Within this semantic space a document vector is equivalent to the vector sum of all the term vectors corresponding to the terms it contains. Similarly, a term vector represents a term in all the different senses in which it may be used within the documents of the collection (Landauer, 2002).

The SVD produces a factorization of the original term-by-document matrix A in three parts: $A = U\Sigma V^T$. Where the rows of matrix U are a set of vectors that correspond to the terms of the collection, and the rows of matrix V are a set of vectors that correspond to the documents of the collection. The nonzero diagonal elements of Σ, the singular values, essentially serve as a set of scaling factors for the dimensions. LSA uses a truncated SVD based on finding the k extremal singular values for the term-by-document matrix A. With a full SVD, it is possible to reconstruct the original matrix A from the three matrices U, Σ, and V, however, the truncated SVD that is computed for LSA generates the best rank-k approximation of A (Golub and Van Loan, 1996). The truncated SVD, defined as $A_k = U_k \Sigma_k V_k^T$, yields a reduced set of k dimensions based on these singular values for A. This process results in the most significant dimensions of the space being used to define the truncated vector space used for the LSA representation (see Figure 2.1). This dimensional reduction has the effect of removing noise from the original representation of A, essentially countering the

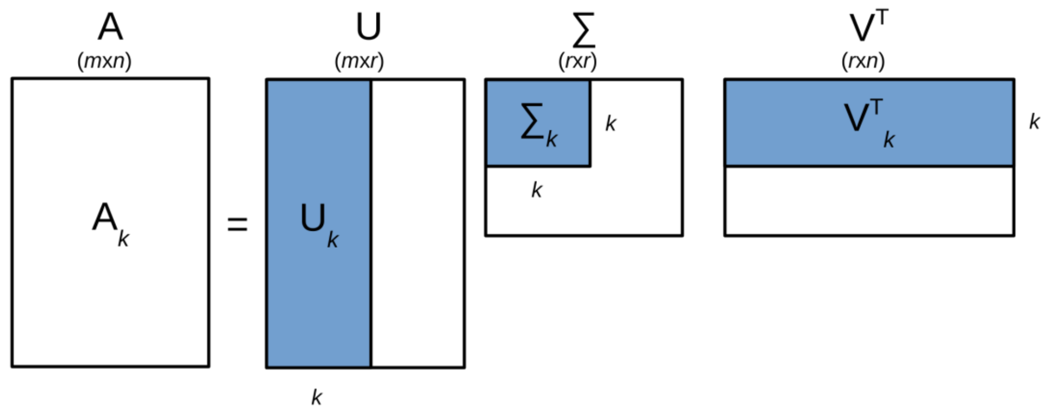

Figure 2.1 - A pictorial representation of the truncated singular value decomposition (Martin and Berry, 2010)

dissimilarity of related documents that use synonymous terms while separating those that contain polysemic terms (Martin et al., 2016).

The resulting vector space produced by the truncated SVD can be considered as a multi-dimensional hyperspace where each item is represented by a vector projecting into this space. This concept can be roughly pictured with a simple 3-dimensional representation, $k = 3$, where the vectors point out into the 3-D space, (see Figure 2.2). This illustration in this figure is extremely simplified to facilitate visualization. In practice, k is typically selected to be anywhere from 300 to 500 dimensions. Empirical testing has shown the selection of the number of dimensions in this range to be most effective in tests for recognition of synonyms (Landauer and Dumais, 1997; Landauer et al., 1998a). Within this hyper-spatial representation, information items are left clustered together based on the latent semantic relationships between them. This representation forms a “semantic space” that provides a mapping of relative meaning for terms and documents as learned from the body of input content.

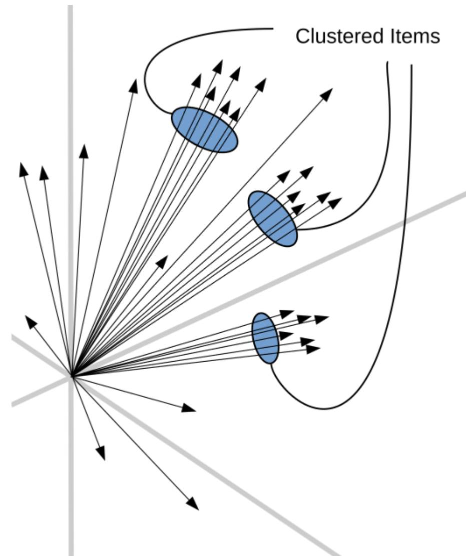

Figure 2.2 - A simplified 3-D visualization of a LSA semantic space depicting the conceptual clustering of similar data items

Query and Analysis Processing

There are two basic forms of analysis that may be performed after the truncated SVD has been computed for a collection of input text. First, any of the items within the space can be compared for semantic similarity using quantitative measures. Second, new document items can be constructed and projected into the semantic space being mapped according to their semantic content as defined by the space. Similarity comparisons may then be performed between these new items and existing items in the space, or between different new items themselves.

Individual items within the space can be compared based on their semantic mappings. Each item, whether a term or document, is represented by a k-dimensional vector describing its mapping. Vectors for any two items may be compared quantitatively by computing a distance measure or a similarity measure indicating their proximity or separation. The similarity measure that is typically used with LSA is the vector cosine similarity which has been shown to be a reliable measure of semantic relatedness within the LSA space (Rehder et al., 1998). Cosine similarity essentially describes the angle between the two vectors being compared, u and v. It is defined as the dot product of vectors u and v divided by the product of their vector lengths: $\frac{u \cdot v}{\|u\|\|v\|}$, (Trefethen and Bau, 1997). Other possible similarity measures include the Euclidean distance measurement or the dot product of two vectors being compared (Martin et al., 2016).

2.2 The LSA Model of Learning

LSA is based on what is now referred to as the Compositionality Constraint: the meaning of a document is a sum of the meaning of its words, and the meaning of a word is defined by all the contexts in which it appears (and does not appear) (Landauer, 2007). A document derives its meaning from the terms it contains. Each term contributes something to the collective meaning contained in the document. At the same time, two documents can be similar in meaning and not contain the same terms. Likewise, terms that appear together in one document do not necessarily have similar meaning in different document contexts. By using the SVD to form a semantic space from a system of simultaneous linear equations where each term, and each document, is mapped by a vector, LSA exploits the concept of mutual constraints on the occurrence of many words across many contexts. The resulting representation allows for similarities to be observed between the words and contexts based on their mapping within this semantic space (Martin et al., 2016). LSA essentially models word association which is an extremely important component of human cognition (Landauer, 2002). A LSA system functions both as way to automatically learn the meaning of words and contexts, and as a computational model for the very process of human learning (Landauer et al., 1998b).

While LSA performs well in many cases, it should be noted that it does not provide a complete model of human knowledge. A LSA system is limited in the information that it receives as it builds its model of meaning from only the body of text that is supplied as input. The functional performance of a LSA system is restricted in the face of inadequate training data that fails to represent the language exposure of a typical person (Landauer et al., 1998b). Human learners also have access to a full range of perceptual information, emotional input, instinct, and other information sources that are not generally available to a LSA system as input. LSA can, however, access knowledge about those processes through the written word and produce a close enough approximation to human-like knowledge to represent those concepts (Landauer et al., 1998a).

There are three aspects of the LSA model that are important to its operation. These are the dimensionality of the semantic space, the orthogonal mapping axes that are produced as the framework for this space, and the individual term meanings that are derived from the input content and mapped in this space.

2.2.1 Dimensionality

As noted in Section 2.1.1, LSA does not use the complete SVD, but rather a truncated SVD that will yield the best rank-k approximation of the term-by-document matrix A. Selection of k, the number of dimensions, is an important consideration in the construction of a LSA semantic space. The inclusion of too small a number of dimensions has the effect of under-differentiating concepts in the data, but using too large a number of dimensions will result in over-differentiation of items in the data, either condition making it difficult to find the underlying important semantic relationships in the data (Deerwester et al., 1990; Landauer, 2007; Martin and Berry, 2015). Identification of the optimal dimensionality currently remains an open research topic.

The dimensional reduction that is obtained using the truncated SVD is an important facet of the LSA model. It has been theorized that this reduction produces the same approximate relations as occur in human cognition, mimicking the structure of the brain along with the statistical structure of experience (Landauer, 1998). Ideally, finding the optimal number of LSA dimensions would yield a representation of the input data analogous to the same dimensionality as the source that generates the semantic space of a human learner (Foltz et al., 1998; Landauer, 2007).

2.2.2 Orthogonal Mapping Axes

Another important product of the LSA process is the set of orthogonal axes that are generated during the SVD computation as the mapping dimensions of the semantic space. These dimensional axes are abstract features that do not correspond to any namable concept or meaning and are not interpreted as such (Landauer, 2002). Instead, they form the framework for the vector space where concepts are mapped. Terms and documents subsequently derive their meaning from their mapping on these axes, but they do not serve to define the axes (Landauer, 2007; Martin et al., 2016). Being that these axes are orthogonal, the mappings defined on them are all universally comparable via straightforward computations.

2.2.3 Term Meanings

The ultimate result of the LSA process is the set of derived definitions of individual term meanings as vector mappings within the semantic space. Since LSA is based on the compositionality constraint, the availability of individual term meanings allows the construction of new documents and their associated placement within the framework of meaning represented in the specific LSA space being used. This is the basic mechanism for individual query processing. A pseudo-document vector is formed using the term vectors corresponding to the terms in the query as described in Section 2.3.1, and a comparison is then made between this pseudo-document vector and other document vectors in the space by computing a similarity measurements between them (Martin and Berry, 2015).

2.3 The Background Space

The mapping system defined by a LSA space can be used in text analysis applications leveraging its learned term meanings and framework of orthogonal axes as a semantic background for analyzing new content. When used in this fashion the semantic background space is referred to as a *Base Interpretive Space* (BIS) (Martin et al., 2016). A BIS is simply a LSA space that is purpose built from a set of content in order to learn the term meanings that will be used for mapping new content items. Once available, it is possible to use it for analysis of new content by constructing mapping vectors for the new items from the term vectors contained in the BIS. This may be used to analyze content that becomes available after the initial LSA space, the BIS, was built. It is also useful for other applications where the dataset of interest is small or narrowly focused, not providing enough material to establish a basis of meaning by constructing an LSA space on that content alone. In these situations the BIS provides a contextual background that augments the meaning represented in the dataset. The BIS forms a consistent representation of meaning that can be used for evaluation and comparison of the semantic relationship between natural language text inputs.

2.3.1 Projection of New Content

Use of the BIS involves the projection of new content items into the semantic space (see Figure 2.3). This process makes use of the term definitions provided by the BIS to calculate the mappings for new document items within the semantic context of the space. This can be used to probe the space as in the case of an information retrieval query, or for simply establishing semantic relationships

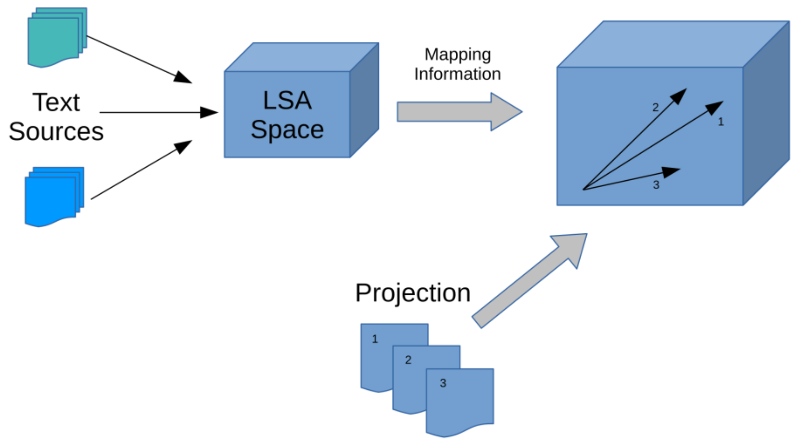

Figure 2.3 – A conceptual illustration of the projection of content into an existing LSA space used as a BIS (Martin et al., 2016)

between items of interest within the context of meaning represented by the BIS. In mathematical terms, a new document projection, a pseudo-document, is computed as the weighted sum of the term vectors corresponding to the terms in the item being projected scaled by the inverse of the singular values, $z = \frac{q^T U_k}{\Sigma_k}$ (Martin and Berry, 2007). The projection of new content is performed in three steps consisting of parsing the text of the projection item, applying a weighting function to the parsed item, and then finally composition of the weighted term vectors for the projection item, each described below.

Parsing

The projection of new content begins by parsing the text of the item to be evaluated. This projection parsing follows the same procedure as the initial parsing of the input content used to form the BIS, enforcing any policy for the handling punctuation, casing of characters, numeric values, etc., and the input text is reduced to a term frequency vector (TFV) where the number of occurrences of each mappable term is counted. Terms that were not present in the input content used to form the BIS have no corresponding term vectors and therefore cannot be mapped. The number of mappable terms, the term “hit-rate”, is monitored to determine if the BIS is sufficient to map the projected content or if there are shortcomings in the vocabulary of the BIS that need to be addressed (Martin et al., 2016).

Weighting

The next step in the projection of new content is the application of a weighting function to the TFV for the projection item. The weighting function applied to the projection item must correspond to the one used in the construction of the BIS. If the typical log-entropy weighting method is used, the weighted value for each term in the TFV for the evaluation item is computed as: $\ln\left(tf_{ij} + 1\right)$. This value is then multiplied by the global entropy value for that term in the BIS which was obtained when the original term-by-document matrix used to form the BIS was weighted.

Global entropy is defined as

$$1 + \sum \frac{p_{ij} \log_2\left(p_{ij}\right)}{\log_2(n)}, \qquad where\ p_{ij} = \frac{tf_{ij}}{gf_i}.$$

In these equations tf_{ij} is the number of times that term i appears in document j and gf_i is the total number of times that term i appears in the entire collection used to build the space into which the projection is being made(Martin and Berry, 2007; Martin et al., 2016).

Composition

The final step in the formation of a projection item vector is the composition of the projection vector, z. z is computed by multiplying the weighted TFV, q, for the projection item with the term vectors U_k from the BIS. The vector product of this multiplication is then divided by the singular values Σ_k from the BIS to obtain the new projection item vector.

The projection vector computation is defined as

$$z = \frac{q^T U_k}{\Sigma_k}.$$

The projection vector z of an item can be compared to other items projected using the same BIS or to term or document items from the BIS using the similarity measures described previously in Section 2.1.1.

2.3.2 Quality of the background space

The meaning represented in a BIS, and therefore the mappings that are generated and the quality of the results returned, is completely dependent on the input content used for its construction. Spaces may be built from corpora thought to represent general knowledge, or supplemented with documents on specific topics thought to be relevant to a particular domain. LSA forms the mappings of linguistic constructs (terms and documents) based on the input information it is given. If this input is insufficient, or narrowly focused, the meanings that are represented in the BIS may be badly distorted.

There are several basic but important considerations when selecting content that will be used in the construction of a BIS that have the potential to affect its overall quality. These include the size of the input corpus, the overall term coverage provided by the input corpus, and the presence of relevant and distributed content within the corpus (Martin et al., 2016).

Corpus Size

LSA learns meaning by representing association between the terms and documents provided to it as input. Early experimental work in the development of LSA was frequently performed using relatively small input collections often consisting of 2,000 documents or less. The results from these initial tests were varied, but generally yielded promising results. Limitations imposed by the available computational resources made the processing of larger corpora impossible for these early experiments, but with the advent of more powerful processors and larger accessible memory and storage resources the ability to process larger input sets became feasible and improvements in the performance of LSA were observed. It is now possible to process input collections consisting of hundreds of thousands of documents or larger.

Large content sets are necessary for LSA to adequately learn meaning, just as large volumes of linguistic stimuli are necessary for a human to learn language. “Greater amounts of text help define the space by providing more contexts in which words can co-occur with other words” (Foltz, 1996). It has been noted that a minimum of 100,000 paragraph-sized passages is needed to represent the language experience of an elementary student (Landauer, 2007). LSA learns the meaning of a term not just by the number of times it has been seen in a context, but also by the generally much larger number of times it has been observed absent from various other contexts. LSA must be trained on a sufficient number of documents before the meaning of any word can be distinguishable from other words and using a small input collection does not provide enough context for meaning to be learned. LSA cannot learn word meanings from contexts or for terms to which it has not been exposed. Small or domain specific corpora are generally not adequate to train a LSA learning system (Landauer, 2002; Landauer, 2007; Martin et al., 2016).

Term Coverage

Another important consideration in the quality of a LSA space is that of term coverage. Only the terms already defined in the vocabulary of a space can be mapped when a new projection item is processed. Terms that do not appear in the vocabulary of the space cannot be included in the computed projection vector and are in effect discarded when they are encountered in a projection item or query. For a space to serve as a usable BIS it is necessary that the vocabulary include a high percentage of the terms that are present in the items being processed, otherwise the interpretation of those items becomes questionable. Obtaining sufficient term coverage for a BIS is generally not a problem, as most of the expected terms for a set of projection items will be typical of the overall language and already in the vocabulary of a reasonably sized corpus. Monitoring the term “hit-rate” as projection items are processed provides a means for identifying individual items that may have a questionable mapping and to provide some indication that the BIS lacks the needed content to handle the domain of the items being processed (Martin et al., 2016)

In most cases an occasional missing word will have little net effect as the meaning that is required will be partially indicated by the surrounding context. LSA actually performs well in tests where a missing word must be selected to complete a text passage (Landauer et al., 1998b). It is only when larger portions of a document have no mappable terms that the overall meaning becomes distorted. This is analogous to the situation that a human reader experiences when attempting to read a document with terms that are simply unknown to him or her. Depending on the number of unknown terms and their importance the reader may misunderstand the meaning of the text or find the document completely unintelligible. In these cases a human reader would generally indicate that there was a problem in understanding the text, where an automated system will simply give a wrong answer unless the term coverage condition was specifically being monitored. Past flawed reports of poor LSA performance have been attributed to this very cause (Landauer, 2002).

Relevant and Distributed Content

In order to form a high quality representation of linguistic meaning, LSA also needs good quality input content that provides concept information that is not just limited to the domain of interest, but also includes enough general information to represent language usage as a whole. As noted previously, LSA forms its representations of meaning from the analysis of text alone. The nature of the content included in the corpus used to construct a BIS influences whether the LSA representation models an amateur, novice, or expert level of knowledge. If LSA is trained on content including highly technical texts of a particular domain, then it will form meaning representations more like an expert in the field and the semantic representation will be much more elaborate (Foltz et al., 1998). Additionally, to build this notion of meaning LSA needs to have many representative textual associations both in the present and in prior knowledge of a potential user of the system (Landauer, 2007; Landauer et al., 1998a). The inclusion of elementary texts is needed to form the foundational representation of simple linguistic meaning that more complex meaning relationships are built upon. The BIS must include not just items in the domain but background meanings representative of the wide range of expression (Martin et al., 2016).

CHAPTER 3

COMPARISON MEASURES

The principal question addressed in this research is the development of a method for determining the difference between two arbitrary LSA spaces, S_1 and S_2, in a quantifiable manner that sufficiently reflects the difference in meaning represented by the spaces. Building on the basic characteristics of a space described in Section 2.3.2, a measurement or suite of measurements is desired that can be easily computed on the spaces S_1 and S_2 themselves that will give a general indication of the degree of difference that is present.

There are several potential methods for comparing two spaces and measuring their differences. These methods fall into three major categories:

- Direct measurement and comparison of properties for the individual spaces
- Comparisons of projected content items across both spaces
- Mathematical rotation to transform the vectors from one space into the mapping of another space

Within these categories there are multiple methods that provide some comparative value, each with its own degree of computability, expense, and amount of insight provided. These measures will be evaluated against each other to determine which ones individually or in combination provide a usable indication of the net differential in meaning between S_1 and S_2.

3.1 Direct Comparison Measures

Direct comparison measures deal with observations that can be immediately made from the individual spaces. Several simple characteristics for a LSA space can be observed and may be employed for basic analysis of a space as well as comparison between spaces. These range from the raw metrics of corpus size, term coverage, and term and document overlap between spaces, to more detailed analysis of the distribution of similarity measures between individual terms, individual documents, and between those items and other points within a space. While these are perhaps inadequate on their own to describe the net difference in represented meaning between two spaces, they do serve as an initial indicator of difference in the content used to construct the spaces in question.

3.1.1 Raw Metrics

The simple raw metrics initially available are the corpus size, the count of documents contained in the collection used to build a space, and the term coverage, the count of terms contained in the collection. These counts exist as the number of document vectors and the number of term vectors contained in an LSA space. As noted previously (Section 2.3.2) the use of a small number of documents for constructing a space has been theorized to provide inadequate information for building a system of meaning that correctly models human understanding of language (Landauer, 2007). Even with large numbers of input documents (in excess of 150,000), however, significant differences in the meaning represented by spaces of similar size may still be observed.

The measures of document overlap and term overlap between two spaces can be derived by counting the number of documents and terms that are commonly represented in two spaces that are being compared. These counts are easily obtained, though in the case of documents such comparison also requires either very primitive forms of the input content (term frequency vectors), or the raw content (the text documents) itself, be available for examination. Overlap of either terms or documents can be expressed as a ratio of the number of common items in both S_1 and S_2 to the total number of unique items in S_1 and S_2.

$$Term\ Overlap\ (S_1, S_2) = \frac{|terms(S_1) \cap terms(S_2)|}{|terms(S_1) \cup terms(S_2)|}$$

While it is expected that term overlap between two spaces must be high if the two spaces possibly represent the same meaning (terms must be mapped in order to be used), it is not necessary for the same documents to be present in the two spaces. In most cases there will be few, if any, overlapping documents in the spaces that are being compared.

3.1.2 Distribution Analysis

Direct investigation of the difference between two spaces is also possible by performing a comparison of the relative distributions of items within S_1 and S_2. One such distribution is a vector-to-vector analysis where the distribution of all pairwise cosines for items in a space, either terms or documents, is computed and examined. Another distribution to be considered is the full space centroid analysis where the distribution of all component items around a general centroid of the space is computed. Similarly, a targeted subset centroid distribution analysis may be used to look at the distribution of a subset of common related items to their subset centroid.

Vector-to-Vector Analysis

For any two item vectors in an LSA space, even though they are individually projected into the hyper-dimensional space, a pairwise comparison can be made which is measured in the plane in which they both exist. Within this plane the similarity measure most frequently used is the vector cosine (Deerwester et al., 1990; Dumais, 1991). A high cosine is an indicator of similarity between the semantic mappings of two term or document vectors. A cosine of 1 occurs when two vectors duplicate each other in meaning within the context of the space. A cosine of zero indicates the absence of any semantic relationship between the two items being compared within the context of the space as the vectors are

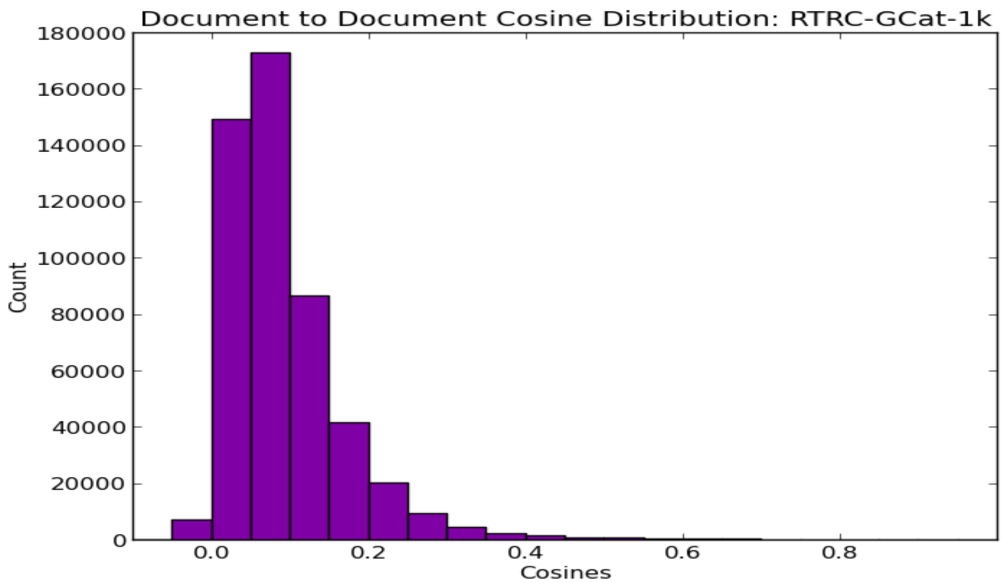

Figure 3.1 – Visualization of the vector-to-vector document cosine distribution for a single space

orthogonal in their mapping. Negative cosine values between two items may be seen occasionally. These items are perhaps related to some degree, albeit in a negative or oppositional way, where items with a zero cosine between them are not.

One way to view the entire space is to calculate for each term and document vector the cosine to all the other term or document vectors in the space. The resulting cosines allow examination of the maximum, the minimum, the average and standard deviation of the cosines between all of the term vectors or the document vectors. The number of cosine calculations required for this analysis is $\binom{n}{2} = \frac{n(n-1)}{2}$, where $n = |S|$. So, while simple in concept, this analysis amounts to a computation with a complexity of $O(n^2)$. Further analysis may be performed by dividing the resulting cosines into bins at a selected level of granularity and observing the distribution pattern of the set of cosines. This analysis may be visualized with a histogram which will depict the generalized curve of the distribution (see Figure 3.1). The compilation of this further information is a straightforward counting operation of the k cosines produced in the previous step with a complexity of $O(n)$, where $n = k$.

These statistics give a picture of the overall distribution of the term or document items within the mapping provided by the space. For example, if the minimum cosine is at the high end of the scale (nearer to 1) then it can be observed that the content of the space is narrowly focused with little difference in the similarity measure between the various components of the space. Similarly, if the maximum cosine is near the lower end of the scale (nearer to 0) then it can be observed that the content of the space is very dissimilar as mapped in the space. If the maximum and minimum cosines are near in value, but are neither particularly at the high end nor the low end of the scale, this indicates that there is little variation in the mapping relationships among the items. Essentially none of the items in the space are mapped as more similar or less similar in meaning to each other according to the cosine similarity measure. If the minimum and the maximum cosines are broadly separated then examination of the average cosine and the standard deviation gives some indication of whether the majority of the content is similar or dissimilar in meaning. Visualizing the cosine data in a histogram gives a better indication of the overall distribution of the items in a space, showing the actual curve of the distribution, and allows for comparison between the analyses of two different spaces through visual inspection. The histogram also makes it possible to more easily note outliers and holes in a space.

Full Space Centroid Analysis

While a full vector-to-vector analysis provides a detailed picture of all of the semantic relationships in a LSA space, it is computationally expensive for large spaces. A more easily accessed big picture view of the distribution may be obtained by computing the centroid vectors for the entire content of the space (one centroid for terms and one for documents) and then examining the distribution of cosines for all the document vectors or term vectors relative to their respective centroid. The centroid for a collection of vectors is obtained by averaging the set of vectors which is accomplished by simply adding the vectors and then dividing through by the number of vectors. As with the vector-to-vector distribution analysis, examining the minimum, maximum, and average similarity to the centroid along with the standard deviation gives a set of easily computed metrics for reviewing the distribution. The number of cosine calculations required for completing this analysis is simply n, where n is the number of terms or the number of documents in the space, thus the complexity for this entire analysis is simply $O(n)$.

These measures and their interpretation differ from those computed for the individual vector-to-vector analysis. When applied to the whole content of the space, the centroid distribution analysis provides an indicator of whether a space is narrowly focused. If the average cosine to the centroid is high, then most of the documents are mapped to similar meanings and there is less potential to differentiate between them. A lower average or maximum cosine, however, is not necessarily indicative of a problem with the space. A widely distributed space should be expected to have a lower average cosine to the centroid. Further, the minimum cosine from this analysis may give some indication of the presence of outliers, items that are not like the general content of the collection. Again, further analysis may be performed by binning the resulting cosine values at a selected level of granularity and observing the distribution pattern of the set of cosines.

Targeted Subset Centroid Analysis

The same computation used in the full space centroid analysis is also easily applied to a targeted subset of the content of the space. This may be performed to assess the cohesion or dispersion of certain topic areas or a special set of items identified within the space. The vectors associated with the subset are simply averaged to produce a representative centroid. Individual vectors in the set are then compared to this subset centroid vector, and these measures are interpreted similarly to the full space centroid analysis. The number of cosine calculations required for completing this analysis is again simply n, where n is the number of items in the subset being examined, thus the complexity for the targeted subset analysis is simply $O(n)$. Depending on the selected subset, n may also be significantly smaller than the size of the item set used for the full space centroid analysis.

Taken in concert with other subset centroids and the full space centroid, the target subset analysis approach allows various aspects of a space to be probed in detail. Use of this method for comparison of two spaces does require that the targeted content be present in both. The overlapping document or term content for S_1 and S_2 is a candidate subset for evaluation with this method, though spaces with little or no overlapping document content would not be amenable to this approach. Since this method requires specific a priori knowledge about the document space or the ability to modify the space to insert such targeted content, its use is not practical in most cases.

3.2 Projected Content Comparisons

The next level of comparison methods, projected content comparisons, is the examination of a selected set of items which is projected into each of the spaces being compared. The method for projecting an item into an existing space was described in Section 2.3.1 and involves using term vectors present in the space to compose new document vectors that were not in the original content of the space. Because of this, projection content comparisons can only be performed on the document portion of the spaces, but they encompass the meanings represented by the term vector portion of the spaces as well.

In the projected content approach, a set of projection items P is mapped into both S_1 and S_2 yielding a set of projection vectors for each space, P_1 and P_2 respectively. Since P_1 and P_2 relate the same set of items, P, differences in the relationships between items in these sets provides direct insight into differences in the meaning being represented in their associated spaces. As with the direct comparison measures, distribution analyses may be applied to P_1 and P_2 to obtain a composite picture of the way the items in P are being mapped. Further examination of the changes in individual relationships between P_1 and P_2 can be accomplished by calculating the degree of three-tuple order changes observed between items in each of the mappings.

Various possibilities exist for selecting items to be included in P. These include the use of a standard set of items, the selection of one or more subsets of items from the collections themselves, the projection of one entire collection into another, or the use of randomly selected items for P. Were the projection items representative of the type of data to be analyzed by the spaces in use, the relationship changes could be anticipated to describe the degree of impact of the difference in the spaces on the data of interest. The identification of a projection set that would give more generalizable results across multiple subject spaces is desired. In any case, one drawback of the projected content approaches in this category is that projection of content requires access to suitable content items. These techniques will not be usable in cases where projection sets are desired to be selected from a space where the original document content is not available.

3.2.1 Projection Set Distributions

Once the projection sets P_1 and P_2 have been produced, a vector-to-vector distribution analysis and a centroid distribution analysis may be performed. Just as these are computed for the direct comparison measures, the vector-to-vector distribution characterizes the pairwise cosines between each of the items in P and the centroid distribution does the same for each of the items respective to the centroid for the set. Results from the direct comparison measures are limited to contrasting the general characteristics of the spaces S_1 and S_2 since they are based on different items to some degree (unless a targeted subset analysis of overlapping items is being performed). Differences in the distributions between P_1 and P_2 however can be plainly observed to reflect the way that the meanings represented in their associated spaces affect the mapping of P. Instead of describing the content of S_1 and S_2 these differences describe the influence of S_1 and S_2 on a common set of projection items.

Since the vector-to-vector distribution and centroid distribution are produced via the same process as in the direct comparison measures, the computational complexity is the same: $O(n^2)$ for the vector-to-vector distribution analysis and $O(n)$ for the centroid distribution analysis. For this application $n = |P|$, which in most cases would be selected to be smaller than the number of items in S_1 or S_2, but might still be of significant size.

3.2.2 Three-tuple Order Comparisons

After computing the entire set of pairwise cosines between the items in P_1 and P_2, the three-tuple relations between the documents can be examined noting the relationship changes from P_1 to P_2. The three-tuples are defined as follows:

$$\left\{(A, B, C) \middle| \, A = p_i, B = p_j, C = p_k, \text{where } i \neq j \neq k, \forall p \in P\right\}$$

The reason for examining three-tuples is to observe changes in the relative mappings between P_1 and P_2 without regard for the magnitude of the specific cosine values involved. Each three-tuple has three relationships with a definite ordering by similarity. A comparison of one such three-tuple is visualized in Figure 3.2 which depicts a case where the three-tuple relationships all change. In P_1, a set of three documents projected in S_1, **BC** represents the closest pair, **AB** the next closest, and **AC** the most distant. P_2, the same documents projected in S_2, have all changed their relative positions such that **AB** is the closest pair, **AC** the next closest, and **BC** is the most distant. These differences represent a complete change in relative meaning for the documents involved based on the space they are projected into, S_1 or S_2. Differences in the cosines between P_1 and P_2 that do not show changes in the relative ordering are less significant as they may simply reflect a general inflation or deflation in the cosine values as a whole and do not indicate a change in the meaning relationships

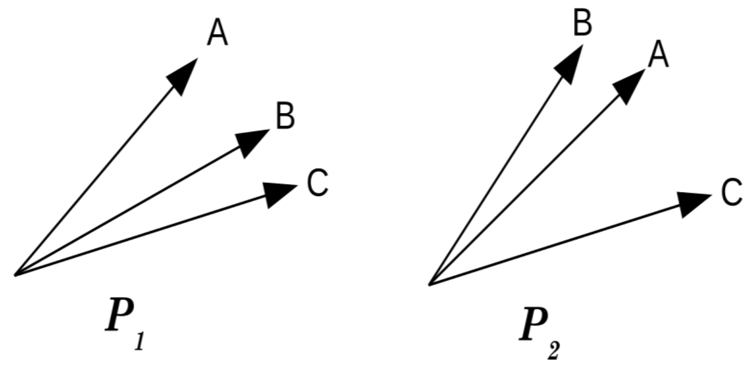

Figure 3.2 – Illustration of changes in relationships between projected items A, B, and C from projection set P as they are projected into spaces S_1 or S_2.

being ascribed to the projection items.

Computation of the three-tuple order comparisons is an expensive process. First, the vector-to-vector cosines between all of the items in P_1 and their counterparts in P_2 must be produced. This results in k cosines for each of the sets P_1 and P_2: $k = \binom{n}{2} = \frac{n(n-1)}{2}$, where $n = |P|$. So producing the k pairwise cosines is $O(n^2)$. Using these k cosines, the total number of three tuples that can be produced for each set is $\binom{k}{3} = \frac{k(k-1)(k-2)}{6}$. Evaluating the changes in relative tuple ordering between P_1 and P_2 requires comparisons for each of these tuples and since k is $O(n^2)$ the final complexity for the entire operation is $O(n^6)$, where $n = |P|$. Therefore, the size of the selected projection set P is critical to the usability of this approach.

3.3 Rotations and Other Transform Comparisons

The third area of investigation for space comparison measures is the use of a rotation method for identifying a transform between the spaces. By defining a set of anchor items A in both spaces (items of known fixed meaning that are mapped in both S_1 and S_2 as A_1 and A_2 respectively), a transform can be derived as a rotation matrix Q that maps items from one space into another using the *Procrustes algorithm* (Golub and Van Loan, 1996). Applying the transform is as simple as taking the vector V_1 mapping the item in question in S_1 and multiplying it by the matrix Q to yield V_2 mapping the same item in S_2 according to its dimensional axes.

There are several major advantages to this approach that make it attractive. First, it deals directly with the spaces themselves without need for access to the underlying raw document content used to create them. Anchor documents may be selected from items already present in the spaces, but this is not necessary as they may be defined as any set of items that are intended to remain fixed in their meaning.

Second, this approach allows several simple direct comparison items between the two spaces using standard similarity measures that simplify the overall difference evaluation. Since items from S_1 can be mapped directly into S_2, the whole space centroids or individual item vectors can be compared by computing their cosine similarity. Net changes in overlapping terms may be examined individually or as a composite, as can test sets of projection items.

Finally, depending on the number of anchor points required, the computability of the rotation matrix and subsequent difference comparisons may be significantly less expensive than other methods available. Production of the Q transform matrix requires a matrix multiplication of ${A_1}^T A_2$, a process with a complexity of $O(|A|^2 n)$, and relies on the dense matrix SVD process which has a widely published complexity of $O(n^3)$ or better depending on the specific algorithm used (Trefethen and Bau, 1997). In this application $n = k$, the smallest rank (number of dimensions) of the two spaces being compared. The value of k is typically in the range of 300 to 500 (see Section 2.1.1) and is much smaller than $|A|$, $|P|$, or $|S|$ in general.

3.3.1 Production of the Q Transform

The rotation matrix Q is obtained by first projecting a set of anchor items A into the spaces being examined, S_1 and S_2, to produce a set of mapping vectors for the anchor items in each space, A_1 and A_2 respectively. Singular Value Decomposition (SVD) is performed on ${A_1}^T A_2$ to yield the matrix factorization of ${A_1}^T A_2 = U \Sigma V^T$. Finally, Σ is discarded and Q is computed as the matrix product of UV^T. Using this process Q is found such that the matrix Frobenius norm $\|A_1Q - A_2\|_F$ is minimized (Golub and Van Loan, 1996).

3.3.2 Comparative Space Centroid Analysis

Once the rotation matrix Q has been produced, it is a simple matter to map individual items from one subject space into the other and calculate a similarity measure between the two. The whole space centroids for $S_1$and S_2 are natural candidates for comparison. The centroid vector from S_1, C_1, is simply multiplied by Q to yield its analog in S_2, $\widehat{C_1}$. Comparison between $\widehat{C_1}$ and C_2, the S_2 centroid, is performed using a simple vector cosine.

Several variations on this theme are possible. Since any of the items from S_1 can be mapped in S_2 using the Q transform, all or some subset of the document or term items from S_1 can be mapped into S_2 to obtain their analog in S_2, $\widehat{S_1}$. A full space centroid analysis may be performed on $\widehat{S_1}$ and the resulting distribution and centroid, $Centroid(\widehat{S_1})$, compared to the S_1 distribution and the remapped centroid $\widehat{C_1}$. Changes in the distribution between S_1 and $\widehat{S_1}$, and differences between $Centroid(\widehat{S_1})$ and $\widehat{C_1}$ are indicative of differences in the meanings represented in S_1 and S_2.

Rotation of the S_1 items into S_2 is a matrix multiplication of S_1, an $n \times k$ matrix, by Q, a $k \times k$ matrix where k is the number of dimensions used in the creation of S_1. This results in a total of nk^2 multiplications but since k is a fixed value (typically in the neighborhood of 300) the complexity of the transform operation is still considered to be $O(n)$. Calculation of the whole space centroids and distribution is of $O(n)$ complexity as described in Section 3.1.2.

3.3.3 Comparative Term Vector Analysis

Since the production of Q minimizes the Frobenius norm (F-Norm) of the two anchor sets, examination of its application to the difference between the sets of overlapping term vectors for S_1 and S_2 gives another indication of the direct difference between the two spaces involved. Calculating the F-norm for the difference between the sets of overlapping terms involves identifying the vectors for the terms common to both S_1 and S_2 and sorting them into two corresponding term matrices, T_1 and T_2 with the same ordering. The F-norm of the difference is then computed as follows (Golub and Van Loan, 1996):

$$\|T_1 Q - T_2\|_F = \left\|\hat{T}\right\|_F = \sqrt{\sum_{i=1}^{|\hat{T}|} \sum_{j=1}^{k} \left|\hat{t}_{i,j}\right|^2}$$

The magnitude of the resulting F-norm corresponds to the net difference between the mappings of T_1 and T_2 and therefore is also indicative of the difference in the way S_1 and S_2 represent meaning. A result of zero would indicate that the two sets of overlapping terms were essentially identical in their mapping of meaning for the set of overlapping terms, T.

Calculation of the F-norm across the difference between the overlapping term vectors is of $O(n)$ complexity. The term vectors must be matched in the same order for the calculation to be performed correctly, but this can be accomplished during the construction of T_1 and T_2 without added cost. The resulting measure was named the Overlapping Term Vector Norm or OTV-Norm.

CHAPTER 4

INVESTIGATIVE EXPERIMENTS AND RESULTS

Six sets of controlled experiments were conducted to explore the relationships of the various comparative metrics available and determine if there was a computationally expedient measure or set of measures that could be used to easily quantify the semantic differences between two LSA spaces, S_1 and S_2. Several different document collections were assembled to be used in the construction of spaces that could be compared. These datasets and their composition are described in Section 4.1, and construction of the corresponding LSA spaces is described in Section 4.2. The various measures described in Chapter 3 were calculated for these spaces individually, where applicable, and in two-way comparison combinations for each of the spaces within the six experimental sets. These experiments and the collection of the measurements and results are described in Section 4.3.

4.1 Datasets

As a data source for the initial experiments, documents were selected from the Reuters Text Retrieval Corpus (RTRC) RCV1 collection (Lewis et al., 2004) to provide varied content deemed to be at a typical adult reading. The content in this set consists of 806,791 English language news articles on various sources published by Reuters in 1996-97 and made publicly available for use in research. A second graded content data source was used for subsequent experiments in order to provide evaluation comparisons while controlling for the influence of varying types and quality of content. This collection consists of over 3.5 million paragraph-sized documents each tagged with a reading complexity level roughly corresponding to US education system grade levels (Landauer and Way, 2012). The use of this set allows for the construction of LSA spaces that can be expected to have some variation in the type of meaning represented in the space as it is possible to control the level of sophistication of the content included in the input to the LSA space. This collection has previously been used in studies evaluating the maturity of vocabulary understanding at different levels of reading complexity (Landauer et al., 2011). Finally, as described in Section 4.1.3 additional textual content was taken from randomly selected texts from various literature collections (Landauer et al., 1998a), as well as a set of content from public driver training manuals used in the NICHD project (Martin et al., 2016). This additional content was used to construct “third-party” projections into two spaces being compared or as anchors for rotational transforms.

4.1.1 RTRC News Articles

An initial group of datasets for comparative analysis was constructed by randomly selecting items from the RTRC corpus. The RTRC corpus consists of newswire articles from 1996-1997 spanning several different categories, and each article is tagged with several category indicators for both the general content category and multiple specific content sub-categories. Many of the articles contain market reports and numerical financial data with little or no narrative commentary. To obtain more general linguistic content, only items from the Government/Social top level category were selected for use as they contain more regular text. This subset consists of 234,873 documents. Four documents from this set were discovered during processing to contain mal-formed UTF-8 character sequences and were eliminated. The remaining 234,869 documents were taken as a whole for one set, and an additional two partially overlapping subsets of 150,000 each were randomly selected from this group. A final set of documents was parted from the first 1,000 articles in the Government/Social document set for use in initial testing, and while a set of this size is too small to be considered adequate for representing linguistic semantics (Landauer, 2007), it is included in some of the analysis results for contrast. The document and term counts for these document sets are detailed in Table 4.1.

Table 4.1 - RTRC document sets created for use in the experiments. The 150k, 150k-B and 1k sets are all subsets of the full RTRC-GCat collection.

Set	Documents	Terms
RTRC-GCat-150k	150,000	215,769
RTRC-GCat-150k-B	150,000	225,754
RTRC-GCat	234,869	272,739
RTRC-GCat-1k	1,000	21,103

4.1.2 Grade Level Series

As a foundation for comparative analysis, four incremental content groups were constructed from the grade level tagged content as series of five additive sets (Series A – D). Each set within a series was built incrementally by amending to the content of the previous set 50,000 additional randomly selected content items representing each grade level in the step being added. This was done in a balanced manner where each grade level was equally represented in the number of content items to the extent possible. This resulted in 20 total sets in four series of five sets each consisting of content through grade levels 3rd, 6th, 9th, 12th, and college (grade 16) respectively as shown in Table 4.2. Each grade level space contained representative texts a reader might be exposed to up to that grade level. Within a series, the grade level 6 set contained all of the content in the grade level 3 set, adding another 50,000 documents selected evenly from the content available for grade levels 4 – 6. Likewise, the grade level 9 set contained all of the content in the grade level 6 set, again adding 50,000 documents from the content available for grades 7 – 9. Due to the limited amount of document content available for the very elementary grade levels, a full 50,000 document selection was not available for the third grade sets, but all

Table 4.2 – Grade Level Document Sets Series A – D created for use in the experiments

Grade Level		Series A	Series B	Series C	Series D
3rd	**Documents**	37,446	37,525	37,367	37,475
	Terms	45,130	44,975	44,944	45,133
6th	**Documents**	87,446	87,525	87,367	87,475
	Terms	65,784	65,806	65,681	66,001
9th	**Documents**	137,446	137,525	137,367	137,475
	Terms	93,677	93,569	93,185	93,727
12th	**Documents**	187,446	187,525	187,367	187,475
	Terms	128,375	128,792	128,456	128,379
College	**Documents**	237,446	237,525	237,367	237,475
	Terms	173,731	174,473	174,126	173,830

subsequent additions in each series consisted of a full 50,000 documents. These incremental series were constructed to mimic the growth of content volume and complexity of a human reader at the various grade levels represented.

Since the content sets within each series increased in size with each additive step, it was possible that the differences in collection size between the incremental sets were impacting the comparison measures rather than the grade level aspect of the sets. To address this question, two additional datasets were created with larger content amounts at the intermediate grade levels. These two collections were randomly selected from texts up to 6th grade and 9th grade reading levels respectively until sets of approximately the same size as the college level sets were identified. A 6th grade set was created consisting of 229,753 documents and 116,586 terms was assembled, as was a 9th grade set

Table 4.3 - Non-overlapping Grade Level Document Set Series A – B created for use in the experiments

Grade Level		NO-Series A	NO-Series B
3rd	Documents	37,468	37,468
	Terms	45,312	45,454
6th	Documents	87,468	87,468
	Terms	66,102	66,165
9th	Documents	137,468	137,468
	Terms	93,979	94,135
12th	Documents	187,468	187,468
	Terms	129,093	129,230
College	Documents	237,468	237,468
	Terms	174,977	174,960

consisting of 240,884 documents and 126,586 terms. The selection of content items for each of these sets was again balanced across grade levels to represent each component grade level with a roughly equivalent amount of material. These sets were partially overlapping in document content both between themselves and with the sets in the Grade Level series.

After the randomly selected grade-level series sets were constructed it was observed that overlapping document content between the series might influence the results. Two additional series sets of non-overlapping documents were constructed to control for this condition as shown in Table 4.3. These sets were constructed in a similar manner as the previously selected series in increments of 50,000 documents, with the exception that a document was only permitted to be included in a single series. The resulting ten total sets in two series (A and B) of

Table 4.4 – Frozen vocabulary document sets created for use in the experiments

Set	Documents	Unique Terms	Term Instances
Base GL 3	28,270	33,069	2,344,511
GL 6A	128,268	33,069	11,546,926
GL 6B	128,272	33,069	11,547,196
GL 12A	228,268	33,069	20,960,487
GL 12B	228,272	33,069	20,963,119

five sets each contained content through grade levels 3, 6, 9, 12, and college respectively. As with the previous series, a full set of 50,000 documents could not be obtained for the third grade sets due to lack of available content at the elementary level.

Finally, to address the question of the impact of term overlap on the measurements, a group of document sets were assembled as a series with a common base group as a foundation. The foundational documents were comprised of 28,270 randomly selected elementary content items from the grade level collection up to 3rd grade. These items were selected in a similar manner as the previous grade level series groups with each grade equally represented in a balanced manner. The term vocabulary was frozen at this point to the 33,069 terms that had been identified in the base set of documents, and two series of document sets were created using this base. The two series were randomly selected to be sets of an additional 100,000 non-overlapping documents up through the 6th grade level and then another 100,000 non-overlapping documents up through the 12th grade level (see Table 4.4). The term vocabulary was held constant at each addition beyond the base, so no new terms were recorded. The

Table 4.5 - Document sets created for use as projection items and anchor items

Set	Documents	Unique Terms	Term Instances
NICHD04	1,060	5,912	70,063
T-500	500	16,317	123,668
T-1000	1,000	24,319	252,372
T-5000	5,000	49,995	1,281,749

subsequent spaces then all maintained 100 percent term overlap with the base space and with each other.

4.1.3 Other Content Sources

The additional content listed in Table 4.5 was also required for these experiments to implement the projection set measures and analysis described in Section 3.2, and as anchor items for the rotational measures described in Section 3.3. The first set of projection items constructed (NICHD04) was a set of 1,060 paragraph sized documents taken from public driver information manuals that were collected in the NICHD Driver training project (Martin, et al., 2016). These documents form a set of items that are all focused on a similar related topic and are expected to exhibit a high degree of semantic similarity. Three additional sets of paragraph sized documents (T-500, T-1000, T-5000) were sourced from randomly chosen texts from various literature collections dispersed over a wide array of topics (Landauer et al., 1998a).

4.2 Space Construction

Each of the 42 document sets assembled were parsed and processed to produce corresponding individual LSA spaces using the LSA_Toolkit™ software provided

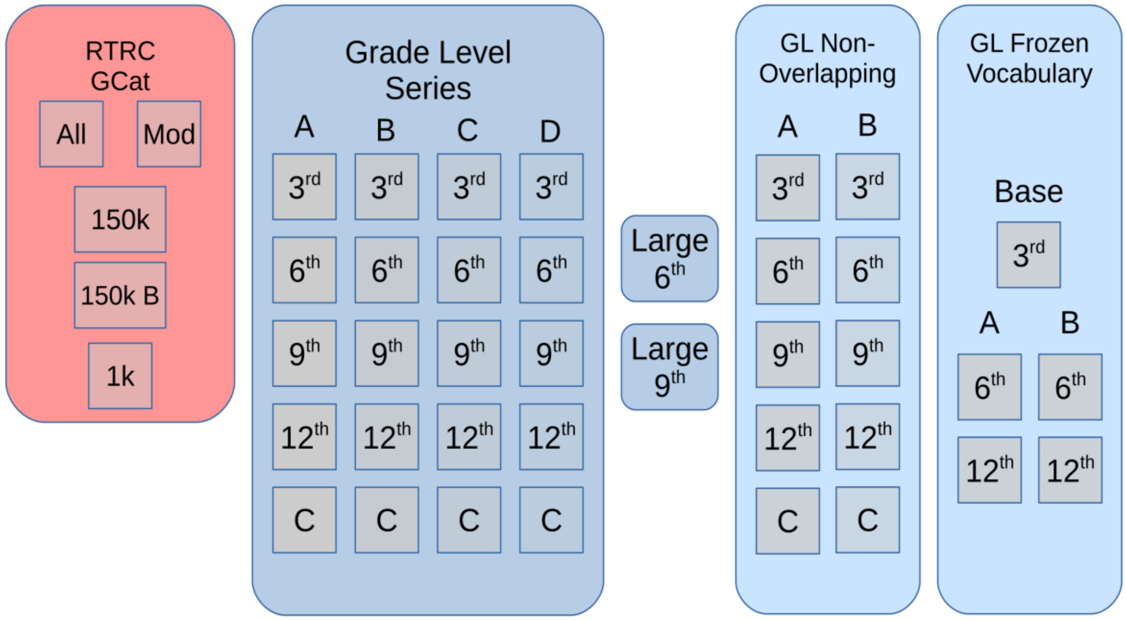

Figure 4.1 –Depiction of the groups and series of LSA spaces created for these experiments

by Small Bear Technologies, Inc., a commercial software package for LSA processing. All of the input documents were XML encoded and had to be parsed and tokenized to remove any markup tags and punctuation. Punctuation within the texts was processed using the common policy of replacing any character that was not in the alphanumeric English character set with a blank (Landauer and Dumais, 1997). All input text was processed, and no stop-lists or filters were used to remove words. The resulting parsed document sets were then weighted using a log-entropy weighting as described in Section 2.1.1, and LSA spaces were then built to exactly 300 dimensions for each set.

As shown in Figure 4.1, one additional control space, RTRC-GCat-Mod, was built to provide a basis for making comparisons across two spaces that were nearly identical. This was done as an early control experiment to verify that the measures being studied were not overly sensitive and that minor differences between two spaces would not show significant differences in the measurements. This control space was constructed by taking the RTRC-GCat set and removing ten randomly selected terms from all of the documents in which they appeared and rebuilding the space. This modified RTRC-GCat space had the same document count as the original space and introduced no new terms. By comparing this modified space to the original unmodified RTRC-GCat space the stability of the various measurements could be tested.

In total, 6,106,470 document instances were processed with 633,490,622 total term instances across those documents to produce the 42 LSA spaces used in these experiments. Individual term and document counts for each of these spaces is included in Table A - 1 in the Appendix.

4.3 Measurements

To begin the analysis process, individual direct measurements were taken for each of the test spaces. These measurements included the simple raw metrics and the more elaborate distribution analysis as described in Section 3.1. The process and details of these measures are discussed in Section 4.3.1, and selected results are included in the Appendix. Once the individual test space analysis were completed, the various sets were compared in pairs on the basis of their individual measures, using the projected content measures described in Section 3.2, and using the rotational measures described in Section 3.3. The process used for each of these measures is described in the following Sections 4.3.2 and 4.3.3 respectively, with selected results again detailed in the Appendix.

The pairwise comparison of spaces began with the initial control experiment comparing each of the measures matching the slightly modified RTRC-GCat-mod space with the original unmodified RTRC-GCat space. In the second experiment set, the four primary RTRC spaces (GCat, GCat-150k, GCat 150k-B, and GCat-1k) were processed against each other using all of the measures. These results provided a baseline for contrasting the results from the other experiments to be performed. The next experiment set to be conducted was the evaluation of the twenty spaces in the grade level series group. For this experiment, each space was compared vertically to the higher level spaces within its own series, and longitudinally to all of the spaces across the other three series. The grade level data sets were compared vertically within each series to observe measurement differences that occurred as a result of growth in the level of sophistication in meaning development simulated by the LSA spaces, both by the increase in content volume and reading complexity. The longitudinal comparisons across the series, both within grade level and to different grade levels, were intended to explore the possibility that changes in content at the same simulated grade level would still produce notable differences in the measurements being tested, though this was not expected to be the case, and to verify that changes in the grade level would be reflected similarly in the measurements involving sets where the base content was not the same.

The fourth experiment set was intended to explore the impact of the volume of content, the total number of documents, on the measurement results. To accomplish this, the large volume 6th and 9th grade spaces were compared to each of the spaces in the grade level series as well as to each other. The large volume 6th and 9th grade spaces increased the amount of content but remained at fixed grade levels to test the impact of collection size on the measurement results while controlling the nature of that content.

The fifth experiment set addressed the impact of document overlap on the measurements between the sets being compared. The ten spaces from the two non-overlapping graded document set series were processed both vertically and longitudinally, comparing each space to the higher level spaces within its own series and then to all of the spaces in the other non-overlapping series. This provided the same view of the spaces as in the previous grade level experiment set while virtually eliminating the presence of overlapping documents between the series. Each space in the non-overlapping series was then also compared to all of the spaces in the original grade level series to provide an additional view of the measures involving those spaces.

The final experiment set examined the influence of term overlap on the comparison measures by comparing the frozen vocabulary spaces against each other. In addition to controlling the term overlap between the spaces, the two series in this group contained no overlapping document content other than the documents that were in the initial base set that was shared between them.

Each of these experiments, except the frozen vocabulary set, was conducted using all four of the different projection sets listed in Table 4.5 as projection and anchor items. The frozen vocabulary set was processed against only the first three projection sets, omitting the large T-5000 set. A total of 592 pairwise space comparisons were performed across the 42 spaces for these experiments. Of these, 582 were processed against four different sets of projection items each, and the remaining ten were processed against three different sets of projection items.

4.3.1 Direct Measures

Direct measurements were collected or computed for each of the LSA spaces being examined individually. The initial measures of document and term count totals were collected during the document parsing process as the spaces were being constructed. Following that, a full distribution analysis was performed on each space completing the individual space metrics. Each pairwise comparison of spaces was then examined to calculate both the document overlap and term overlap ratios.

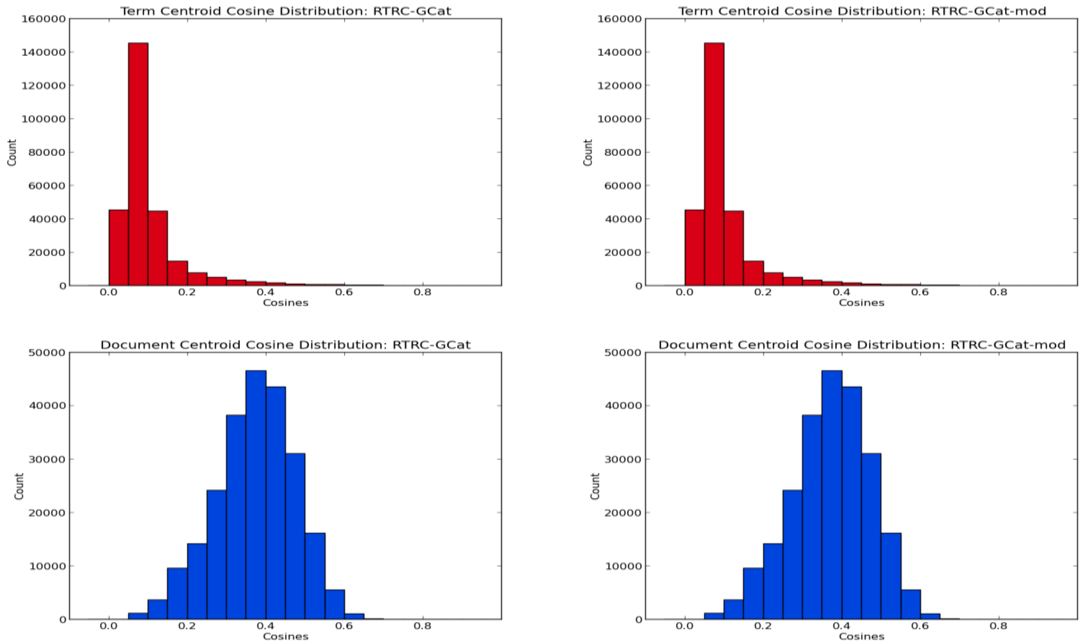

Figure 4.2 – Term centroid and document centroid distribution comparison between RTRC-GCat space (left) and the control RTRC-GCat-mod space (right)

Distribution Analysis

A simple analysis program was used to produce the distribution analysis measures for the document space centroid, the term space centroid, and the document to document distributions described in Section 3.1.2. First, centroid vectors were computed for both the full set of document vectors and the full set of term vectors in the space. Cosines between the individual vectors to their and respective centroid were computed and collected in total for the purpose of generating a histogram, as well as calculating the mean and standard deviation for each of these distributions. A full vector-to-vector distribution analysis was also performed for the document items in the space by computing the cosines between every pairwise combination of the document vectors. Again these cosines were collected for generating a histogram as well as calculating the mean and standard deviation for the document-to-document distribution. The resulting output for each analyzed space was graphed, as depicted in Figure 4.2 for the control comparison of the RTRC-GCat space with the RTRC-GCat-Mod space, and recorded to be used for comparison between spaces. Details for the distribution analysis results are included in the Appendix in Table A - 2.

Document overlap ratios

For each comparison between two spaces the document overlap ratio was calculated. While it is possible to achieve this through careful cataloging of the document collections in question, document overlap for these experiments was obtained by direct examination of the term frequency vectors in the document sets programmatically to eliminate the need for a single document cataloging system, avoid the possibility of mismatched catalog items, and to account for the possibility of the same document content being introduced multiple times under different document identifiers. Document overlap counts as well as the computed ratio were recorded and graphed both as individual comparison reports and across the grade level series sets.

All of the grade level series sets contained overlapping documents within their series since each successive grade level set included all of the documents from the previous grade level sets. This can be seen in Figure 4.3 which shows the document overlap comparison between the Series A Grade Level 3 space and the other grade level spaces in the series. There were also overlapping documents with the sets in the other series due to the nature of the random selection process used to construct the sets which were all taken from the same pool of documents. The non-overlapping grade level sets contained overlapping documents within their series since again each successive grade level set included all of the documents from the previous grade level sets. There were however, no overlapping documents between the two series as they were constructed specifically to be non-overlapping in document content. The RTRC document sets, being from an entirely different source collection contained no

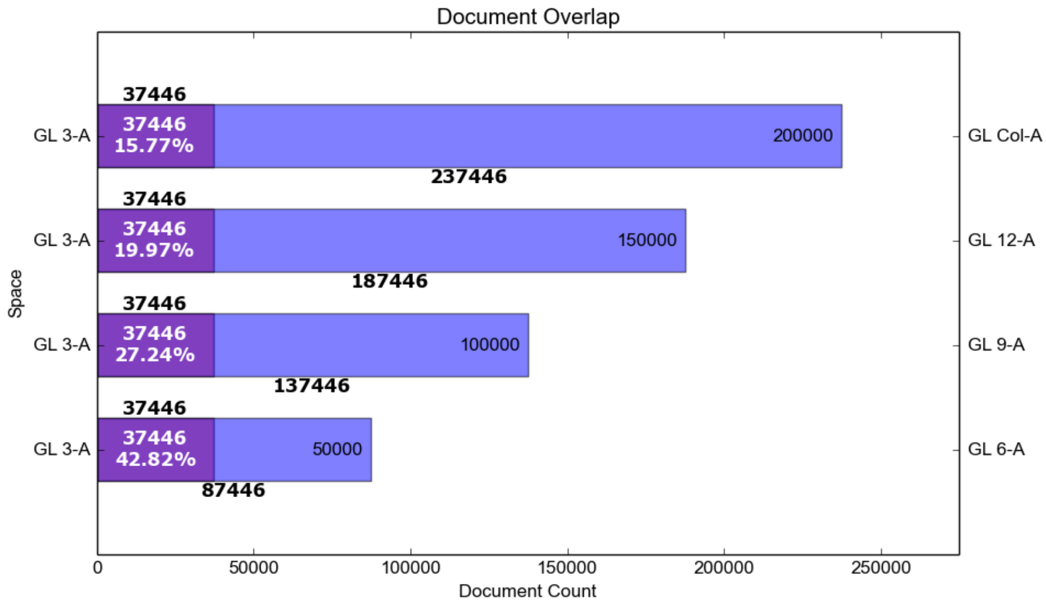

Figure 4.3 - Document overlap comparison between the Series A Grade Level 3 space and the other Series A spaces. Each bar depicts the total number of documents in each space, the number of overlapping documents between the spaces, and the document overlap ratio.

overlapping documents with any of the grade level series sets or non-overlapping grade level series sets. This fact was specifically verified programmatically by examining the term frequency vectors in the parsed document sets directly. The RTRC document sets did contain overlapping documents with each of the other sets from the RTRC source, and the control space, RTRC-GCat-Mod, completely overlapped the RTRC-GCat space as only terms were modified between the two spaces. Details of the document overlap metrics for each of the compared spaces are included in Table A - 3 in the Appendix.

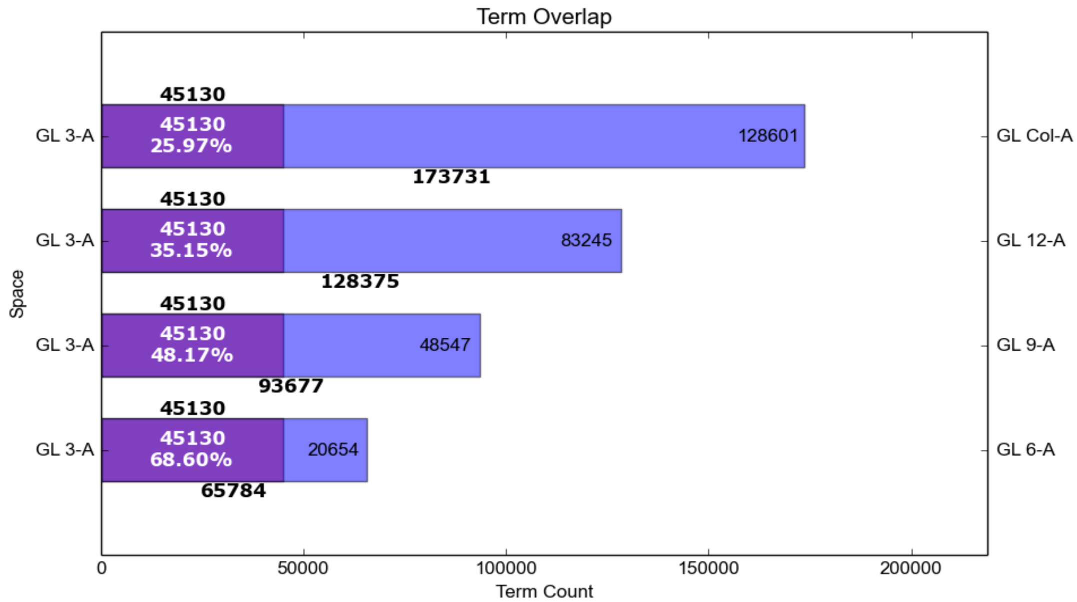

Figure 4.4 - Term overlap comparison between the Series A Grade Level 3 space and the other Series A spaces. Each bar depicts the total number of terms in each space, the number of overlapping terms between the spaces, and the term overlap ratio.

Term overlap ratios

Finally, the term overlap ratio was computed for each of the LSA spaces being compared. As noted in Section 3.1.1, the term overlap for two spaces is easily obtained by simply counting the number of terms that are duplicated between them. This was done by accessing the term dictionaries for each of the spaces being compared. Term overlap counts as well as the computed overlap ratio were recorded and graphed both as individual comparison reports for each of the spaces compared, and across the grade level series spaces as a group for correlation with the other measures. Figure 4.4 shows the term overlap comparison between the Series A Grade Level 3 space (GL 3-A) and the other grade level spaces in the series. In these series all of the terms in the GL 3-A space appear in the other spaces and the non-overlapping terms appear only in the spaces being compared to the GL 3-A space. Details of the term overlap metrics for each of the compared spaces are included in Table A - 3 in the Appendix.

4.3.2 Projected Content Measures

Projected content measures use a third set of content items to generate projections into two spaces being compared as described in Section 3.2. By examining the positioning of the projected items relative to each other in each of the two spaces, differences in the way the two spaces affect the semantic mapping of the projected items may be observed. Several different sets of items were used as projected items varying in both size and content. These sets included:

- Items with related content
    - NICHD04 – 1,060 text items on driving
- Randomly associated content items
    - T500 – 500 paragraphs from assorted literature
    - T1000 – 1,000 paragraphs from assorted literature
- Larger item sets
    - T5000 – 5,000 paragraphs from assorted literature

Each of these content sets was parsed and projected into the spaces being compared according to the process described in Section 2.3.1. As each set was parsed, the number of mappable terms for the set, the term "hit-rate", was recorded. A term is mappable within a space if there exists a term vector in the space corresponding to the term in question. The term itself must have been present in the original content used to construct the space. Once the projections were complete, the relationships between the projected item vectors were analyzed by performing a distribution analysis on the projected item vectors and a three-tuple order comparison on the projected item vectors.

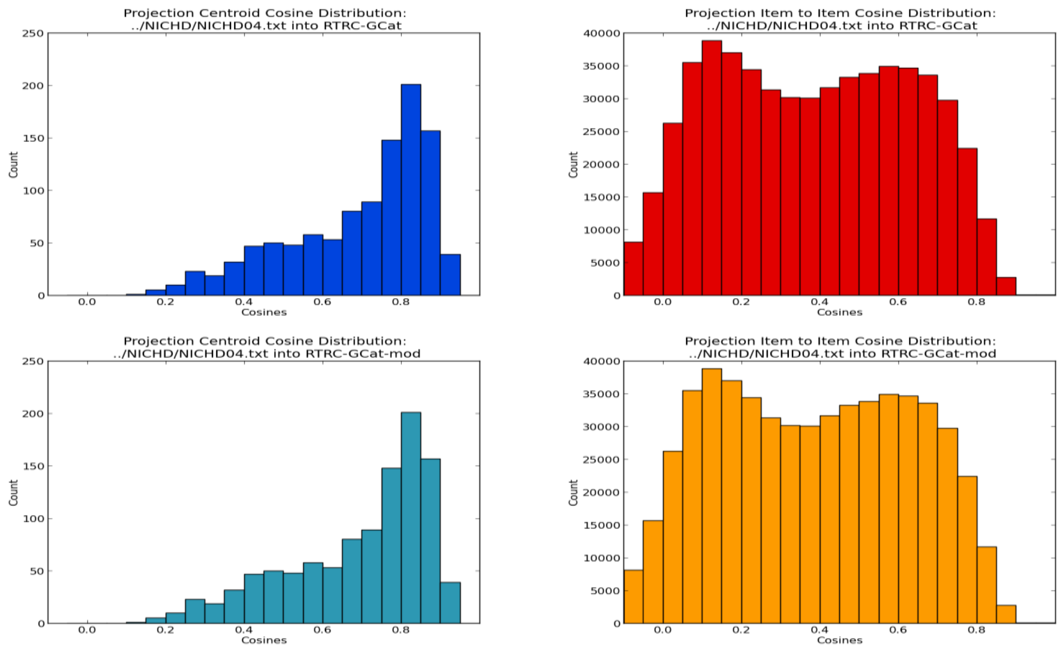

Figure 4.5 - Projection set distribution comparison between RTRC-GCat space and the control RTRC-GCat-mod space showing the projection centroid cosine distributions (left) and the projection item-to-item cosine distributions (right)

Projection set distributions

As with the distribution analysis performed on the individual spaces themselves, the analysis of the projection set distribution was performed by comparing the projection vectors to a centroid of the projected items. The projection centroid vector was computed for the projection items and then cosines between the individual projection item vectors and the centroid were computed and collected. These were used to generate histograms depicting the distribution of the items relative to the centroid, as shown in Figure 4.5, as well as calculating the mean and standard deviation for each of this distribution. A full vector-to-vector analysis was also performed for the projection items by computing the cosines between every pairwise combination of the projected item vectors. Likewise, these cosines were collected for generating a histogram as well as calculating the mean and standard deviation for the item-to-item distribution. The resulting output for the projection set in each of the spaces being analyzed was recorded to be used for comparison between sets and correlation with the other measures.

Three-tuple order comparisons

After the pairwise projection item cosines were computed for the projection set in each of the spaces being compared, the three-tuple comparisons were performed using the pairwise cosines to form the three-tuples as described in Section 3.2.2. Each three-tuple from the projected items in the first space was compared with its corresponding three-tuple in the second space to determine if any of the relative relationships changed from one space to the other. If a change in ordering was observed, the type of relationship change was noted as the maximum similarity pair changing (MaxRel), the minimum similarity pair changing (MinRel), or both the maximum and minimum similarity pairs changing (BothRel). During processing, these changes were recorded by keeping three simple counters for the number of instances where MaxRel, MinRel, or BothRel changed. Since both the MaxRel changes and the MinRel changes were included with the changes counted in BothRel, the total number of changes was computed as the sum of the MaxRel counter and the MinRel counter less the BothRel counter. These relationship changes were then examined across the spaces in the various grade level series as illustrated in Figure 4.6, both within a single series and across the different series and among the different projection sets. The total relationship change percentage values observed for each of the experiments is included in the Appendix in Table A - 3. Additionally, the top five instances of each type of change, as measured by magnitude of the cosine difference, were also stored and reported individually with the associated document identifiers and cosine values for further examination.

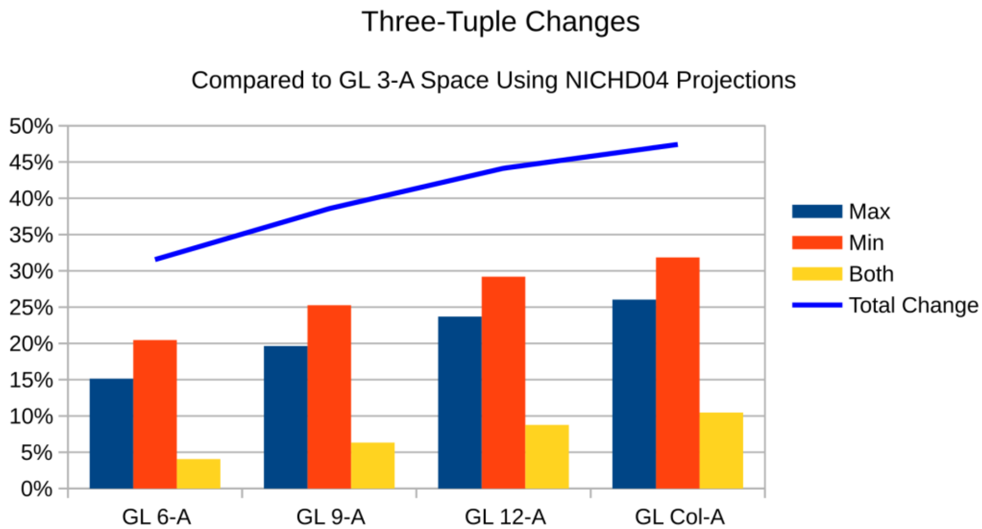

Figure 4.6 - Three-tuple comparison of the Series A Grade Level 3 space and the other Series A spaces using the NICHD04 projection set

4.3.3 Rotational Measures

The final comparison measures evaluated in these experiments were the methods based on the identification of a rotational transform between the two spaces being compared. As described in Section 3.3, the *Procrustes algorithm* was used to produce a rotation matrix Q that could be used to transform items from their mapping in one space to a mapping in another space. Using this rotation matrix, comparisons were made between spaces using the remapped term and document space centroids, and by performing a comparative term vector analysis for the overlapping terms between the spaces being compared.

Since rotational transformation between two spaces is a directional operation, that is, one space is being rotated to fit the dimensional axes of the other, some tests were performed in both directions to examine the stability of the measures. For these tests, space S_1 was rotated into space S_2, and then likewise space S_2 was rotated into space S_1. The rotation-based measurements were taken using both transformations and examined to determine if they were similar, or if the direction of the transform produced significant differences.

Construction of the rotation transform

As described in Section 3.3.1, production of the rotation matrix Q for transforming the mappings of items in one space into the mapping of another space is based on the definition of a set of anchor items that are taken to have the same mapping in both spaces. This anchor set can be any group of content items whose relative mappings are desired to remain constant in both spaces. For these experiments, anchor items were randomly selected from content documents that were not present in either of the spaces being compared. Since the LSA spaces for these experiments were all created at 300 dimensions, a minimum size for the anchor set of 300 items was required in order to produce a transform that covered the available dimensionality of the spaces. Comparison tests were performed using as anchor sets the full 1060 items from the NICHD04 document set, a subset of 500 of the NICHD04 items, as well as the T-500 and T-1000 document sets. A final set of 5,000 randomly chosen texts was also used in a several series of tests to examine whether the measures were significantly affected by the use of a larger number of anchors.

For each comparison test, these anchor sets were systematically projected into each of the two LSA spaces being compared. The term hit-rate for the sets of anchor projections was recorded for both sets. The projection vectors were then used as the input to the *Procrustes algorithm* detailed in Section 3.3.1, requiring a relatively small (300 by 300) dense SVD be performed to produce the rotation matrix for the specific spaces in that comparison.

Comparative Space centroid analysis

The comparison of the space centroids performed for these experiments was limited to the analysis of the full space term and document centroids that were

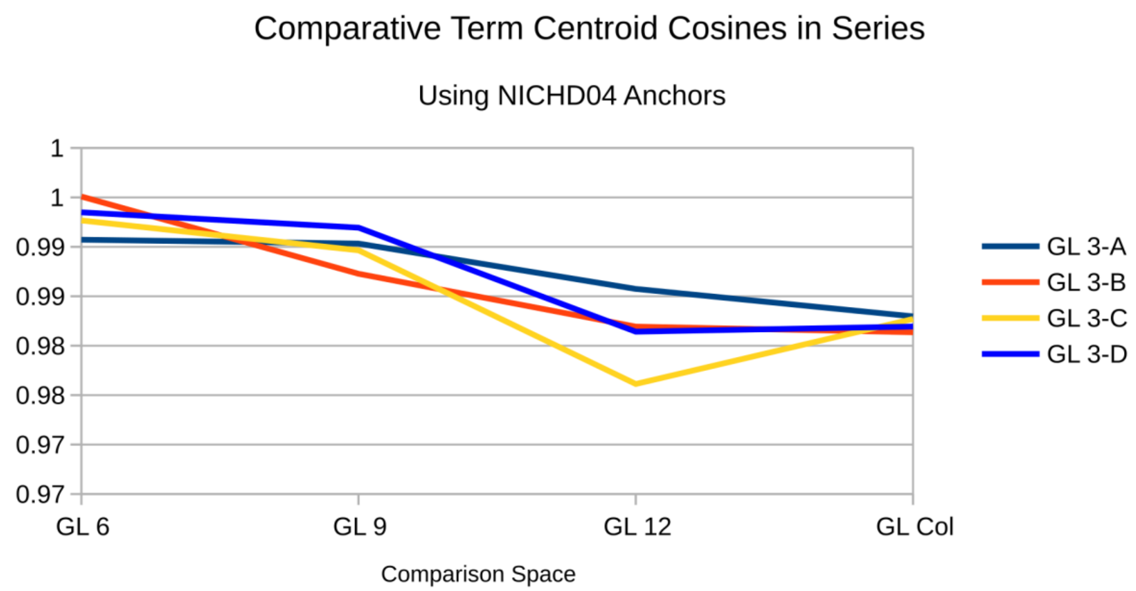

Figure 4.7 - Term Centroid cosine comparison between Grade Level 3 spaces and the other Grade Level spaces in Series A through D using the NICHD04 anchors

computed as part of the direct measures described in Section 4.3.1 for each of the spaces being compared. Using the process detailed in Section 3.3.2, the centroids from S_1 were mapped into S_2 by applying the transform Q to each centroid vector and the resulting remapped centroids $\widehat{C_1}$ were compared to their respective counterparts in S_2, C_2, using the cosine similarity measure. This analysis was performed using each of the available anchor sets to produce the rotation matrix. The comparative space centroid cosines were recorded for each pair of spaces that was examined and were correlated with the other measures for analysis. Figure 4.7 shows the relationship of the comparative term vector cosines for each of the grade level series with that series GL 3 space. Selected details for the individual measures are included in the Appendix in Table A - 4 and Table A - 5.

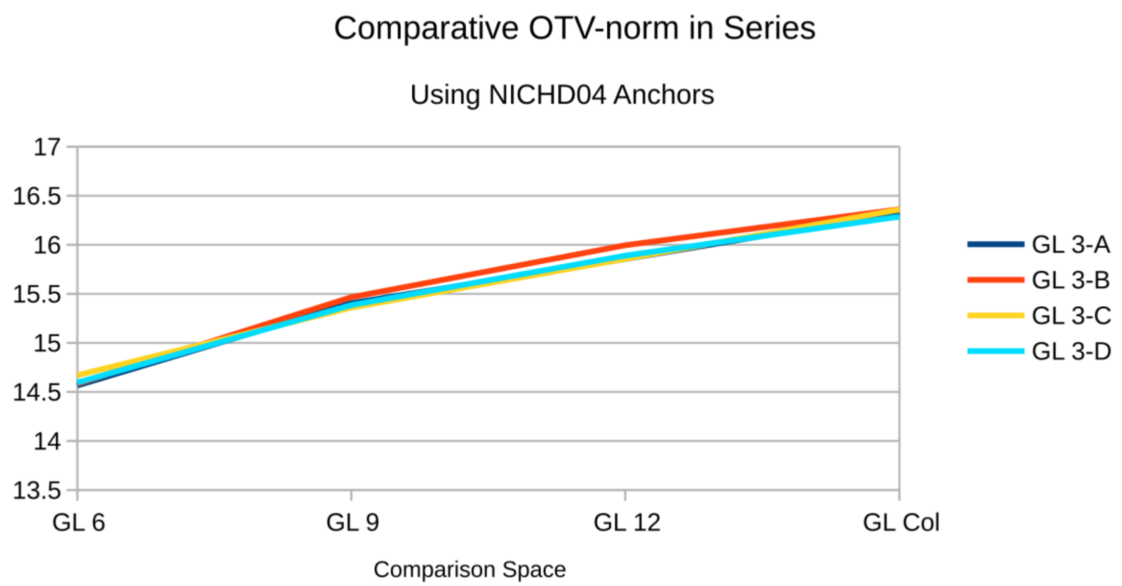

Figure 4.8 – Comparative Term Vector F-Norm for the Grade Level 3 spaces compared to the other Grade Level spaces within their respective series using the NICHD04 anchors

Comparative Term Vector analysis

The comparative term vector analysis is computed as a single measure on the difference between the sets of overlapping term vectors between S_1 and S_2 as described in Section 3.3.3. After identifying the overlapping term set between the two spaces, the transform Q was simply applied to each of the overlapping term vectors from S_1 and then the difference was computed between them and their corresponding vectors in S_2 to produce the difference vectors $\hat{T}$. The F-norm of $\hat{T}$, $\|\hat{T}\|_F$, was then calculated by taking the square root of the sum of the squared vector components to yield the comparative term vector measurement between S_1 and S_2, which we will call the Overlapping Term Vector Norm or OTV-Norm. The OTV-Norm measurements were recorded for the pair of spaces, along with the term overlap ratio described in Section 4.3.1 and the term hit-rate described in Section 4.3.2 for the projected anchors within each of the two spaces. This analysis was performed using each of the available anchor sets for the production of the rotation matrix. These results were then compared across the group of grade level series spaces both within each series, as shown in Figure 4.8, and longitudinally across the grade level series as well as across the tested anchor sets to evaluate performance of the measure. The measures were also computed between the RTRC spaces and each of the grade level series spaces. Selected measures for all of the compared spaces are included in the Appendix in Table A - 4 and Table A - 5.

CHAPTER 5

OBSERVATIONS AND DISCUSSION

Reviewing the results from the space comparison experiments provided several different data points to be considered. In all, 42 different LSA spaces were compared in 592 total pairings that were each related across four different sets of projection items in six separate experiments. Measurements were collected for each space individually and as part of the space-to-space comparisons. A number of observations were made about the individual spaces and the individual spaces measures, as well as the data from the comparison experiments and the performance of the specific metrics being evaluated.

5.1 Observations about the Spaces

There were 15 specific measures captured for each individual LSA space after they were constructed. Aside from the basic counts of documents, terms, and total term instances, the individual space distribution measures showed little variability regardless of the size or content of the space. This was unexpected given the wide assortment of content that was used to build the spaces. These additional individual space measures included the average, maximum, and minimum cosines as well as the standard deviation of those cosines for all of the documents to the document centroid, the terms to the term centroid, and the document-to-document pairs within each space.

Except for the small 1,000 document test space, the RTRC spaces all exhibited nearly identical average cosines to their respective term and document centroids, and the differences between their document-to-document cosine distributions were very small as well (see Table 5.1). The Grade Level series spaces all demonstrated similar characteristics as well, with very slight variability in their document and term centroid average cosines and their average document-to-document cosine which can be seen in Figure 5.1. These space distribution

Table 5.1 – Average cosines for document and term items to their respective centroids and the average of all pairwise document-to-document cosines for the RTRC spaces that were analyzed along with the standard deviation of each

	Space			
	GCat-150k	**GCat-150k-B**	**GCat**	**GCat-1k**
Doc Centroid Avg. Cosine	0.037055	0.372007	0.372702	0.305389
DC Std. Dev.	0.099875	0.100081	0.101144	0.100406
Term Centroid Avg. Cosine	0.102212	0.102214	0.100769	0.115367
TC Std. Dev.	0.083612	0.084662	0.082230	0.105327
Doc-to-Doc Avg. Cosine	0.137934	0.139012	0.139532	0.093876
DD Std. Dev.	0.088118	0.088180	0.088513	0.082031

measures correlated strongly (0.965 for document centroid average cosine, 0.866 for term centroid average cosine, and 0.966 for document-to-document average cosine) with the number of documents used to construct the space, but possessed a total variation of less than 0.022 for any of the three measures. Similar correlation of the average cosines to the content volume was also observed for the RTRC spaces when considered in a group by themselves. Taken as a whole group, however, the distribution measures did not serve as a consistent indicator of content volume for the spaces.

For the group of grade level spaces, differences in grade level, apart from the document count, did not appear to have a notable or consistent effect on the individual space distribution measures. Both the large 6th grade and 9th grade spaces measured more like the similarly sized 12th grade and College level spaces than they did to their 6th and 9th grade counterparts, but again the total difference between the distribution measures for this group was extremely small.

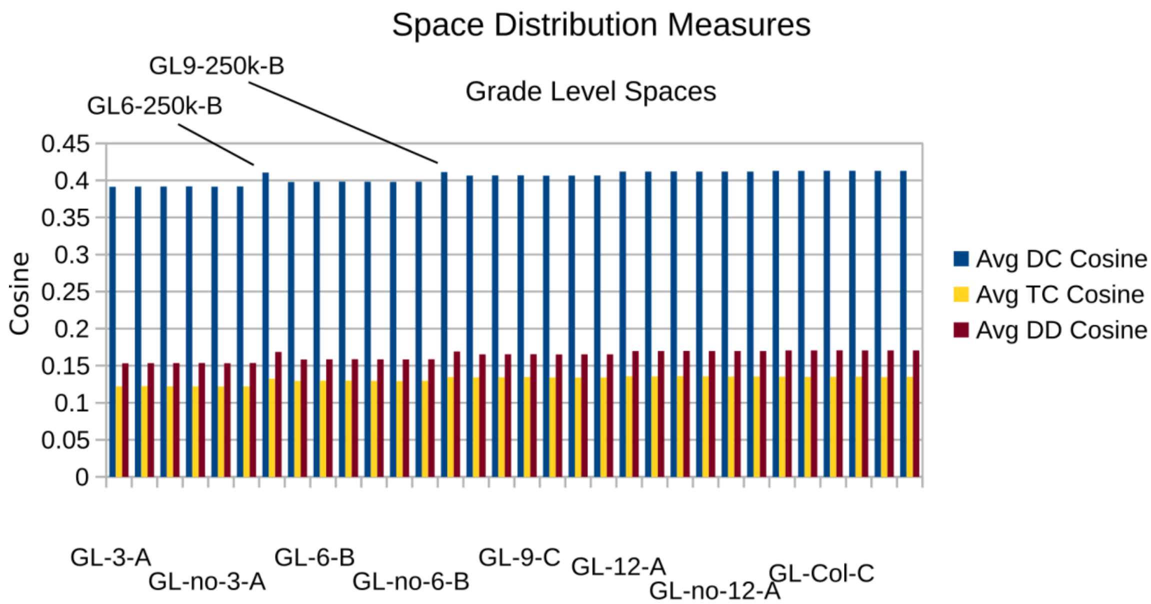

Figure 5.1 – Grade Level Series individual space measures for average cosines to the space document centroid, term centroid, and average document-to-document cosine for each space

Because the distribution measures for the individual spaces showed so little differentiation between the spaces, further analysis of them in comparisons between spaces was discontinued.

5.2 Observations from the Experiments

The six different experiment sets were performed to verify the concept that the meaning embodied within spaces was different, that this difference was related to qualities of the content used to build the spaces, and that certain measures would be useful in describing the degree of difference between two spaces. Each of the experiment sets were designed to address particular questions related to these objectives.

The results did indicate significant difference between the spaces in the way they related projected items to each other which is the operational explanation of meaning that was being sought. Those differences were demonstrated to be dependent on the quality of the content being used rather than simply its volume, or due to chance.

5.2.1 The Control Experiment

Comparison of the GCat space, formed from the 234,869 documents in the Government/Social top level category of the Reuters news articles collection, with a slightly modified version of the same document set, GCat-Mod, was performed as a control experiment to verify that the measures being studied were not overly sensitive to minor changes in the document collection. The GCat-Mod space differed from the original GCat space only in the removal of 10 terms from the collection vocabulary. This resulted in the net reduction of 50 term instances across the dataset amounting to less than 0.00012% of the total term instances. The document count remained unchanged, so the document overlap ratio between the two spaces was 1.0. With the ten missing terms the term overlap ratio between GCat and GCat-Mod was 0.999927.

The measures collected for the GCat to GCat-Mod pairing showed very small amounts of comparative change between the spaces, which was as expected. The measurements also varied only slightly across the four comparison sets used, as shown in Figure 5.2. The total three-tuple relationship change percentage (TC%) measured between the spaces amounted to less than 0.003% of the total relationships in the space using any of the four projection sets. Likewise, the F-Norm of the Overlapping Term Vector differences (OTV-Norm) measured consistently 0.0047 for all four of the projection sets used as anchors. The comparative space centroid analysis showed the remapped term and document centroids from one space to the other to have similarity cosines of 1.0, indicating that they were mapping identically. The OTV-Norm was a very small 0.0047 for all four projection sets. The only notable difference in the performance of the projection sets was in the projection distribution measures

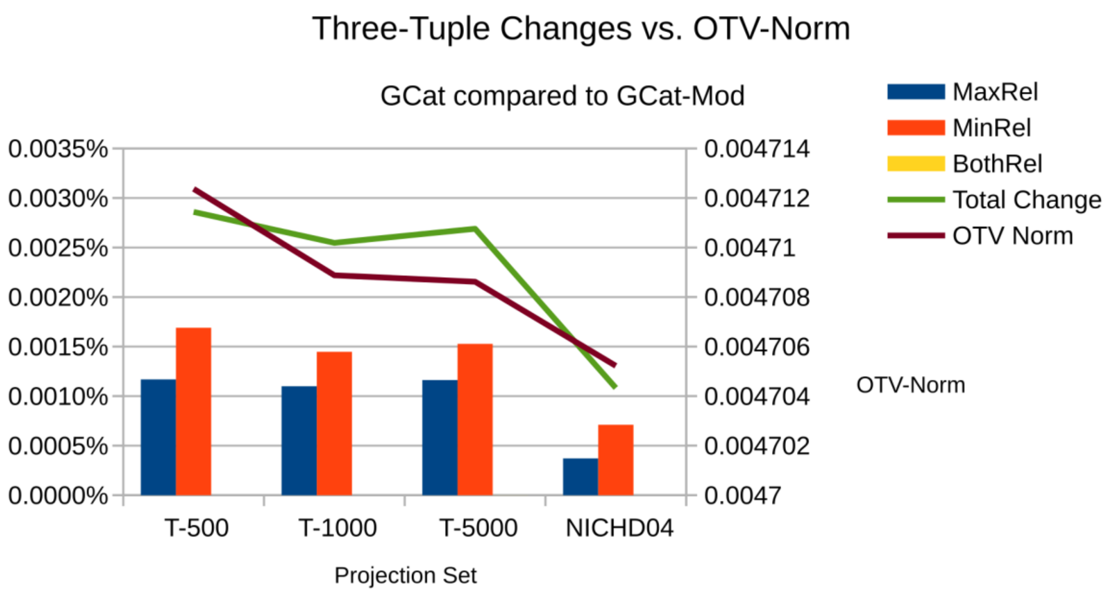

Figure 5.2 – Comparison measurements from the control experiment showing the tuple-change percentages and OTV norm obtained using the different projection sets and anchor sets

themselves, where the NICHD04 set showed slightly lower centroid cosines and larger standard deviation numbers than the other three sets as seen in Table 5.2. These measures were consistent across both spaces, however. Rotational measures were computed both by mapping the GCat-Mod space into the GCat space, and in the reverse direction, mapping GCat into GCat-Mod. There were no observable differences in the results using the two rotation directions.

5.2.2 The General Experiment

The second experiment set was a more general test using four different spaces constructed from the RTRC collection (see Section 4.1.1). Two of the sets contained 150,000 documents each, and of the remaining sets one was significantly larger (234,869 documents) and the other significantly smaller (1,000 documents). The small space, GCat-1k, was used primarily to test initial comparison programs and produces widely different comparison measures from

Table 5.2 – Differences in the Projection Set Centroid distributions for the RTRC-GCat space (S_1) compared to the RTRC-GCat-Mod Space (S_2)

		Projection Centroid in S_1	Projection Centroid in S_2
T-500	**Cosine**	0.703681	0.703683
	Std. Dev.	0.120599	0.120602
T-1000	**Cosine**	0.715860	0.715864
	Std. Dev.	0.128651	0.128653
T-5000	**Cosine**	0.713321	0.713326
	Std. Dev.	0.120797	0.120798
NICHD04	**Cosine**	0.694126	0.694127
	Std. Dev.	0.178146	0.178146

the other spaces in all categories. While the other spaces all exhibit total three-tuple relationship changes in the range of 10-16% across the four projection sets, GCat-1k has total changes in the 73-75% range. The term hit-rate for all of the projection sets used in this experiment was over 99% for each of the spaces tested except for GCat-1k, but even with its small amount of content all four projection sets maintained a term hit-rate above 94.7%. These differences between GCat-1k and the other GCat spaces parallel most of the measures collected in this experiment.

Since the two 150,000 document GCat spaces were selected from the set used to create the full Gcat space, they both had an identical document overlap ratio of 0.638654 between them and the full GCat space. The document overlap ratio between the two of them was measured at 0.566727. Among these three spaces, the GCat-150k-B space and the larger GCat space were measured as most similar regardless of the projection set used. These two sets had the highest term-overlap ratio (0.8278), as well as the lowest total three-tuple relationship change rate (TC%), which was in the 10.1% to 10.6% range, and OTV-Norm measurements ranging from 12.45 to 12.63, as shown in Figure 5.3.

All three space comparisons revealed a total three-tuple relationship change rate exceeding ten percent, as can be seen in Figure 5.3. This change demonstrated that different LSA spaces do exhibit a measurable degree of difference in the way they map content items that are projected into their framework of meaning even though they were constructed with items taken from the same pool of content.

5.2.3 Grade Level Series Experiment

The Grade Level series experiment was intended to investigate the semantic mapping relationships between spaces representing a series of content evolving

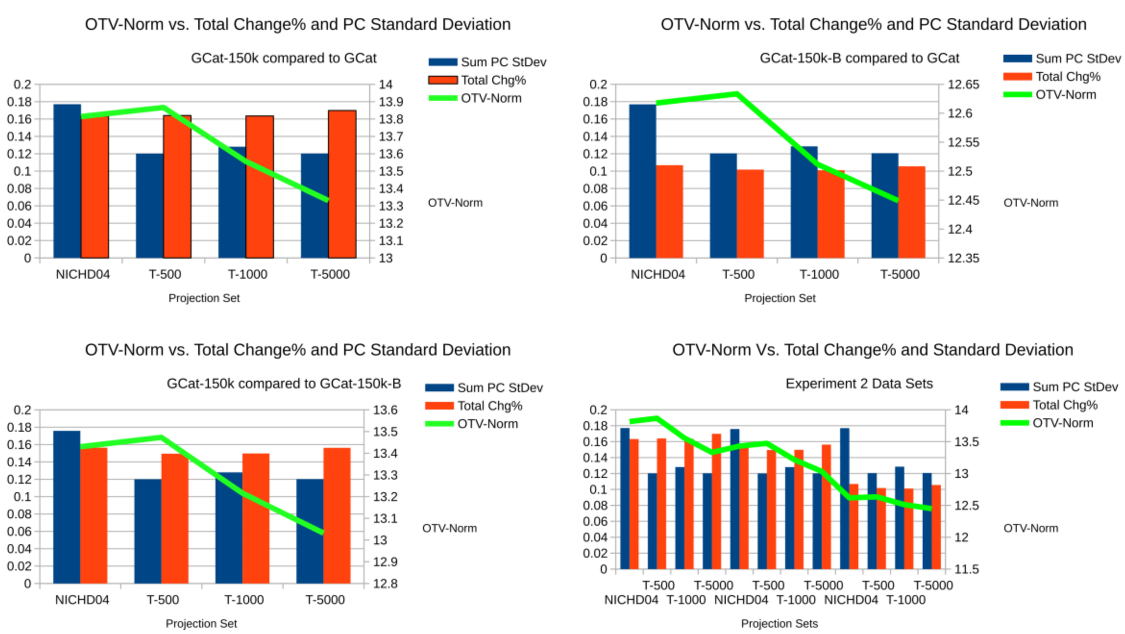

Figure 5.3 - General experiment measures for individual RTRC spaces and combined view of the results for all three spaces (lower right)

in its reading complexity from elementary texts through college level material. Spaces were compared vertically within each series to examine the change in sematic mapping as the complexity of the material increased. Longitudinal comparisons across the grade level series at the same grade level were performed to test the theory that the semantic mappings would exhibit less change when constructed from sets of content of similar reading levels.

Throughout all of the comparisons none of the projection sets had a term hit rate (percentage of mappable terms) lower than 96.6%. The term overlap ratio, however varied from 0.23 to 0.73 among the different spaces being compared.

Vertical comparisons

Each LSA space of the four series of grade level spaces was compared to the higher level spaces within the same series. All four series of spaces performed similarly in terms of their observed measurements. For each given space in a series, the document overlap ratio between spaces steadily decreased as the comparison space was increased in grade level (Figure 5.4). The term overlap ratio decreased similarly as well. The amount of difference indicated by the TC% comparisons increased as the grade level distance between the compared

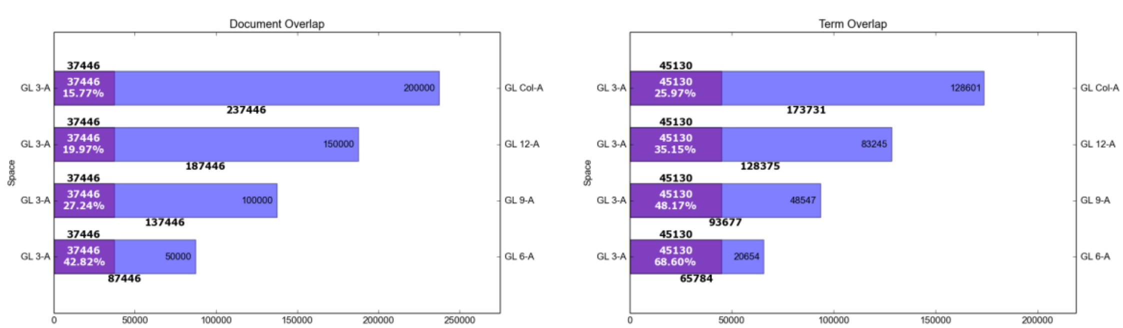

Figure 5.4 - Document overlap ratios (left) and term overlap ratios (right) for the GL-3A space to other spaces within series A

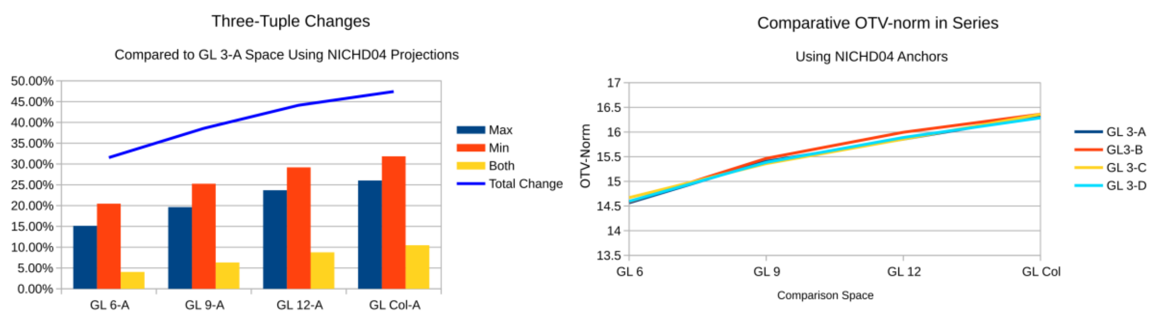

Figure 5.5 – Three-tuple changes and OTV-Norm vertical Grade Level series comparisons

spaces increased (See Figure 5.5). The TC% measurement correlated strongly, 0.88 or better, with the difference in grade level between the spaces across all four of the projection sets tested.

The OTV-Norm exhibited a high degree of correlation, 0.94 or better, with the total three-tuple change percentage across all projection sets in the vertical series comparisons when each projection set was viewed individually and nearly as strong a correlation when all projection sets were viewed together (0.86).

These vertical grade level comparisons successfully demonstrated the presence of significant semantic differences between spaces related to the inclusion of content of different complexity levels. The fact that the measurements correlated to the amount of difference in grade level indicates the influence of the quality of the content items has on the meaning represented in a LSA space.

Longitudinal Comparisons

Examining the collected measurements across the series of grade level spaces, there was again very little difference in the observed trends between the series. The spaces had very similar document overlap ratios across the series even though the document content for each was randomly chosen. This might be attributed to the limited amount of content available for selection at the elementary grade levels below 3rd grade which constitute the bulk of the overlapping items. The document overlap ratio for each set peaked at approximately 0.335 for each set when a space is compared across series with its corresponding spaces at the same grade level. Figure 5.6 illustrates the document overlap comparison across series for the spaces in GL Series A with the spaces in the other series. The pattern demonstrated in the Series A spaces was repeated for the remaining grade level series with very similar values.

Term overlap between the series behaved similarly to the document overlap with the ratio reaching as high as .70 for spaces compared across series with other spaces at the same grade level. The term overlap ratio correlates very strongly, 0.98, with the grade level difference in the spaces being compared.

Three-tuple changes across the series were very similar at grade level across the four series as depicted in Figure 5.7. In general a wider range of difference was observable when examining the three-tuple relationship changes between spaces at the lower grade levels, with TC% rates in the 25%-55% range. These differences decreased progressively at each higher grade level until the college level spaces were exhibiting TC% rates in the 18%-20% range. Overall the TC% measure varied from 18.40% to 54.07% across all four series when compared with spaces at different grade levels. There was a strong correlation, 0.89-0.95, between the OTV-Normand the difference in grade level between the spaces being compared. Overall, the OTV-Norm tracked with the total three-tuple change rate for the spaces with a degree of correlation of 0.92 or better for all four projection sets used as anchors, ranging from 12.50 to 16.62 across the projection sets.

5.2.4 The Large Volume Experiment

Given the varying sizes of content used to produce the spaces in the grade level series with the higher grade level spaces containing more documents, the next

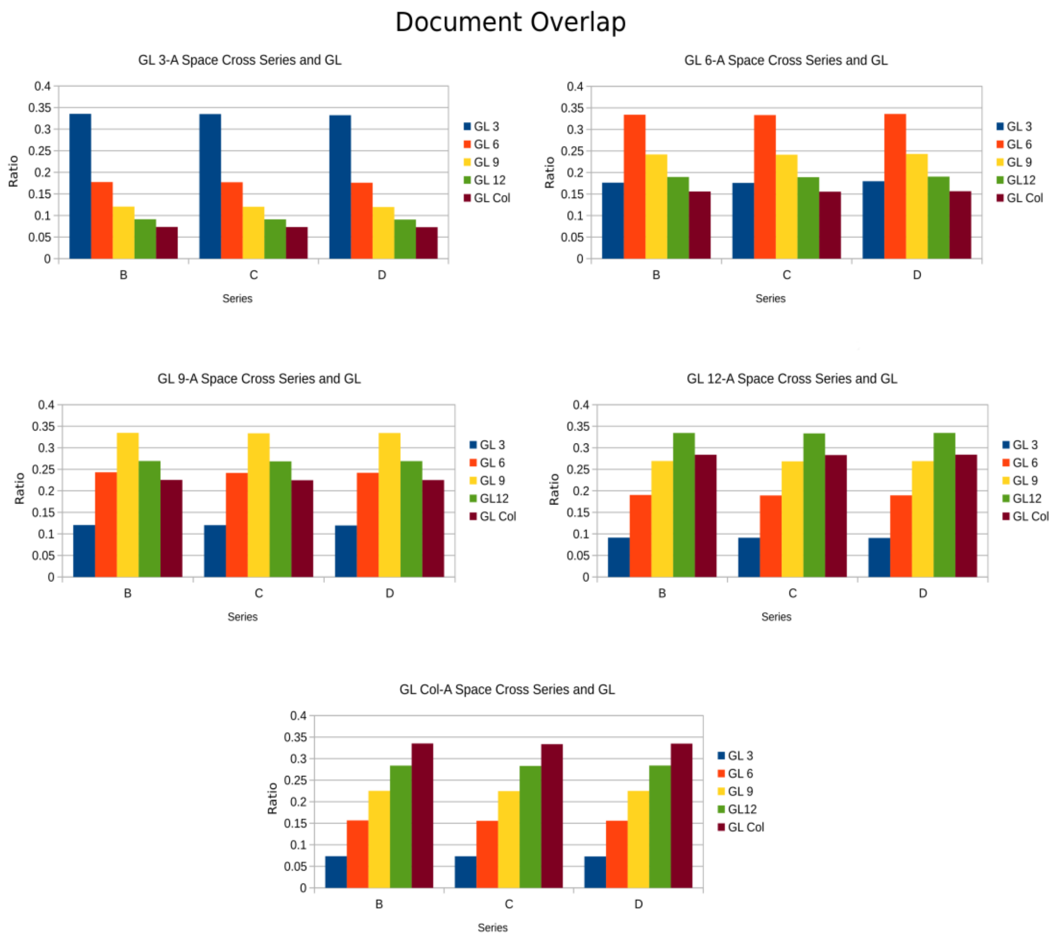

Figure 5.6 – Individual cross series document overlap ratios for each of the Grade Level spaces in Series A with all the grade level spaces from the other series used in the experiments

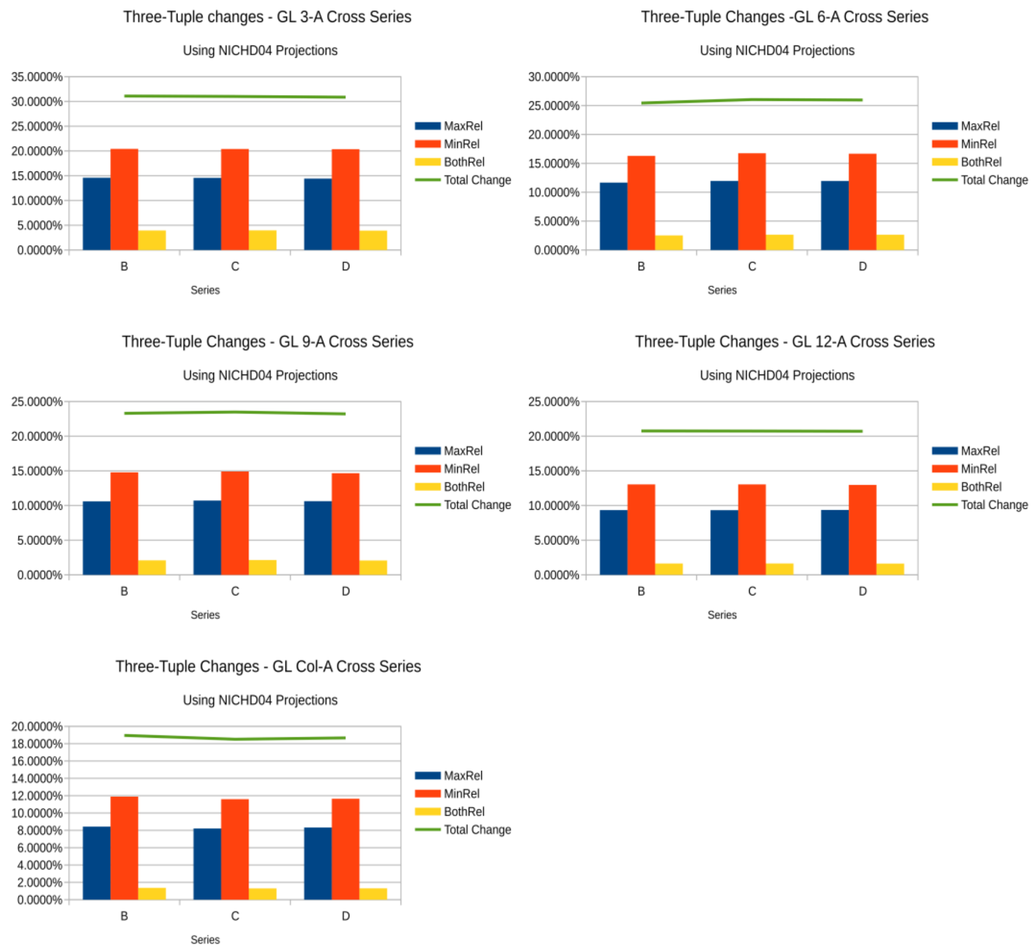

Figure 5.7 – Three-tuple changes for the Series A Grade Level spaces compared longitudinally across series to their Grade Level equivalent spaces in the other series

experiment was designed to explore the impact of the volume of content, the total number of documents, on the measurement results where the complexity of the content was held constant. This was required to demonstrate that the changes in the spaces observed in the vertical comparisons in the Grade Level Series experiment were not simply attributable to the volume of content included in the spaces rather than the quality of the content. Larger spaces filled with intermediate grade level material were constructed to represent a 6^{th} grade level (L6) and a 9^{th} grade level (L9) of complexity at high volume, with the spaces containing approximately the same amount of content as the college level spaces from the Grade Level (GL) series. These large spaces were then compared against the various GL series spaces.

The document overlap observed between these large spaces and the GL series spaces was much lower than in the grade level series experiment, being in the range 0.07-0.16 for comparisons with the L6 space and 0.07-0.22 with the L9 space. As with the longitudinal comparisons in the GL series, the document overlap ratio was highest when comparing the L6 space to the other 6^{th} grade spaces, and likewise for the L9 space when compared to the other 9^{th} grade spaces.

Term overlap ratio measurements were much lower in this experiment than in the original GL series comparisons, ranging from 0.35-0.53 for the L6 space when compared to the spaces in the grade level series, and 0.32-0.58 when the L9 space was used in the comparisons. The term overlap ratio measurements for the L6 space were highest when compared to the 9^{th} grade spaces in the GL series. For the L9 space the term overlap ratio was highest when compared with other 9^{th} grade spaces.

The measured TC% ranges comparing the L6 and L9 spaces to the GL series ranged from 23.47% to 44.45% across the four projection sets with the lowest degree of difference seen between the spaces at the same or next highest grade

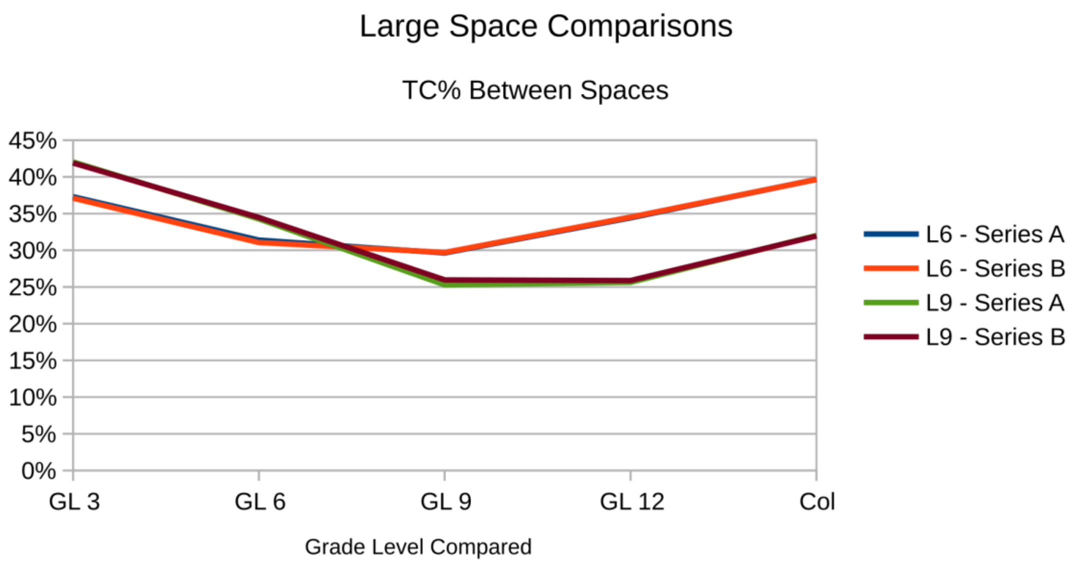

Figure 5.8 - Observed TC% values for the L6 and L9 spaces compared to the Grade Level spaces from Series A and B

level (see Figure 5.8). The change in TC% showed a moderate correlation with the difference in grade level. Interestingly, if the direction of the grade level difference was included (absolute value not used) the correlation of the grade level difference with the TC% disappeared, but the correlation of the grade level difference with the OTV-Norm increased to above 0.5. The lowest TC% measurements (31% or below) were observed with comparisons of the large spaces to GL series spaces at the same grade level or to spaces one step up in the series (L6 to GL9 or L9 to GL12). The comparison of the L9 to GL9 or GL12 spaces yielded TC% measurements below 26% for all four projection sets. These measurements are similar to those observed in the GL Series experiment.

Examining the OTV-Norm for both the L6 and L9 comparisons to the GL series revealed a strong correlation with the measured TC% of 0.81 or higher across all four projection sets. With each space considered separately, however, the correlations differed for comparisons against the L6 space, ranging from 0.53 to

0.65. The OTV-Norm correlated much more strongly with the TC% for the L9 space comparisons separately with the value ranging from 0.89 to 0.96.

In all cases, the lowest term overlap ratios and higher TC% measurements were seen when comparing the L9 space down to the GL 3 spaces.

5.2.5 Non-overlapping Series Experiment

To answer the question of how much impact document overlap was having on the difference measurements, a set of non-overlapping grade level spaces was constructed. These consisted of two series of spaces similar to the grade level series used in the previous experiments, with the condition that no document was allowed to be selected for inclusion in both series. Comparisons were made between these series of spaces both vertically within each series, and longitudinally across series, similar to the regular Grade Level experiment.

Vertical comparisons

Performance in vertical comparisons for the non-overlapping series was very similar to the observations in the vertical comparisons for regular Grade Level Series spaces. The document overlap ratio for each series ranged from 0.157781 to 0.789445. Both series exhibited the same document overlap ratios between spaces in the same steps due to the way the spaces were constructed. Just as with the regular GL series, the document overlap ratio was higher with the smaller spaces at the lower grade levels and diminished as the grade level and space size increased. The term-overlap ratio measurements varied in a similar pattern from 0.25896 to 0.738626, although the values did differ between the series. In terms of the TC% measurements, both series measured similarly ranging from a low of 21.96% to 47.24%, correlating strongly (0.88 or higher) with the difference in grade level between the spaces being compared. The OTV-Norm measured between spaces ranged from 12.7 to 16.28 for all projection sets. These measurements correlated with the observed TC% measurements across all four projection sets with a value of 0.95 or greater. These measurements also correlated strongly, 0.81 or better, with the grade level differences between the spaces being compared.

Longitudinal comparisons

Document overlap was virtually eliminated between the series of spaces used for this experiment. A small number of documents with the same term frequency vector were again identified between the series, but this amounted to less than 0.002 document overlap ratio in the largest case.

The term overlap ratio for the non-overlapping sets was not as high as that seen in the regular GL series experiment, averaging 0.4 and falling in the range 0.22 to 0.52 across all the comparisons, as opposed to the average or 0.5 and range of 0.23 to 0.73 from the regular GL series comparisons. However, just as in the regular GL series, the term overlap ratio measurements for the non-overlapping series did show a strong correlation of 0.98 to the grade level difference (absolute value) between the spaces being compared.

Mirroring the observations in the regular GL Series experiment, the TC% measurements across the non-overlapping series were consistently in the range of 23.44% to 54.45% for all four projection sets tested. These measurements correlated almost perfectly (0.99) with the difference in grade level between the spaces being compared. The bottom end of these measures was slightly higher than that of the GL series comparisons, which was 18.4%, but matched the top end of the TC range seen in those tests measured at 54.07%. Likewise, the measured OTV-norm ranged from 13.1 to 16.74 across the projection sets which matched the range seen in the regular GL experiment. These measures correlated to the observed TC% at 0.87 or better regardless of the anchor set used in the rotation. Overall, the elimination of the document overlap between the series did not appear to alter the overall pattern of measurements that had been seen in the regular GL Series experiment or in the Large Volume experiment.

5.2.6 Frozen Vocabulary Experiment

Finally, to address the issue of term overlap on the remaining measures, a set of comparisons was performed on a series of spaces that were constructed using a controlled vocabulary. The set of terms was frozen after an initial base GL3 space was created. This space was then used as a foundation for building two series of spaces at the 6th and 12th grade levels. These two series used the same vocabulary of 33,069 words that were present in the initial GL3 space. Additional documents selected for the series were only allowed to be included in one series, so the new content being added was virtually non-overlapping. Again as in the larger non-overlapping series construction, some documents were found that resulted in the same term frequency vector, resulting in some of the added documents effectively overlapping between the series.

The base set of 28,270 documents which were used to establish the fixed vocabulary was included in all of the spaces, accounting for practically all of the document overlap between the series. The resulting document-overlap ratios for the cross series comparison is shown in Table 5.3 for both the vertical and

Table 5.3 - Document overlap ratio between the Frozen Vocabulary series of spaces

	Base Space	**6A Space**	**12A Space**
6A Space	0.220398	1.0	0.561918
12A Space	0.123846	0.561918	1.0
6B Space	0.220391	0.137453	0.095243
12B Space	0.123843	0.095233	0.073381

longitudinal comparisons in the series. Since the vocabulary was fixed, the term overlap ratio between all spaces in the experiment was 1.0.

The measured TC% for this experiment ranged from 20.95% to 48.9% across the three projection sets tested. This experiment was processed using the three projection sets (NICHD04, T-500, and T-1000) omitting the larger T-5000 projection set. This observed range is similar to the range seen in both the non-overlapping series experiments and in the regular GL series tests. Individual comparison measurements between the Frozen Vocabulary spaces were likewise a few percentage points lower than the TC% measures seen in the previous experiments but followed the same pattern where same grade space comparisons showed lower TC% measurements. The OTV-Norm measurements also exhibited very similar results, ranging from 12.79 to 16.89 and correlating strongly, 0.96 or better, with the TC% measures across all three projection sets. Term overlap being identical between spaces did not strongly affect the TC% or OTV-Norm measurements being observed.

5.3 Observations about the Measures

Three groups of measurements were captured for each space to space comparison performed. These were the direct measures, the projected content measures, and the rotation-based measures. Each group had measures that provided some insight into the differences between the spaces being examined.

Of the direct measures, comparative examination of the individual space metrics such as content volume, vocabulary size, or the more complicated distribution analysis did not yield much insight into the relative difference between two spaces. Looking at the individual space metrics gave no indication of the degree of difference that was present between two spaces as evidenced by all of the experiments that were performed. Certainly having a sufficient volume of content in the space matters in the quality of the space, but similarly-sized spaces are not necessarily any more alike than others that are of notably different size as demonstrated in the Large Volume experiment.

Document overlap also does not appear to be a particularly useful measure on its own. While 100% document overlap could only be the case if two spaces were identical, the experiments showed that the lack of document overlap between two spaces is not an indicator of dissimilarity of the relational mappings contained within the spaces. The RTRC spaces had absolutely no document overlap with any of the grade level spaces and yet indicated less than 70% change in the three-tuple relationships when compared to any of the grade level spaces outside of the detailed experiments.

The term-overlap ratio measurement, however, provided very useful information on the difference between two spaces. Term overlap is a strong indicator of the degree of similarity between two spaces primarily because of the way the spaces represent meaning is by defining term vectors that map within the space. If a term is not present in both spaces, its meaning cannot be represented in both of them or compared between them. Term overlap does not tell the whole story by itself as shown in the Frozen Vocabulary experiment where all of the spaces shared 100% of their terms and significant differences were still observed.

The projection-based evaluation of the three-tuple relationship changes, the TC% measurement, supplied direct evidence of internal relational changes for differentiating spaces. Examining the three-tuple relationship changes appeared to be the most easily understood measure of difference between two spaces. However, these measures are expensive computationally and may be more susceptible to differences between the content of the projection sets than other measures. Another measure in this group, distribution of the projection vectors within the spaces being compared showed some moderate indication of degree of difference, but these measures are not easily quantified or compared. It is perhaps more of an indicator of how well the projection set will do in demonstrating the difference. Wider distribution of the projection items, as indicated by a larger standard deviation, seemed prone to show more change than narrower distribution. Further research in this area is required.

Three rotation-based measures were collected for all the experiments. The first two were the cosine similarity measures of the rotated term centroid and document centroid from one space into the other. The results from these centroid comparison measures provided surprisingly little insight into the degree of difference between two spaces. In general the range of values observed for these measures across many of the space comparisons was very small, often varying by as little as 0.03 across entire series of the Grade Level spaces being compare where notable change in the TC% values and OT-Norm were observed. In the few cases where a wider degree of difference was seen in centroid comparison measures, their values did not correlate with the other difference measures, such as the TC% value or the OTV-Norm.

The third rotational measure, the OTV-Norm was a strong indicator of the degree of difference between two spaces. This measure correlates strongly, 0.9346 or better, with the observed TC% measure across all experiment comparisons and all projection sets (see Figure 5.9). This measure is definitely affected by the term overlap ratio between two spaces being compared. Terms that are not present in both of the spaces cannot be used to construct documents or define meaning in both spaces and are not part of the OTV-Norm calculation. The higher the degree of term overlap, the more term vectors are available as input for computing the OTV-Norm which in turn results in a higher the portion of the spaces being reflected in the value. Differences in the anchor sets used to compute the rotation transform upon which the OTV-Norm computation is based also affect the resulting value of the measure, but the observed variation is small across the four anchor sets tested. All of the anchor sets produced OTV-Norm

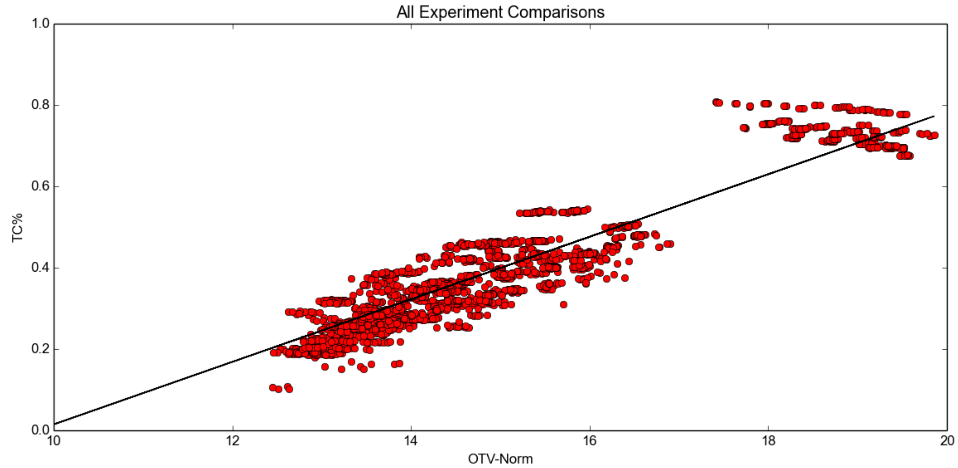

Figure 5.9 - OTV-Norm values plotted against the TC% observations from all the experiments

measurements that had a high degree of correlation with the TC% measurements using any of the projection sets.

Further examination of the variance in the OTV-Norm revealed an interesting trend in the OTV-Norm values related to the observed total change percentage for the projection set and the standard deviations of the projection centroid cosines for the projection set. This can be seen in Figure 5.3 where the results from the General Experiment are shown for the three RTRC spaces. For each comparison of spaces in these experiments there were two projection item centroids produced, one in each space. As illustrated in the figure, the sum of the standard deviation of projection item cosines around the centroid for each of the two spaces (Sum PC StDev) is shown in relation to the OTV-Norm and to the TC% measure observed for a given projection set. Notably, the lower values of the OTV-Norm coincided with lower values for the standard deviation sum and an increase in the size of the projection set used as anchors. The number of anchors and the standard deviation of those anchors from their projection centroids in each of the sets being compared seems to have some impact on the value of the OTV-norm produced using those anchors, albeit not large. This relationship appeared to be present to some degree in all of the experiments that were conducted. The mechanics of this relationship and how it bears on the value of the OTV-Norm or its accuracy are a subject for further research.

5.4 Computational Performance

These experiments were conducted with the primary goal of assessing the usefulness of the values generated by the measurements being examined. Computational performance of the measurement calculations was not timed under controlled benchmarking conditions for the tests that were run. General expectations based on the algorithmic complexities of each measure discussed in Chapter 3 were that the three-tuple relationship comparisons performed as part of the projection based measure would prove to be cost prohibitive in their computational demands as the number of projection items was increased. This indeed was the case for the larger T-5000 projection set. Wall-clock timings of the three-tuple relationship comparisons to produce the TC% value using the T-5000 projection set almost exceeded 40 minutes while the computation of the OTV-Norm across an overlapping set of term vectors that was five times larger completed in just over one minute using the same 5,000 items as anchors to compare the same pair of spaces[1]. Similar differences in the wall clock timings were seen across the entire experiment set when using the larger projection set. Smaller projection sets were processed for three-tuple relationship comparisons in shorter times more comparable to those seen in the computation of the OTV-Norm as processing the T-1000 set was completed in roughly the same wall-clock time for either method.

[1] These timings were performed using a single-thread process for all computations on an AMD FX8150 – an 8 core 64 bit processor running at 3.6GHz with 16MB total on chip cache. The system also had an available 32GB of physical memory and had a minimal processing load.

5.5 OTV-Norm and Term Overlap Ratio

A strong relationship between the term overlap ratio (TOR) and the OTV-Norm measurements can be recognized from the results of all six experiments. These two measurements both exhibit a strong correlation with the total number of three-tuple relationship changes observed in the comparison of any two spaces tested. Since the OTV-Norm is calculated using the identified overlapping terms from the two spaces being compared, the TOR also provides information about the portion of the space reflected in the OTV-Norm value. Taken together, the two measures are a good candidate for a metric describing the relative semantic difference between two spaces. Computation of both the TOR and OTV-Norm measures are relatively inexpensive compared to the production of the three-tuple relationship analysis across two spaces.

Analyzing the combined results of these measures by performing a linear regression on the two variables of TOR and OTV-Norm yields a simple model for prediction of the total three-tuple change percentage using these two inputs:

$$TC\% \approx -0.207882 + 0.050719(OTVNorm) + -0.339339(TOR)$$

The plot of this model is shown in Figure 5.10, and the comparison of the observed TC% versus the TC% predicted from the model is shown in Figure 5.11. Addition of the term in the equation for term overlap increases the correlation of the predicted TC% values with the observed TC% values to 0.946 versus 0.918 using the OTV-Norm alone. This combination of these data measures is what will be called the Semantic Measurement Model (SMM). Further refinement of this model might be necessary to adjust for the variability in different anchor sets that could be used and perhaps further account for the differences in the TC% measurement observed between different projection sets using the projection centroid standard deviation information mentioned in Section 5.3 and is suggested as a topic for further research.

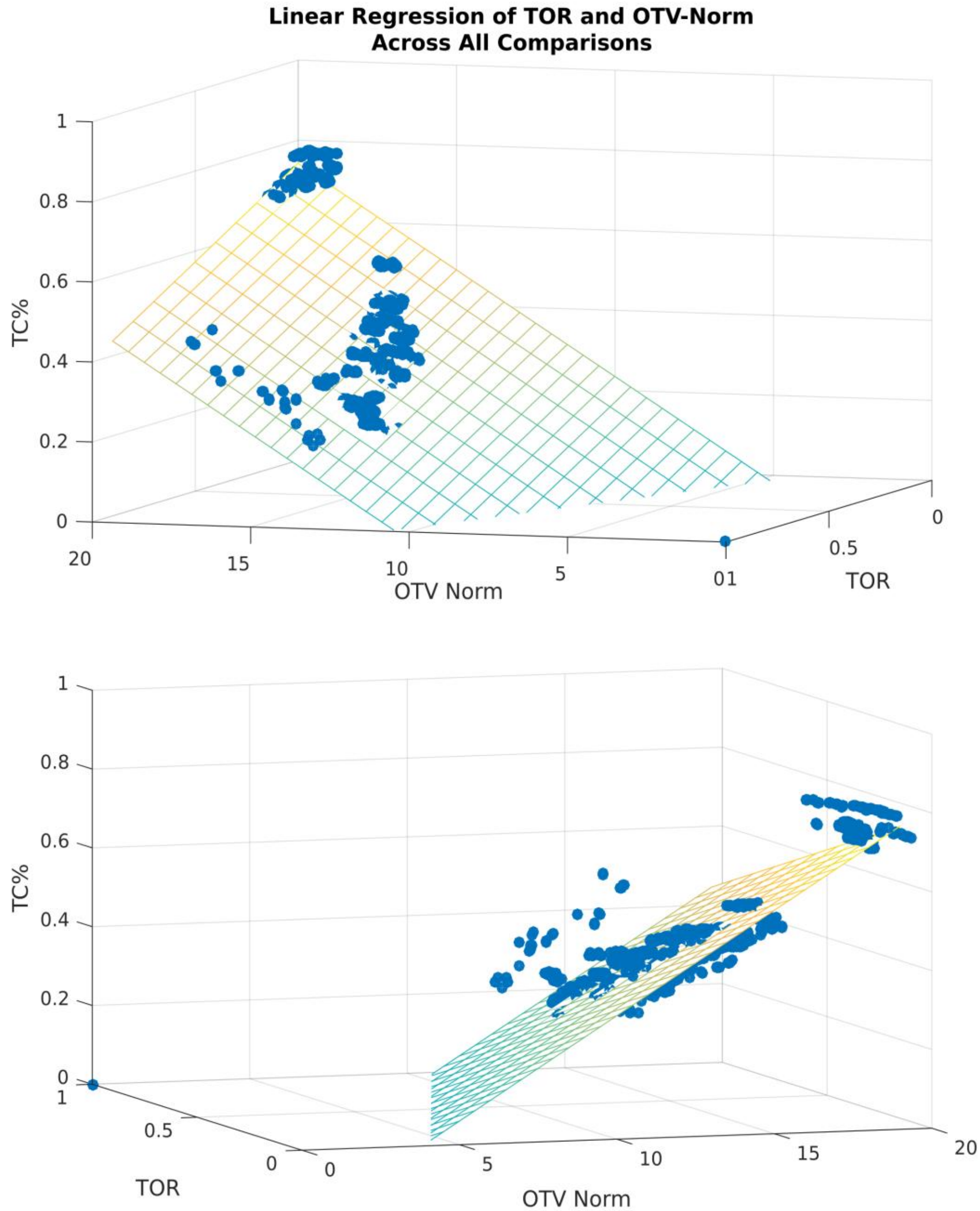

Figure 5.10 - Linear regression model of the TOR and OTV-Norm related to the observed TC% for all comparisons across the experiments

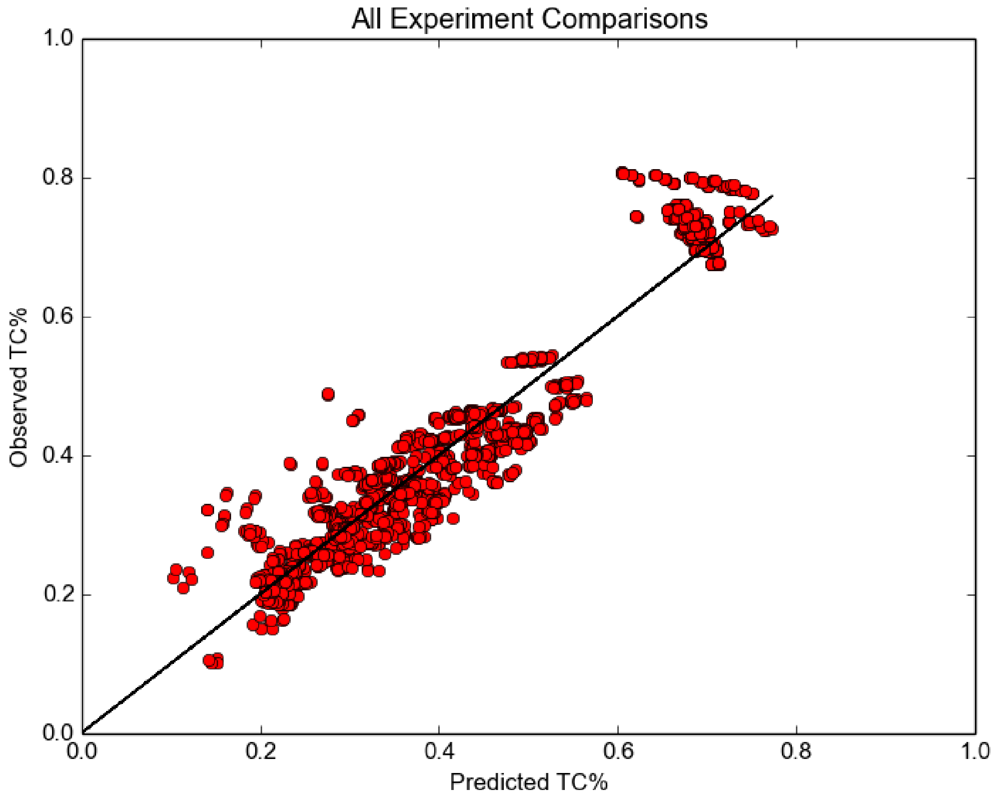

Figure 5.11 – Observed TC% value versus the predicted TC% value using the Semantic Measurement Model combining the OTV-Norm and TOR model for all experiments

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

This work has set about addressing both practical and theoretical questions in the use of Latent Semantic Analysis for natural language processing. LSA is a robust and established technology with a solid theoretical background as discussed in Chapter 2. The work here has focused on improving the usefulness of LSA by providing better understanding of the semantic mapping system that is at its core. The measures put forth in Chapter 3 and the experiments described in Chapters 4 and 5 answered several question about the semantic representation within a hyper-dimensional LSA space and how it relates to the content used in its construction. As is often the case, these findings have also raised new questions for further research, but also provide some tools for addressing them.

6.1 Findings

This research has shown that semantic differences between LSA spaces are observable and measureable, and that these differences are related to the quality of the meanings carried in the input content from which these spaces are built. A new measure was developed, the Overlapping Term Vector Norm (OTV-Norm) described in Section 3.3.3, which was demonstrated to be a viable, easily computed measure for quantifying the difference between two spaces. Further augmenting the information provided by the OTV-Norm with information about the term overlap ratio between the two spaces being compared improved the results obtained from the measure.

In the process of evaluating the several different possible measures enumerated in Chapter 4, a number of items were also identified as providing little useful information for differentiating two spaces. Among these were the individual space distribution measures which showed surprisingly little variation across all of the LSA spaces examined as described in Section 5.1.

Additionally it was demonstrated in the experiments that similarly sized spaces do not necessarily represent the same mapping of meaning and that increasing the volume of content does not necessarily induce change 1n the meanings represented if the content is all of the same quality. It was also shown that document overlap between spaces is not needed for similar meanings to be represented. Related to that fact, it was also shown that while common vocabulary is required for two spaces to provide meaning for the same terms, having complete vocabulary overlap does not necessarily produce the same mapping of meaning for those terms.

6.2 Further Research

The main purpose of this research was to provide tools for further examining the characteristics of LSA spaces and the meaning they contain. With the leverage provided by these objective quantitative measures a number of different questions may now be better investigated. Among those are the issues of selecting the optimal dimensional representation for a given space, the identification of key content items or sets of items that contribute significant meaning to the space, and starting to determine not just how much a space has changed when its content is modified but also how to intentionally modify the content of a space to obtain a desired representation of meaning.

Refinement of the Semantic Measurement Model to adjust for the variability in different anchor sets and perhaps further account for the differences in the TC% measurement observed between different projection sets is a subject for additional future research. Observations during the experiments indicated that the number of anchors used had some influence on the OTV-Norm value that was produced. The observed effects were minor, but it is possible that adjusting the number of anchors and making use of information about the projection centroid standard deviation for those anchor sets could improve the quality or stability of the measure, perhaps be used in providing confidence intervals for the results of the metric.

In the course of this work additional questions have also been raised. One unexpected finding was lack of variation seen in the individual space term and document distributions around their centroids. This characteristic and the observation that the rotated document and term centroids showed so little difference in comparison with spaces that were in many cases quite different brings up questions about the hyper-spatial mapping that is representing the terms and documents in a LSA space. Are these mappings all essentially balanced around some common center? Is this characteristic related to the dimensionality of the space? Can the little motion there is be used to answer other questions about the properties of a space?

Another question remains related to the amount of document overlap between two spaces. The measure of document overlap between LSA spaces was determined to provide little usable information in the discernment of the degree of semantic difference between two spaces at levels of overlap ranging from moderate to none. Other than the Control Experiment which had a document overlap ratio of 1.0, the highest document overlap ratio seen in this work was a ratio of 0.638654 in the General Experiment described in Section 5.2.2. Further investigation of semantic differences in highly overlapping spaces still remains to be examined.

6.3 Conclusion

The end product of this research has been labeled the Semantic Measurement Model (SMM) which is a metric that quantifies the whole-space semantic difference between two LSA spaces. The SMM is composed of the OTV-Norm and the term overlap ratio between two spaces used together to describe the degree of semantic difference as defined at the end of Chapter 5. This measure is necessary for the purpose of understanding the semantic mappings that are defined by these hyper-dimensional spaces because the nature of these high-dimensional relationships is difficult for practitioners to reason about intuitively. This measure will provide reasoning leverage for analyzing the characteristics of a LSA space related to the overall body of meaning that it represents.

LIST OF REFERENCES

Berry, M. W., & Browne, M. (2005). *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. SIAM.

Biemiller, A., Rosenstein, M., Sparks, R., Landauer, T. K., & Foltz, P. W. (2014). Models of Vocabulary Acquisition: Direct Tests and Text-Derived Simulations of Vocabulary Growth. Scientific Studies of Reading, 18(2), 130–154.

Deerwester, S., Dumais, S. T., Furnas, G., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of The American Society for Information Science*, *41*(6), 391–407.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, *23*(2), 229–236.

Dumais, S. T., Furnas, G., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using Latent Semantic Analysis To Improve Access To Textual Information. In *SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS* (pp. 281–285). ACM.

Foltz, P. W. (1996). Latent Semantic Analysis for Text-based Research. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 197–202.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *28*(2-3), 285–307.

Foltz, P. W., Laham, R. D., & Derr, M. (2003). Automated speech recognition for modeling team performance. In *Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting*. Denver, CO.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent Essay Accessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, *1*(2).

Foltz, P. W., & Martin, M. A. (2008). Automated communication analysis of teams. In E. Salas, G. F. Goodwin, & S. Burke (Eds.), *Team Effectiveness in Complex Organizations and Systems: Cross-disciplinary perspectives and approaches*. New York: Routledge.

Foltz, P. W., Oberbreckling, R. J., & Laham, R. D. (2012). Analyzing Job and Occupational Content Using Latent Semantic Analysis. In M. A. Wilson, W. Bennett, S. G. Gibson, & G. M. Alliger (Eds.), *The Handbook of Work Analysis Methods, Systems, Applications and Science of Work Measurement in Organizations*. New York, NY: Taylor & Francis Group.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Accessor. In M. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation* (pp. 68–88). Routledge, NY: Routledge.

Golub, G. H., & Van Loan, C. F. (1996). *Matrix computations* (3rd ed.). Baltimore: Johns Hopkins University Press.

Hearst, M. A. (2000). The debate on automated essay grading. *IEEE Intelligent Systems and Their Applications*, *15*(5), 22–37.

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary Street: Computer-Guided Summary Writing. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kurby, C. A., Wiemer-Hastings, K., Ganduri, N., Magliano, J. P., Millis, K. K., & McNamara, D. S. (2003). Computerizing reading training: Evaluation of a latent semantic analysis space for science text. *Behavior Research Methods, Instruments, & Computers*, *35*(2), 244–250.

Laham, D., Bennett, W., & Landauer, T. K. (2000). An LSA-Based Software Tool for Matching Jobs, People, and Instruction. *Interactive Learning Environments*, *8*(3), 171–185.

Landauer, T. K. (1998). Learning and Representing Verbal Meaning: The Latent Semantic Analysis Theory. *Current Directions in Psychological Science*, *7*(5), 161–164.

Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. In *Psychology of Learning and Motivation* (Vol. 41, pp. 43–84). Elsevier Inc.

Landauer, T. K. (2007). LSA as a Theory of Meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*(2), 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*(2-3), 259–284.

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word Maturity: A New Metric for Word Knowledge. *Scientific Studies of Reading*, *15*(1), 92–108.

Landauer, T. K., Laham, D., & Foltz, P. W. (1998). Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems* (pp. 45–51). MIT Press.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated Scoring and Annotation of Essays with the Intelligent Essay Accessor. In M. D. Shermis & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-disciplinary Perspective* (pp. 87–112). Lawrence Erlbaum Associates.

Landauer, T. K., Lochbaum, K. E., & Dooley, S. (2009). A New Formative Assessment Technology for Reading and Writing. *Theory Into Practice*, *48*(1), 44–52.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Landauer, T. K., & Way, D. (2012). Improving text complexity measurement through the Reading Maturity Metric. Presented at the National Council on Measurement in Education, Vancouver, Canada.

LaVoie, N., Foltz, P., Rosenstein, M., Oberbreckling, R., Chatham, R. & Psotka, J. (2008, December). *Automated support for AARs: Exploiting communication to assess team performance*. Paper presented at I/ITSEC, Orlando, FL.

Lewis, D., Yang, Y., Rose, T., & Li, F. (2004). RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, *5*, 361–397.

Martin, D. I., & Berry, M. W. (2007). Mathematical Foundations Behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, &

W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Martin, D. I., & Berry, M. W. (2010). Latent Semantic Indexing. In M. J. Bates & M. N. Maack (Eds.), *Encyclopedia of Library and Information Sciences (ELIS)* (Third, pp. 2195–3204). Oxford: Taylor & Francis.

Martin, D. I., & Berry, M. W. (2015). Text mining. In N. Higham (Ed.), *Princeton Companion to Applied Mathematics*. Princeton, NJ: Princeton University Press.

Martin, D. I., Martin, J. C., & Berry, M. W. (2016). The Application of LSA to the Evaluation of Questionnaire Responses.

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to Assess Knowledge: Some Technical Considerations. *Discourse Processes*, *25*(2-3), 337–354.

Streeter, L., Lochbaum, K., Lavoie, N., & Psotka, J. (2007). Automated Tools for Collaborative Learning Environments. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Swygert, K., Margolis, M., King, A., Siftar, T., Clyman, S., Hawkins, R., & Clauser, B. (2003). Evaluation of an automated procedure for scoring patient notes as part of a clinical skills examination. *Academic Medicine, 78*(10), S75-77.

Trefethen, L. N., & Bau, D., III. (1997). *Numerical Linear Algebra*. SIAM.

APPENDIX

Summary of Appendix Items

Algorithm 1 describes the process for computing the OTV-Norm comparing two LSA spaces based on the mapping term vectors of the terms that they share in common. The OTV-Norm is described in Section 3.3.3 and is used in the experiments discussed in Chapters 4 and 5.

Table A - 1 on page 104 provides information on the LSA spaces constructed for the experiments as described in Section 4.2. This individual space information includes the number of terms, docs, and non-zeroes which are the term instances for all of the documents in the space. The space names indicate the primary experiments for which each space was constructed. The Group column values indicate the following: RTRC = Reuters texts used in the Control and General experiments (Sections 5.2.1 and 5.2.2), GL = grade level texts used in the grade level experiment (Section 5.2.3), Large = large size 6th and 9th grade sets used in the Large Volume experiment (Section 5.2.4), GLNO = grade level non-overlapping texts used in the Non-overlapping experiment (Section 5.2.5), FixV = fixed vocabulary sets used in the Frozen Vocabulary experiment (Section 5.2.6). The Set column values indicate the grade level or specific RTRC set used in the comparison. The Series column differentiates between multiple spaces within the same Group and Set.

Table A - 2 on page 105 gives the detailed centroid cosine and document-to-document cosine analysis information that was computed for each of the individual LSA spaces constructed for the experiments as described in Section 4.3.1. The Group, Set, and Series columns are the same as those defined for Table A - 1.

Table A - 3 starting on page 106 gives a portion of the observed measurement details for each of the pairs of LSA spaces compared for the experiments described in Section 4.3. These details include the document overlap figures for both the total count and overlap ratio, as well as the term overlap figures for both the total count and overlap ratio which are described in section 4.3.1. Also included are the observed total change percentage for three-tuple relationships between projected items as described in Section 4.3.2 for the four different sets of projection items that were used in the experiments. The Group, Set, and Series columns are the same as those defined for Table A - 1, describing both LSA spaces that were used in the comparison.

Table A - 4 starting on page 126 gives more of the observed measurement details for each of the pairs of LSA spaces compared for the experiments described in Section 4.3. These details include the OTV-Norm value as well as the cosine values between the document centroids and term centroids of the two spaces being compared as defined in Section 4.3.3. These values are based on the rotation matrices produced by using the NICHD04 anchors and the T-500 anchors. The Group, Set, and Series columns are the same as those defined for Table A - 1, describing both LSA spaces that were used in the comparison.

Table A - 5 starting on page 147 gives the remaining observed measurement details for each of the pairs of LSA spaces compared for the experiments described in Section 4.3. These details include the OTV-Norm value as well as the cosine values between the document centroids and term centroids of the two spaces being compared as defined in Section 4.3.3. These values are based on the rotation matrices produced by using the T-500 anchors and the T-5000 anchors. The Group, Set, and Series columns are the same as those defined for Table A - 1, describing both LSA spaces that were used in the comparison.

Algorithm 1: Algorithm for computing the OTV-Norm comparing two LSA spaces based on a set of anchor items

Given:

two LSA spaces for comparison, S_1 and S_2
one set of anchor items A

Step 1: Compute the rotation matrix Q to rotate S_1 into S_2 using Procrustes Algorithm (Golub and VanLoan, 1996):

For each a_i in A:
Project a_i into S_1 and append resulting vector to A_1
Project a_i into S_2 and append resulting vector to A_2

Compute SVD of ${A_1}^T A_2 \rightarrow USV^T$
$Q = UV^T$

Step 2: Compute the term overlap, $\hat{T}$, between S_1 and S_2:

For each term vector t_i in S_1:
If t_i is has a corresponding term vector w_j present in S_2:
Compute $t_i Q - w_j$ and append result vector to $\hat{T}$

Step 3: Compute the F-Norm for $\hat{T}$:

For each vector v_i in $\hat{T}$:
For each element e_j in v_i:
$sum = sum + (e_j)^2$

OTV-Norm = $\sqrt{sum}$

Table A - 1: Term and document information for LSA spaces used in experiments

Group	Set	Series	Terms	Docs	Non-zeros
RTRC	GCat	mod	272,729	234,869	42,097,799
RTRC	GCat150k	A	215,769	150,000	26,769,968
RTRC	GCat150k	B	225,754	150,000	26,865,391
RTRC	GCat		272,739	234,869	42,097,849
RTRC	GCat-1k		21,103	1,000	184,547
GL	GL 3	A	45,130	37,446	3,280,173
GL	GL 6	A	65,784	87,446	7,874,522
GL	GL 9	A	93,677	137,446	12,679,004
GL	GL 12	A	128,375	187,446	17,594,743
GL	Col	A	173,731	237,446	22,613,898
GL	GL 3	B	44,975	37,525	3,283,854
GL	GL 6	B	65,806	87,525	7,877,496
GL	GL 9	B	93,569	137,525	12,680,400
GL	GL 12	B	128,792	187,525	17,596,247
GL	Col	B	174,473	237,525	22,610,421
GL	GL 3	C	44,944	37,367	3,276,395
GL	GL 6	C	65,681	87,367	7,873,107
GL	GL 9	C	93,185	137,367	12,677,966
GL	GL 12	C	128,456	187,367	17,597,043
GL	Col	C	174,126	237,367	22,614,089
GL	GL 3	D	45,133	37,475	3,282,208
GL	GL 6	D	66,001	87,475	7,875,001
GL	GL 9	D	93,727	137,475	12,678,889
GL	GL 12	D	128,379	187,475	17,592,658
GL	Col	D	173,830	237,475	22,611,988
Large	GL 6	250k	116,586	229,753	21,319,141
Large	GL 9	250k	126,586	240,884	22,527,963
NOGL	GL 3	A	45,312	37,468	3,280,886
NOGL	GL 6	A	66,102	87,468	7,875,404
NOGL	GL 9	A	93,979	137,468	12,678,187
NOGL	GL 12	A	129,093	187,468	17,594,848
NOGL	Col	A	174,977	237,468	22,609,853
NOGL	GL 3	B	45,454	37,468	3,283,269
NOGL	GL 6	B	66,165	87,468	7,877,158
NOGL	GL 9	B	94,135	137,468	12,681,495
NOGL	GL 12	B	129,230	187,468	17,596,617
NOGL	Col	B	174,960	237,468	22,617,906
FixV	GL 3	A	33,069	28,270	2,344,511
FixV	GL 6	A	33,069	128,268	11,546,926
FixV	GL 12	A	33,069	228,268	20,960,487
FixV	GL 6	B	33,069	128,272	11,547,196
FixV	GL 12	B	33,069	228,272	20,963,119

Table A - 2: Term and document centroid information for LSA spaces used in experiments

Group	Set	Series	Avg. DC Cos	DC Std. Dev.	Avg. TC Cos.	TC Std. Dev.	Avg. DD Cos.	DD Std. Dev.
RTRC	GCat	mod	0.372702	0.101143	0.100770	0.082232	0.139533	0.088513
RTRC	GCat150k	A	0.370547	0.099875	0.102212	0.083612	0.137934	0.088118
RTRC	GCat150k	B	0.372007	0.100081	0.102214	0.084662	0.139012	0.088180
RTRC	GCat		0.372702	0.101144	0.100769	0.082230	0.139532	0.088513
RTRC	GCat-1k		0.305389	0.100406	0.115367	0.105372	0.093876	0.082031
GL	GL 3	A	0.391371	0.076122	0.122156	0.090548	0.153249	0.068743
GL	GL 6	A	0.397888	0.070962	0.129482	0.099025	0.158386	0.067662
GL	GL 9	A	0.406532	0.066354	0.134276	0.104184	0.165334	0.066664
GL	GL 12	A	0.411929	0.062068	0.135709	0.107537	0.169772	0.066205
GL	Col	A	0.412938	0.060771	0.135232	0.107757	0.170650	0.067310
GL	GL 3	B	0.391704	0.076271	0.122514	0.090638	0.153516	0.068913
GL	GL 6	B	0.398201	0.071433	0.129778	0.099006	0.158639	0.067910
GL	GL 9	B	0.406715	0.066552	0.134370	0.104178	0.165484	0.066751
GL	GL 12	B	0.411989	0.061966	0.135544	0.107340	0.169821	0.066159
GL	Col	B	0.412957	0.060670	0.134896	0.107511	0.170666	0.067275
GL	GL 3	C	0.391787	0.075573	0.122113	0.090588	0.153575	0.068604
GL	GL 6	C	0.398371	0.070602	0.129852	0.099185	0.158773	0.067549
GL	GL 9	C	0.406846	0.066321	0.134780	0.104378	0.165589	0.066671
GL	GL 12	C	0.412127	0.061912	0.135814	0.107644	0.169935	0.066135
GL	Col	C	0.413047	0.060677	0.135104	0.107756	0.170739	0.067264
GL	GL 3	D	0.391977	0.075734	0.122011	0.090459	0.153724	0.068673
GL	GL 6	D	0.398171	0.070805	0.129444	0.098907	0.158613	0.067676
GL	GL 9	D	0.406458	0.066135	0.134241	0.104069	0.165273	0.066571
GL	GL 12	D	0.411921	0.061960	0.135741	0.107491	0.169766	0.066158
GL	Col	D	0.412982	0.060880	0.135142	0.107640	0.170686	0.067359
Large	GL 6	250k	0.410560	0.064301	0.132341	0.108460	0.168614	0.065102
Large	GL 9	250k	0.411302	0.063557	0.134677	0.108875	0.169224	0.065697
NOGL	GL 3	A	0.391588	0.075826	0.121935	0.090653	0.153421	0.068682
NOGL	GL 6	A	0.397989	0.070821	0.129349	0.099005	0.158468	0.067645
NOGL	GL 9	A	0.406566	0.066300	0.134142	0.104236	0.165361	0.066643
NOGL	GL 12	A	0.411976	0.061909	0.135355	0.107403	0.169810	0.066139
NOGL	Col	A	0.412963	0.060801	0.134656	0.107521	0.170670	0.067298
NOGL	GL 3	B	0.391899	0.075408	0.122026	0.090624	0.153662	0.068523
NOGL	GL 6	B	0.398260	0.070750	0.129711	0.099005	0.158683	0.067639
NOGL	GL 9	B	0.406601	0.066071	0.134188	0.104050	0.165390	0.066572
NOGL	GL 12	B	0.411907	0.061812	0.135450	0.107358	0.169754	0.066099
NOGL	Col	B	0.412938	0.060547	0.134970	0.107576	0.170651	0.067240
FixV	GL 3	A	0.372345	0.086530	0.121443	0.088033	0.138746	0.072655
FixV	GL 6	A	0.405730	0.071498	0.204630	0.130633	0.164687	0.067694
FixV	GL 12	A	0.405709	0.071656	0.204520	0.130483	0.164670	0.067766
FixV	GL 6	B	0.414790	0.063113	0.254630	0.147631	0.172128	0.066243
FixV	GL 12	B	0.414797	0.063343	0.254196	0.147592	0.172133	0.066336

Table A - 3: Space comparison results for all experiments detailing document overlap, term overlap, and the TC% measure for each of the four projection sets used, NICHD04, T-500, T-1000, and T-5000. Group column values: GL = grade level, GLNO = grade level non-overlapping, Large = large size 6th and 9th grade sets, FixV = fixed vocabulary sets. Set column indicates the grade level or specific RTRC set used in the comparison.

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
RTRC	GCat		RTRC	GCat	mod	234,869	1.000000	272,729	0.999927	0.00%	0.00%	0.00%	0.00%
RTRC	GCat150k	B	RTRC	GCat		150,000	0.638654	225,754	0.827729	10.68%	10.18%	10.11%	10.56%
RTRC	GCat150k		RTRC	GCat150k	B	108,518	0.566727	192,452	0.772679	15.63%	14.95%	14.97%	15.63%
RTRC	GCat150k		RTRC	GCat		150,000	0.638654	215,769	0.791119	16.32%	16.40%	16.35%	16.98%
GL	3	A	GL	6	A	37,446	0.428219	45,130	0.686033	31.55%	32.44%	34.51%	34.32%
GL	3	A	GL	9	A	37,446	0.272442	45,130	0.481762	38.57%	39.05%	41.66%	41.22%
GL	3	A	GL	12	A	37,446	0.199770	45,130	0.351548	44.13%	42.88%	46.12%	45.51%
GL	3	A	GL	Col	A	37,446	0.157703	45,130	0.259769	47.42%	49.91%	53.70%	53.40%
GL	3	A	GL	3	B	18,833	0.335477	36,869	0.692558	31.08%	25.42%	27.33%	26.06%
GL	3	A	GL	6	B	18,834	0.177450	39,813	0.559777	34.76%	34.42%	36.66%	36.26%
GL	3	A	GL	9	B	18,834	0.120625	41,382	0.425229	39.96%	39.94%	42.57%	42.21%
GL	3	A	GL	12	B	18,834	0.091366	42,220	0.320572	45.10%	43.38%	46.67%	46.07%
GL	3	A	GL	Col	B	18,834	0.073531	42,556	0.240366	47.85%	50.30%	53.94%	53.66%
GL	3	A	GL	3	C	18,772	0.334969	36,888	0.693566	31.00%	25.34%	27.11%	26.06%
GL	3	A	GL	6	C	18,772	0.177026	39,838	0.561312	35.10%	34.19%	36.39%	36.14%
GL	3	A	GL	9	C	18,772	0.120302	41,337	0.426251	40.34%	40.06%	42.60%	42.18%
GL	3	A	GL	12	C	18,772	0.091108	42,227	0.321463	45.25%	43.61%	46.76%	46.13%
GL	3	A	GL	Col	C	18,772	0.073316	42,582	0.241020	48.04%	50.38%	54.01%	53.82%
GL	3	A	GL	3	D	18,686	0.332284	36,863	0.690318	30.86%	25.34%	27.16%	25.95%
GL	3	A	GL	6	D	18,686	0.175893	39,860	0.559274	34.69%	34.35%	36.55%	36.15%
GL	3	A	GL	9	D	18,686	0.119602	41,375	0.424437	39.96%	39.98%	42.59%	42.07%
GL	3	A	GL	12	D	18,686	0.090605	42,211	0.321490	44.81%	43.33%	46.53%	45.84%
GL	3	A	GL	Col	D	18,686	0.072925	42,577	0.241389	47.68%	50.15%	53.93%	53.68%
GL	6	A	GL	9	A	87,446	0.636221	65,784	0.702243	26.66%	24.09%	25.79%	24.82%
GL	6	A	GL	12	A	87,446	0.466513	65,784	0.512436	35.81%	31.86%	34.26%	33.17%
GL	6	A	GL	Col	A	87,446	0.368277	65,784	0.378654	40.93%	42.40%	45.31%	44.91%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GL	6	A	GL	3	B	18,834	0.177450	39,670	0.558033	35.48%	34.65%	36.72%	36.35%
GL	6	A	GL	6	B	44,000	0.335952	54,274	0.701976	25.44%	21.61%	23.25%	21.92%
GL	6	A	GL	9	B	44,002	0.243147	57,550	0.565308	30.61%	27.30%	29.16%	28.26%
GL	6	A	GL	12	B	44,002	0.190510	59,188	0.437173	37.66%	33.17%	35.67%	34.67%
GL	6	A	GL	Col	B	44,002	0.156608	59,986	0.332755	41.87%	43.19%	45.97%	45.59%
GL	6	A	GL	3	C	18,773	0.177037	39,722	0.559418	35.77%	34.66%	36.74%	36.49%
GL	6	A	GL	6	C	43,751	0.333819	54,179	0.701020	26.04%	21.76%	23.33%	22.08%
GL	6	A	GL	9	C	43,753	0.241649	57,374	0.564733	31.26%	27.58%	29.46%	28.34%
GL	6	A	GL	12	C	43,753	0.189358	58,996	0.436219	37.81%	33.40%	35.80%	34.69%
GL	6	A	GL	Col	C	43,753	0.155671	59,882	0.332626	42.03%	43.28%	46.08%	45.78%
GL	6	A	GL	3	D	18,687	0.175904	39,732	0.558151	35.44%	34.88%	36.95%	36.45%
GL	6	A	GL	6	D	43,817	0.334216	54,348	0.701835	25.97%	21.57%	23.17%	21.91%
GL	6	A	GL	9	D	43,819	0.241958	57,606	0.565291	30.95%	27.32%	29.21%	27.98%
GL	6	A	GL	12	D	43,819	0.189609	59,234	0.439001	37.51%	33.22%	35.69%	34.44%
GL	6	A	GL	Col	D	43,819	0.155883	60,062	0.334510	41.65%	43.02%	45.97%	45.59%
GL	9	A	GL	12	A	137,446	0.733257	93,677	0.729714	23.95%	21.11%	22.64%	21.82%
GL	9	A	GL	Col	A	137,446	0.578852	93,677	0.539207	32.49%	35.15%	37.46%	37.29%
GL	9	A	GL	3	B	18,834	0.120625	41,267	0.423751	40.01%	39.88%	42.42%	42.00%
GL	9	A	GL	6	B	44,001	0.243140	57,625	0.565739	31.10%	27.50%	29.68%	28.24%
GL	9	A	GL	9	B	68,950	0.334675	76,819	0.695654	23.29%	20.12%	21.60%	20.50%
GL	9	A	GL	12	B	68,950	0.269314	80,588	0.567997	28.16%	24.71%	26.60%	25.68%
GL	9	A	GL	Col	B	68,950	0.225311	82,503	0.444408	34.27%	36.52%	38.84%	38.60%
GL	9	A	GL	3	C	18,773	0.120309	41,320	0.424662	40.74%	39.95%	42.53%	42.18%
GL	9	A	GL	6	C	43,753	0.241649	57,547	0.565234	31.05%	27.54%	29.50%	28.20%
GL	9	A	GL	9	C	68,749	0.333629	76,531	0.693649	23.48%	20.01%	21.44%	20.33%
GL	9	A	GL	12	C	68,749	0.268484	80,312	0.566291	28.50%	25.04%	26.67%	25.72%
GL	9	A	GL	Col	C	68,749	0.224623	82,309	0.443729	34.43%	36.68%	38.94%	38.83%
GL	9	A	GL	3	D	18,687	0.119609	41,357	0.424379	40.17%	40.00%	42.62%	42.09%
GL	9	A	GL	6	D	43,819	0.241958	57,689	0.565639	30.99%	27.71%	29.67%	28.43%
GL	9	A	GL	9	D	68,883	0.334322	76,840	0.694982	23.21%	19.95%	21.42%	20.28%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GL	9	A	GL	12	D	68,883	0.269034	80,617	0.569977	27.92%	25.02%	26.74%	25.69%
GL	9	A	GL	Col	D	68,883	0.225080	82,533	0.446187	33.89%	36.37%	38.83%	38.60%
GL	12	A	GL	Col	A	187,446	0.789426	128,375	0.738930	22.29%	27.41%	28.89%	29.10%
GL	12	A	GL	3	B	18,834	0.091366	42,092	0.320681	44.60%	43.42%	46.52%	45.94%
GL	12	A	GL	6	B	44,001	0.190505	59,224	0.438836	37.55%	33.32%	35.96%	34.57%
GL	12	A	GL	9	B	68,950	0.269314	80,473	0.568830	28.25%	24.97%	26.79%	25.67%
GL	12	A	GL	12	B	93,958	0.334355	104,015	0.679162	20.75%	18.91%	20.17%	19.14%
GL	12	A	GL	Col	B	93,958	0.283850	108,323	0.556859	25.98%	30.10%	31.84%	31.81%
GL	12	A	GL	3	C	18,773	0.091113	42,176	0.321603	45.23%	43.36%	46.51%	46.03%
GL	12	A	GL	6	C	43,753	0.189358	59,209	0.439083	37.39%	33.26%	35.78%	34.45%
GL	12	A	GL	9	C	68,749	0.268484	80,204	0.567390	28.02%	24.83%	26.64%	25.54%
GL	12	A	GL	12	C	93,719	0.333408	103,760	0.677855	20.73%	19.10%	20.13%	19.15%
GL	12	A	GL	Col	C	93,719	0.283059	108,136	0.556355	26.19%	30.35%	31.85%	32.05%
GL	12	A	GL	3	D	18,687	0.090611	42,217	0.321553	44.80%	43.52%	46.65%	46.01%
GL	12	A	GL	6	D	43,819	0.189609	59,335	0.439385	37.66%	33.49%	35.98%	34.77%
GL	12	A	GL	9	D	68,885	0.269044	80,563	0.569193	28.30%	24.76%	26.75%	25.70%
GL	12	A	GL	12	D	93,975	0.334495	103,706	0.677604	20.71%	18.83%	20.07%	19.05%
GL	12	A	GL	Col	D	93,977	0.283966	107,950	0.555713	25.58%	29.73%	31.64%	31.72%
GL	Col	A	GL	3	B	18,834	0.073531	42,457	0.240892	48.05%	50.28%	54.02%	53.76%
GL	Col	A	GL	6	B	44,001	0.156604	60,090	0.334862	42.18%	43.08%	46.30%	45.59%
GL	Col	A	GL	9	B	68,950	0.225311	82,472	0.446209	34.56%	36.50%	38.95%	38.49%
GL	Col	A	GL	12	B	93,962	0.283865	108,471	0.558979	26.18%	30.06%	31.72%	31.47%
GL	Col	A	GL	Col	B	119,257	0.335261	138,255	0.658517	18.95%	19.02%	20.24%	19.34%
GL	Col	A	GL	3	C	18,773	0.073321	42,538	0.241505	48.02%	49.97%	53.78%	53.54%
GL	Col	A	GL	6	C	43,753	0.155671	60,048	0.334783	41.84%	43.07%	46.13%	45.41%
GL	Col	A	GL	9	C	68,749	0.224623	82,123	0.444405	33.95%	36.22%	38.71%	38.22%
GL	Col	A	GL	12	C	93,721	0.283066	108,142	0.557304	25.73%	29.73%	31.55%	31.39%
GL	Col	A	GL	Col	C	118,775	0.333602	138,022	0.657764	18.51%	18.74%	19.92%	19.13%
GL	Col	A	GL	3	D	18,687	0.072929	42,587	0.241591	47.84%	50.27%	54.01%	53.70%
GL	Col	A	GL	6	D	43,819	0.155883	60,193	0.335264	42.04%	43.34%	46.34%	45.86%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GL	Col	A	GL	9	D	68,885	0.225088	82,520	0.446204	34.14%	36.25%	38.96%	38.67%
GL	Col	A	GL	12	D	93,977	0.283966	108,102	0.557204	25.91%	29.93%	31.68%	31.62%
GL	Col	A	GL	Col	D	119,139	0.334865	137,828	0.657159	18.66%	18.44%	19.93%	19.07%
GL	3	B	GL	6	B	37,525	0.428735	44,975	0.683448	31.70%	32.31%	34.31%	34.12%
GL	3	B	GL	9	B	37,525	0.272859	44,975	0.480661	38.46%	38.90%	41.48%	41.26%
GL	3	B	GL	12	B	37,525	0.200107	44,975	0.349206	44.12%	42.78%	45.95%	45.49%
GL	3	B	GL	Col	B	37,525	0.157983	44,975	0.257776	47.59%	50.13%	53.67%	53.50%
GL	3	B	GL	3	C	18,855	0.336474	36,809	0.693071	31.56%	25.85%	27.56%	26.39%
GL	3	B	GL	6	C	18,855	0.177815	39,654	0.558491	35.36%	34.34%	36.41%	36.15%
GL	3	B	GL	9	C	18,855	0.120837	41,193	0.424815	40.32%	40.17%	42.56%	42.15%
GL	3	B	GL	12	C	18,855	0.091513	42,056	0.320122	45.00%	43.69%	46.66%	46.09%
GL	3	B	GL	Col	C	18,855	0.073642	42,402	0.239967	48.16%	50.54%	54.07%	53.92%
GL	3	B	GL	3	D	19,004	0.339381	36,926	0.694333	31.12%	25.65%	27.23%	26.05%
GL	3	B	GL	6	D	19,004	0.179290	39,830	0.559835	35.14%	34.36%	36.39%	36.08%
GL	3	B	GL	9	D	19,004	0.121824	41,308	0.424133	40.04%	40.02%	42.50%	42.03%
GL	3	B	GL	12	D	19,004	0.092254	42,135	0.321104	44.60%	43.28%	46.45%	45.81%
GL	3	B	GL	Col	D	19,004	0.074236	42,475	0.240884	47.71%	50.20%	53.92%	53.77%
GL	6	B	GL	9	B	87,525	0.636430	65,806	0.703288	26.69%	23.93%	25.89%	25.02%
GL	6	B	GL	12	B	87,525	0.466738	65,806	0.510948	36.14%	31.85%	34.45%	33.39%
GL	6	B	GL	Col	B	87,525	0.368488	65,806	0.377170	41.40%	42.54%	45.61%	45.10%
GL	6	B	GL	3	C	18,857	0.177838	39,708	0.558937	35.59%	34.51%	36.61%	36.39%
GL	6	B	GL	6	C	43,769	0.333801	54,222	0.701767	25.67%	21.75%	23.47%	22.07%
GL	6	B	GL	9	C	43,771	0.241667	57,423	0.565365	31.41%	27.59%	29.67%	28.32%
GL	6	B	GL	12	C	43,771	0.189386	59,080	0.437040	38.01%	33.50%	36.04%	34.75%
GL	6	B	GL	Col	C	43,771	0.155702	59,935	0.332978	42.40%	43.25%	46.39%	45.85%
GL	6	B	GL	3	D	19,007	0.179323	39,855	0.560675	35.04%	34.72%	36.71%	36.33%
GL	6	B	GL	6	D	44,045	0.336337	54,461	0.704122	25.59%	21.75%	23.23%	21.98%
GL	6	B	GL	9	D	44,046	0.243410	57,676	0.566245	31.22%	27.45%	29.49%	28.11%
GL	6	B	GL	12	D	44,047	0.190718	59,274	0.439356	37.68%	33.24%	35.92%	34.50%
GL	6	B	GL	Col	D	44,047	0.156777	60,062	0.334469	41.88%	42.97%	46.25%	45.66%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GL	9	B	GL	12	B	137,525	0.733369	93,569	0.726513	24.52%	21.15%	22.71%	21.99%
GL	9	B	GL	Col	B	137,525	0.578992	93,569	0.536295	32.82%	35.24%	37.46%	37.30%
GL	9	B	GL	3	C	18,857	0.120851	41,282	0.424577	40.49%	39.93%	42.47%	42.26%
GL	9	B	GL	6	C	43,770	0.241660	57,546	0.565818	30.77%	27.38%	29.32%	28.28%
GL	9	B	GL	9	C	68,603	0.332558	76,415	0.692548	23.75%	20.14%	21.40%	20.48%
GL	9	B	GL	12	C	68,603	0.267678	80,261	0.566159	28.83%	25.23%	26.80%	25.77%
GL	9	B	GL	Col	C	68,603	0.223981	82,251	0.443536	34.85%	36.69%	38.96%	38.74%
GL	9	B	GL	3	D	19,007	0.121845	41,384	0.425245	40.01%	40.06%	42.60%	42.23%
GL	9	B	GL	6	D	44,046	0.243410	57,695	0.566331	30.89%	27.54%	29.43%	28.53%
GL	9	B	GL	9	D	69,109	0.335658	76,830	0.695508	23.38%	19.91%	21.37%	20.43%
GL	9	B	GL	12	D	69,110	0.270077	80,585	0.570057	28.37%	24.93%	26.77%	25.70%
GL	9	B	GL	Col	D	69,110	0.225931	82,501	0.446197	34.14%	36.29%	38.80%	38.46%
GL	12	B	GL	Col	B	187,525	0.789496	128,792	0.738177	22.40%	27.53%	28.86%	29.04%
GL	12	B	GL	3	C	18,857	0.091523	42,118	0.320002	45.45%	43.29%	46.53%	46.04%
GL	12	B	GL	6	C	43,770	0.189381	59,208	0.437719	37.68%	33.16%	35.77%	34.58%
GL	12	B	GL	9	C	68,603	0.267678	80,179	0.565445	28.39%	24.78%	26.55%	25.63%
GL	12	B	GL	12	C	93,727	0.333352	104,157	0.680360	20.85%	18.76%	19.95%	19.02%
GL	12	B	GL	Col	C	93,728	0.283026	108,577	0.558693	26.57%	30.32%	31.74%	31.81%
GL	12	B	GL	3	D	19,007	0.092270	42,271	0.321076	45.07%	43.53%	46.66%	46.07%
GL	12	B	GL	6	D	44,046	0.190713	59,375	0.438457	37.83%	33.46%	35.92%	34.94%
GL	12	B	GL	9	D	69,109	0.270072	80,608	0.568018	28.45%	24.67%	26.65%	25.78%
GL	12	B	GL	12	D	94,160	0.335280	103,934	0.678257	20.79%	18.71%	20.06%	19.04%
GL	12	B	GL	Col	D	94,162	0.284617	108,222	0.556698	25.91%	29.77%	31.54%	31.47%
GL	Col	B	GL	3	C	18,857	0.073650	42,451	0.239882	48.01%	50.07%	53.76%	53.52%
GL	Col	B	GL	6	C	43,770	0.155698	60,005	0.333085	41.90%	43.21%	46.15%	45.50%
GL	Col	B	GL	9	C	68,603	0.223981	82,077	0.442270	33.98%	36.34%	38.70%	38.28%
GL	Col	B	GL	12	C	93,727	0.283022	108,423	0.557428	25.87%	29.88%	31.65%	31.56%
GL	Col	B	GL	Col	C	118,780	0.333547	138,349	0.658021	18.72%	18.72%	19.90%	19.12%
GL	Col	B	GL	3	D	19,007	0.074248	42,614	0.240768	47.74%	50.37%	53.97%	53.72%
GL	Col	B	GL	6	D	44,046	0.156773	60,198	0.333921	42.04%	43.49%	46.31%	45.94%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GL	Col	B	GL	9	D	69,109	0.225927	82,534	0.444529	33.97%	36.39%	38.93%	38.74%
GL	Col	B	GL	12	D	94,160	0.284609	108,268	0.556408	25.81%	29.96%	31.77%	31.72%
GL	Col	B	GL	Col	D	119,333	0.335519	138,209	0.657844	18.44%	18.68%	20.08%	18.98%
GL	3	C	GL	6	C	37,367	0.427702	44,944	0.684277	32.56%	31.85%	34.09%	34.18%
GL	3	C	GL	9	C	37,367	0.272023	44,944	0.482309	39.57%	39.10%	41.60%	41.39%
GL	3	C	GL	12	C	37,367	0.199432	44,944	0.349879	44.87%	42.92%	46.06%	45.62%
GL	3	C	GL	Col	C	37,367	0.157423	44,944	0.258112	47.76%	49.83%	53.52%	53.42%
GL	3	C	GL	3	D	18,784	0.335082	36,942	0.695248	31.56%	25.73%	27.43%	26.22%
GL	3	C	GL	6	D	18,784	0.177111	39,860	0.560737	35.65%	34.25%	36.47%	36.21%
GL	3	C	GL	9	D	18,784	0.120366	41,358	0.425000	40.53%	39.88%	42.48%	42.11%
GL	3	C	GL	12	D	18,784	0.091159	42,170	0.321533	45.36%	43.24%	46.47%	45.86%
GL	3	C	GL	Col	D	18,784	0.073358	42,508	0.241158	47.93%	49.92%	53.77%	53.58%
GL	6	C	GL	9	C	87,367	0.636012	65,681	0.704845	27.34%	24.29%	26.01%	25.00%
GL	6	C	GL	12	C	87,367	0.466288	65,681	0.511311	36.31%	31.96%	34.38%	33.19%
GL	6	C	GL	Col	C	87,367	0.368067	65,681	0.377204	41.32%	42.63%	45.57%	45.09%
GL	6	C	GL	3	D	18,784	0.177111	39,877	0.562147	35.65%	34.40%	36.54%	36.21%
GL	6	C	GL	6	D	43,797	0.334213	54,538	0.706964	26.23%	21.78%	23.37%	22.08%
GL	6	C	GL	9	D	43,798	0.241919	57,763	0.568282	31.18%	27.45%	29.43%	28.10%
GL	6	C	GL	12	D	43,798	0.189566	59,359	0.440672	37.61%	33.24%	35.77%	34.36%
GL	6	C	GL	Col	D	43,798	0.155840	60,141	0.335290	41.75%	43.03%	46.17%	45.52%
GL	9	C	GL	12	C	137,367	0.733144	93,185	0.725423	24.30%	21.08%	22.50%	21.73%
GL	9	C	GL	Col	C	137,367	0.578711	93,185	0.535158	32.61%	35.21%	37.43%	37.30%
GL	9	C	GL	3	D	18,784	0.120366	41,416	0.427401	40.66%	40.19%	42.65%	42.20%
GL	9	C	GL	6	D	43,799	0.241926	57,711	0.568721	31.60%	27.80%	29.65%	28.54%
GL	9	C	GL	9	D	68,853	0.334256	76,840	0.698089	23.75%	19.80%	21.31%	20.21%
GL	9	C	GL	12	D	68,853	0.268969	80,488	0.570529	28.47%	25.00%	26.70%	25.61%
GL	9	C	GL	Col	D	68,853	0.225018	82,313	0.445653	33.92%	36.21%	38.74%	38.31%
GL	12	C	GL	Col	C	187,367	0.789356	128,456	0.737719	22.33%	27.27%	28.61%	29.13%
GL	12	C	GL	3	D	18,784	0.091159	42,249	0.321677	45.31%	43.76%	46.76%	46.13%
GL	12	C	GL	6	D	43,799	0.189571	59,304	0.438792	38.22%	33.70%	36.04%	34.94%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GL	12	C	GL	9	D	68,854	0.268974	80,594	0.569211	28.80%	24.80%	26.67%	25.83%
GL	12	C	GL	12	D	93,886	0.334166	103,896	0.679330	20.80%	18.90%	20.05%	19.16%
GL	12	C	GL	Col	D	93,887	0.283685	108,087	0.556579	25.86%	29.74%	31.63%	31.60%
GL	Col	C	GL	3	D	18,784	0.073358	42,615	0.241248	48.02%	50.46%	54.04%	53.85%
GL	Col	C	GL	6	D	43,799	0.155844	60,173	0.334380	42.39%	43.62%	46.47%	46.16%
GL	Col	C	GL	9	D	68,854	0.225022	82,538	0.445393	34.44%	36.53%	39.06%	38.98%
GL	Col	C	GL	12	D	93,889	0.283693	108,317	0.557795	26.31%	30.25%	31.82%	32.02%
GL	Col	C	GL	Col	D	119,014	0.334471	137,962	0.656981	18.70%	18.97%	20.11%	19.24%
GL	3	D	GL	6	D	37,475	0.428408	45,133	0.683823	31.41%	32.25%	34.32%	34.13%
GL	3	D	GL	9	D	37,475	0.272595	45,133	0.481537	38.41%	39.12%	41.64%	41.21%
GL	3	D	GL	12	D	37,475	0.199893	45,133	0.351561	43.96%	42.91%	46.01%	45.38%
GL	3	D	GL	Col	D	37,475	0.157806	45,133	0.259639	47.06%	49.94%	53.65%	53.48%
GL	6	D	GL	9	D	87,475	0.636298	66,001	0.704183	27.11%	24.59%	26.21%	25.19%
GL	6	D	GL	12	D	87,475	0.466596	66,001	0.514111	36.09%	32.12%	34.48%	33.39%
GL	6	D	GL	Col	D	87,475	0.368355	66,001	0.379687	40.98%	42.67%	45.68%	45.40%
GL	9	D	GL	12	D	137,475	0.733298	93,727	0.730080	24.27%	21.11%	22.83%	22.10%
GL	9	D	GL	Col	D	137,475	0.578903	93,727	0.539188	32.09%	34.94%	37.63%	37.59%
GL	12	D	GL	Col	D	187,475	0.789452	128,379	0.738532	21.81%	26.96%	28.67%	29.09%
Large	6	250kB	GL	3	A	18,812	0.075737	42,193	0.353012	37.30%	36.18%	38.50%	38.21%
Large	6	250kB	GL	6	A	43,807	0.160235	58,943	0.477554	31.39%	28.41%	30.00%	28.77%
Large	6	250kB	GL	9	A	43,807	0.135461	72,951	0.531279	29.63%	27.22%	28.75%	27.55%
Large	6	250kB	GL	12	A	43,807	0.117322	81,786	0.501216	34.42%	30.46%	32.55%	31.49%
Large	6	250kB	GL	Col	A	43,807	0.103467	86,601	0.425107	39.66%	40.28%	42.98%	42.43%
Large	6	250kB	GL	3	B	18,784	0.075591	42,111	0.352541	37.11%	36.02%	38.40%	38.13%
Large	6	250kB	GL	6	B	43,722	0.159828	59,003	0.478187	31.07%	28.29%	29.94%	28.54%
Large	6	250kB	GL	9	B	43,722	0.135130	73,043	0.532725	29.71%	27.36%	28.69%	27.61%
Large	6	250kB	GL	12	B	43,722	0.117043	81,973	0.501655	34.52%	30.49%	32.47%	31.48%
Large	6	250kB	GL	Col	B	43,722	0.103226	86,709	0.424316	39.66%	40.50%	42.99%	42.55%
Large	6	250kB	GL	3	C	18,714	0.075336	42,134	0.352893	37.87%	36.31%	38.46%	38.25%
Large	6	250kB	GL	6	C	43,683	0.159755	58,968	0.478252	31.31%	28.08%	29.63%	28.14%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
Large	6	250kB	GL	9	C	43,685	0.135066	72,750	0.530941	30.23%	27.33%	28.68%	27.36%
Large	6	250kB	GL	12	C	43,685	0.116982	81,699	0.500168	34.92%	30.71%	32.56%	31.47%
Large	6	250kB	GL	Col	C	43,685	0.103168	86,484	0.423468	39.98%	40.60%	43.03%	42.72%
Large	6	250kB	GL	3	D	18,729	0.075369	42,219	0.353297	37.25%	36.09%	38.48%	38.14%
Large	6	250kB	GL	6	D	43,667	0.159624	59,164	0.479360	30.79%	28.33%	30.06%	28.84%
Large	6	250kB	GL	9	D	43,668	0.134961	73,119	0.532961	29.76%	27.36%	28.91%	27.60%
Large	6	250kB	GL	12	D	43,668	0.116897	82,057	0.503701	34.31%	30.40%	32.56%	31.41%
Large	6	250kB	GL	Col	D	43,668	0.103098	86,800	0.426293	39.31%	40.21%	42.93%	42.56%
Large	9	250kB	GL	3	A	18,661	0.071865	42,276	0.326607	42.05%	41.79%	44.42%	44.04%
Large	9	250kB	GL	6	A	43,834	0.154076	59,458	0.447349	34.26%	31.75%	33.66%	32.91%
Large	9	250kB	GL	9	A	68,849	0.222466	81,439	0.586635	25.29%	23.47%	24.81%	24.03%
Large	9	250kB	GL	12	A	68,852	0.191533	90,961	0.554640	25.63%	23.79%	25.23%	24.66%
Large	9	250kB	GL	Col	A	68,852	0.168146	96,217	0.471421	32.02%	33.75%	36.06%	35.64%
Large	9	250kB	GL	3	B	18,600	0.071591	42,060	0.324785	41.89%	41.76%	44.35%	44.01%
Large	9	250kB	GL	6	B	43,524	0.152777	59,396	0.446600	34.47%	31.86%	33.82%	32.81%
Large	9	250kB	GL	9	B	68,548	0.221222	81,267	0.585126	25.96%	23.93%	25.08%	24.21%
Large	9	250kB	GL	12	B	68,553	0.190501	90,956	0.553186	25.89%	23.87%	25.22%	24.59%
Large	9	250kB	GL	Col	B	68,553	0.167261	96,144	0.469190	31.94%	33.90%	36.04%	35.73%
Large	9	250kB	GL	3	C	18,703	0.072060	42,182	0.326113	42.53%	41.81%	44.39%	44.08%
Large	9	250kB	GL	6	C	43,821	0.154066	59,501	0.448164	34.74%	31.85%	33.72%	32.78%
Large	9	250kB	GL	9	C	68,911	0.222768	80,935	0.582954	26.41%	23.65%	24.84%	24.07%
Large	9	250kB	GL	12	C	68,914	0.191781	90,686	0.551766	26.51%	24.04%	25.28%	24.70%
Large	9	250kB	GL	Col	C	68,914	0.168355	95,961	0.468672	32.27%	34.03%	36.03%	35.90%
Large	9	250kB	GL	3	D	18,568	0.071473	42,249	0.326323	42.09%	41.84%	44.45%	44.03%
Large	9	250kB	GL	6	D	43,518	0.152780	59,622	0.448404	34.58%	32.14%	34.02%	33.24%
Large	9	250kB	GL	9	D	68,448	0.220863	81,299	0.584826	25.92%	23.75%	25.23%	24.53%
Large	9	250kB	GL	12	D	68,453	0.190197	91,051	0.555480	25.82%	23.84%	25.43%	24.96%
Large	9	250kB	GL	Col	D	68,453	0.166997	96,188	0.470983	31.67%	33.61%	35.97%	35.72%
GLNO	3	A	GL	6	250kB	18,637	0.074973	42,386	0.354659	37.11%	36.29%	38.47%	38.21%
GLNO	3	A	GL	9	250kB	18,567	0.071471	42,424	0.327664	41.63%	41.75%	44.25%	44.00%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GLNO	3	A	GL	3	A	18,890	0.337177	37,178	0.697995	30.66%	25.17%	27.16%	25.82%
GLNO	3	A	GL	6	A	18,890	0.178167	40,014	0.562927	34.82%	34.59%	36.70%	36.37%
GLNO	3	A	GL	9	A	18,890	0.121071	41,597	0.427109	39.55%	39.73%	42.34%	41.99%
GLNO	3	A	GL	12	A	18,890	0.091688	42,424	0.323198	44.44%	43.34%	46.48%	45.96%
GLNO	3	A	GL	Col	A	18,890	0.073782	42,798	0.242832	47.61%	50.21%	53.91%	53.67%
GLNO	3	A	GL	3	B	18,744	0.333233	36,885	0.690704	31.05%	25.47%	27.07%	25.87%
GLNO	3	A	GL	6	B	18,745	0.176427	39,893	0.560098	34.65%	34.37%	36.38%	36.13%
GLNO	3	A	GL	9	B	18,745	0.119970	41,480	0.425868	39.58%	39.84%	42.36%	42.14%
GLNO	3	A	GL	12	B	18,745	0.090886	42,335	0.321282	44.77%	43.28%	46.45%	45.98%
GLNO	3	A	GL	Col	B	18,745	0.073152	42,696	0.241099	47.75%	50.30%	53.86%	53.67%
GLNO	3	A	GL	3	C	18,665	0.332295	37,012	0.695139	31.29%	25.16%	27.21%	26.13%
GLNO	3	A	GL	6	C	18,665	0.175803	39,930	0.561896	35.17%	34.14%	36.33%	36.12%
GLNO	3	A	GL	9	C	18,665	0.119517	41,454	0.427171	40.24%	39.99%	42.43%	42.11%
GLNO	3	A	GL	12	C	18,665	0.090532	42,315	0.321902	45.02%	43.49%	46.60%	46.07%
GLNO	3	A	GL	Col	C	18,665	0.072862	42,682	0.241474	47.89%	50.39%	53.95%	53.83%
GLNO	3	A	GL	3	D	19,029	0.340326	37,125	0.696268	30.90%	25.36%	27.34%	26.07%
GLNO	3	A	GL	6	D	19,029	0.179665	40,051	0.562025	34.59%	34.10%	36.26%	36.00%
GLNO	3	A	GL	9	D	19,029	0.122048	41,600	0.426934	39.59%	39.78%	42.31%	41.95%
GLNO	3	A	GL	12	D	19,029	0.092412	42,435	0.323300	44.50%	43.24%	46.41%	45.80%
GLNO	3	A	GL	Col	D	19,029	0.074357	42,806	0.242752	47.52%	50.15%	53.85%	53.70%
GLNO	3	A	GLNO	6	A	37,468	0.428362	45,312	0.685486	31.46%	31.99%	33.95%	33.93%
GLNO	3	A	GLNO	9	A	37,468	0.272558	45,312	0.482150	38.24%	38.73%	41.28%	41.08%
GLNO	3	A	GLNO	12	A	37,468	0.199863	45,312	0.351003	43.87%	42.78%	46.00%	45.53%
GLNO	3	A	GLNO	Col	A	37,468	0.157781	45,312	0.258960	47.11%	50.04%	53.67%	53.51%
GLNO	3	A	GLNO	3	B	108	0.001443	30,810	0.513877	37.55%	30.91%	33.16%	31.69%
GLNO	3	A	GLNO	6	B	108	0.000865	35,491	0.467073	37.20%	36.09%	38.33%	37.98%
GLNO	3	A	GLNO	9	B	108	0.000618	38,260	0.378112	40.94%	40.55%	43.14%	42.74%
GLNO	3	A	GLNO	12	B	108	0.000480	39,772	0.295110	45.16%	43.85%	46.97%	46.37%
GLNO	3	A	GLNO	Col	B	108	0.000393	40,426	0.224781	47.90%	50.28%	53.98%	53.78%
GLNO	6	A	GL	6	250kB	43,614	0.159404	59,223	0.479674	31.53%	28.33%	29.93%	28.56%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GLNO	6	A	GL	9	250kB	43,555	0.152933	59,713	0.449054	34.33%	31.93%	33.82%	32.98%
GLNO	6	A	GL	3	A	18,890	0.178167	39,992	0.561370	35.31%	34.28%	36.49%	36.07%
GLNO	6	A	GL	6	A	43,951	0.335599	54,658	0.707748	25.59%	21.81%	23.23%	21.97%
GLNO	6	A	GL	9	A	43,951	0.242873	57,946	0.569030	30.67%	27.50%	29.47%	28.20%
GLNO	6	A	GL	12	A	43,951	0.190295	59,549	0.441339	37.23%	33.44%	35.87%	34.56%
GLNO	6	A	GL	Col	A	43,951	0.156430	60,378	0.336452	41.81%	43.24%	46.23%	45.66%
GLNO	6	A	GL	3	B	18,744	0.176416	39,734	0.556943	35.51%	34.34%	36.32%	35.95%
GLNO	6	A	GL	6	B	43,826	0.334124	54,495	0.703952	25.07%	21.34%	22.96%	21.73%
GLNO	6	A	GL	9	B	43,827	0.241916	57,839	0.567985	30.38%	27.32%	29.24%	28.28%
GLNO	6	A	GL	12	B	43,827	0.189591	59,487	0.439320	37.53%	33.34%	35.81%	34.73%
GLNO	6	A	GL	Col	B	43,827	0.155876	60,284	0.334371	41.93%	43.39%	46.21%	45.74%
GLNO	6	A	GL	3	C	18,666	0.175814	39,794	0.558497	36.00%	34.23%	36.43%	36.22%
GLNO	6	A	GL	6	C	43,711	0.333356	54,491	0.705002	25.89%	21.54%	23.11%	21.81%
GLNO	6	A	GL	9	C	43,713	0.241346	57,677	0.567631	31.18%	27.68%	29.52%	28.30%
GLNO	6	A	GL	12	C	43,713	0.189134	59,298	0.438400	37.73%	33.62%	35.98%	34.76%
GLNO	6	A	GL	Col	C	43,713	0.155495	60,154	0.334052	42.08%	43.52%	46.35%	45.93%
GLNO	6	A	GL	3	D	19,031	0.179687	39,939	0.560186	35.66%	34.40%	36.50%	36.13%
GLNO	6	A	GL	6	D	43,990	0.335922	54,552	0.703434	26.01%	21.55%	23.12%	21.87%
GLNO	6	A	GL	9	D	43,991	0.243109	57,853	0.567320	31.19%	27.50%	29.49%	28.18%
GLNO	6	A	GL	12	D	43,991	0.190477	59,494	0.440739	37.66%	33.39%	35.88%	34.53%
GLNO	6	A	GL	Col	D	43,991	0.156578	60,307	0.335738	41.70%	43.13%	46.18%	45.73%
GLNO	6	A	GLNO	9	A	87,468	0.636279	66,102	0.703370	26.39%	23.95%	25.75%	24.77%
GLNO	6	A	GLNO	12	A	87,468	0.466576	66,102	0.512049	35.83%	32.02%	34.50%	33.35%
GLNO	6	A	GLNO	Col	A	87,468	0.368336	66,102	0.377775	40.93%	42.87%	45.76%	45.37%
GLNO	6	A	GLNO	3	B	109	0.000873	35,473	0.466241	38.23%	36.26%	38.37%	37.90%
GLNO	6	A	GLNO	6	B	237	0.001357	45,439	0.523322	31.56%	26.64%	28.41%	26.95%
GLNO	6	A	GLNO	9	B	240	0.001068	50,768	0.463766	33.85%	29.88%	31.97%	30.61%
GLNO	6	A	GLNO	12	B	240	0.000874	53,618	0.378354	38.71%	34.81%	37.20%	35.86%
GLNO	6	A	GLNO	Col	B	240	0.000739	55,123	0.296457	42.43%	43.49%	46.59%	46.00%
GLNO	9	A	GL	6	250kB	43,616	0.134782	73,246	0.533400	29.79%	27.20%	28.58%	27.29%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GLNO	9	A	GL	9	250kB	68,733	0.221992	81,516	0.586239	26.01%	23.84%	25.14%	24.20%
GLNO	9	A	GL	3	A	18,891	0.121078	41,547	0.425852	40.28%	39.88%	42.50%	42.04%
GLNO	9	A	GL	6	A	43,954	0.242893	57,874	0.568010	30.87%	27.38%	29.22%	28.12%
GLNO	9	A	GL	9	A	68,925	0.334605	77,213	0.699121	23.16%	20.10%	21.40%	20.30%
GLNO	9	A	GL	12	A	68,925	0.269250	80,904	0.571962	27.93%	25.17%	26.86%	25.68%
GLNO	9	A	GL	Col	A	68,925	0.225253	82,810	0.447864	34.07%	36.48%	38.87%	38.37%
GLNO	9	A	GL	3	B	18,745	0.119970	41,304	0.422980	40.02%	39.86%	42.36%	41.96%
GLNO	9	A	GL	6	B	43,828	0.241923	57,820	0.567057	30.91%	27.48%	29.52%	28.22%
GLNO	9	A	GL	9	B	68,843	0.333946	77,107	0.698174	23.20%	20.03%	21.41%	20.34%
GLNO	9	A	GL	12	B	68,843	0.268760	80,899	0.570225	28.19%	24.88%	26.63%	25.63%
GLNO	9	A	GL	Col	B	68,843	0.224867	82,792	0.445933	34.16%	36.57%	38.84%	38.41%
GLNO	9	A	GL	3	C	18,666	0.119524	41,356	0.423873	40.60%	39.77%	42.38%	42.09%
GLNO	9	A	GL	6	C	43,713	0.241346	57,766	0.566922	30.80%	27.21%	29.14%	27.94%
GLNO	9	A	GL	9	C	68,661	0.333025	76,922	0.697756	23.27%	19.82%	21.06%	20.04%
GLNO	9	A	GL	12	C	68,661	0.268025	80,706	0.569439	28.31%	25.10%	26.70%	25.66%
GLNO	9	A	GL	Col	C	68,661	0.224255	82,644	0.445614	34.37%	36.67%	38.93%	38.62%
GLNO	9	A	GL	3	D	19,032	0.122070	41,510	0.425299	40.34%	39.95%	42.53%	42.05%
GLNO	9	A	GL	6	D	43,994	0.243129	57,834	0.566190	31.23%	27.70%	29.58%	28.45%
GLNO	9	A	GL	9	D	69,048	0.335355	77,033	0.696041	23.88%	20.10%	21.66%	20.50%
GLNO	9	A	GL	12	D	69,048	0.269829	80,802	0.570813	28.36%	25.14%	26.89%	25.78%
GLNO	9	A	GL	Col	D	69,048	0.225725	82,742	0.447092	33.82%	36.33%	38.76%	38.37%
GLNO	9	A	GLNO	12	A	137,468	0.733288	93,979	0.727995	23.90%	21.03%	22.81%	21.96%
GLNO	9	A	GLNO	Col	A	137,468	0.578891	93,979	0.537093	32.28%	35.46%	37.71%	37.46%
GLNO	9	A	GLNO	3	B	110	0.000629	38,345	0.379323	41.63%	40.77%	43.38%	42.95%
GLNO	9	A	GLNO	6	B	241	0.001073	50,963	0.466775	33.74%	29.79%	31.71%	30.41%
GLNO	9	A	GLNO	9	B	439	0.001599	63,758	0.512705	28.86%	25.06%	26.61%	25.41%
GLNO	9	A	GLNO	12	B	439	0.001353	69,855	0.455515	31.00%	28.08%	29.77%	28.51%
GLNO	9	A	GLNO	Col	B	439	0.001172	73,163	0.373708	35.33%	37.27%	39.80%	39.20%
GLNO	12	A	GL	3	A	18,891	0.091694	42,318	0.320822	45.08%	43.41%	46.74%	46.09%
GLNO	12	A	GL	6	A	43,954	0.190310	59,412	0.438578	37.49%	33.32%	35.83%	34.65%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GLNO	12	A	GL	9	A	68,928	0.269265	80,840	0.569577	27.92%	24.67%	26.52%	25.50%
GLNO	12	A	GL	12	A	93,875	0.334028	104,394	0.681984	20.30%	19.12%	20.15%	19.10%
GLNO	12	A	GL	Col	A	93,876	0.283581	108,737	0.560249	26.04%	30.12%	31.76%	31.58%
GLNO	12	A	GL	3	B	18,745	0.090886	42,122	0.319237	44.73%	43.40%	46.57%	46.01%
GLNO	12	A	GL	6	B	43,829	0.189601	59,411	0.438496	37.65%	33.28%	36.00%	34.69%
GLNO	12	A	GL	9	B	68,844	0.268765	80,799	0.569557	28.27%	24.71%	26.64%	25.62%
GLNO	12	A	GL	12	B	93,763	0.333403	104,515	0.681457	20.92%	18.74%	20.08%	19.13%
GLNO	12	A	GL	Col	B	93,763	0.283075	108,812	0.558715	26.11%	30.19%	31.73%	31.67%
GLNO	12	A	GL	3	C	18,666	0.090537	42,146	0.319552	45.29%	43.30%	46.59%	46.09%
GLNO	12	A	GL	6	C	43,713	0.189134	59,336	0.438105	37.53%	33.19%	35.81%	34.49%
GLNO	12	A	GL	9	C	68,662	0.268030	80,475	0.567513	28.18%	24.59%	26.44%	25.47%
GLNO	12	A	GL	12	C	93,890	0.334194	104,356	0.681206	20.74%	18.98%	20.16%	19.29%
GLNO	12	A	GL	Col	C	93,892	0.283710	108,744	0.559167	26.34%	30.34%	31.75%	31.89%
GLNO	12	A	GL	3	D	19,032	0.092428	42,334	0.320975	44.97%	43.54%	46.74%	46.08%
GLNO	12	A	GL	6	D	43,994	0.190492	59,423	0.437993	37.81%	33.62%	36.08%	34.91%
GLNO	12	A	GL	9	D	69,051	0.269844	80,711	0.567951	28.36%	24.55%	26.73%	25.70%
GLNO	12	A	GL	12	D	94,220	0.335633	104,263	0.680528	20.61%	18.89%	20.00%	19.00%
GLNO	12	A	GL	Col	D	94,221	0.284895	108,538	0.558366	25.82%	29.82%	31.56%	31.54%
GLNO	12	A	GLNO	Col	A	187,468	0.789445	129,093	0.737771	22.36%	27.78%	29.13%	29.41%
GLNO	12	A	GLNO	3	B	110	0.000489	39,891	0.296244	45.71%	43.90%	47.18%	46.63%
GLNO	12	A	GLNO	6	B	242	0.000881	53,790	0.380227	38.99%	34.43%	37.03%	35.71%
GLNO	12	A	GLNO	9	B	442	0.001362	69,843	0.455344	31.49%	27.27%	29.32%	28.18%
GLNO	12	A	GLNO	12	B	612	0.001635	84,572	0.486743	25.88%	23.75%	25.14%	23.94%
GLNO	12	A	GLNO	Col	B	613	0.001445	91,491	0.430420	28.51%	31.61%	33.56%	33.25%
GLNO	12	A	Large	6	250kB	43,616	0.116744	82,036	0.501311	34.57%	30.49%	32.61%	31.49%
GLNO	12	A	Large	9	250kB	68,739	0.191147	91,148	0.553987	25.91%	23.89%	25.39%	24.83%
GLNO	Col	A	GL	6	250kB	43,616	0.102964	86,751	0.423564	39.54%	40.55%	43.13%	42.68%
GLNO	Col	A	GL	9	250kB	68,739	0.167814	96,299	0.469147	31.72%	34.03%	36.27%	35.93%
GLNO	Col	A	GL	3	A	18,891	0.073786	42,661	0.240417	47.85%	50.30%	54.05%	53.77%
GLNO	Col	A	GL	6	A	43,954	0.156442	60,240	0.333701	41.77%	43.31%	46.18%	45.80%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GLNO	Col	A	GL	9	A	68,928	0.225265	82,777	0.445332	34.02%	36.69%	39.01%	38.82%
GLNO	Col	A	GL	12	A	93,875	0.283577	108,638	0.557936	25.77%	30.38%	32.02%	32.04%
GLNO	Col	A	GL	Col	A	119,003	0.334362	138,729	0.660680	18.61%	18.65%	19.84%	19.08%
GLNO	Col	A	GL	3	B	18,745	0.073152	42,478	0.239348	47.92%	50.44%	54.09%	53.86%
GLNO	Col	A	GL	6	B	43,829	0.155884	60,237	0.333638	41.95%	43.29%	46.47%	45.87%
GLNO	Col	A	GL	9	B	68,844	0.224871	82,733	0.445249	34.26%	36.66%	39.07%	38.74%
GLNO	Col	A	GL	12	B	93,764	0.283079	108,833	0.558301	26.29%	30.31%	31.91%	31.84%
GLNO	Col	A	GL	Col	B	118,864	0.333767	138,890	0.659622	18.88%	18.72%	19.93%	19.12%
GLNO	Col	A	GL	3	C	18,666	0.072866	42,491	0.239480	47.98%	50.10%	53.87%	53.66%
GLNO	Col	A	GL	6	C	43,713	0.155495	60,135	0.333115	41.82%	43.30%	46.33%	45.70%
GLNO	Col	A	GL	9	C	68,662	0.224259	82,343	0.443136	34.07%	36.55%	38.90%	38.53%
GLNO	Col	A	GL	12	C	93,890	0.283703	108,582	0.557257	25.96%	30.13%	31.86%	31.82%
GLNO	Col	A	GL	Col	C	118,904	0.334065	138,795	0.659961	18.49%	18.61%	19.70%	18.84%
GLNO	Col	A	GL	3	D	19,032	0.074370	42,691	0.240622	47.74%	50.40%	54.10%	53.80%
GLNO	Col	A	GL	6	D	43,994	0.156591	60,271	0.333529	41.97%	43.61%	46.56%	46.18%
GLNO	Col	A	GL	9	D	69,051	0.225737	82,677	0.444435	34.03%	36.61%	39.23%	39.01%
GLNO	Col	A	GL	12	D	94,222	0.284899	108,601	0.557629	25.76%	30.27%	31.96%	31.96%
GLNO	Col	A	GL	Col	D	119,290	0.335411	138,516	0.658687	17.98%	18.51%	19.69%	18.70%
GLNO	Col	A	GLNO	3	B	110	0.000400	40,569	0.225556	48.39%	50.67%	54.45%	54.20%
GLNO	Col	A	GLNO	6	B	242	0.000745	55,315	0.297669	42.87%	43.89%	46.92%	46.32%
GLNO	Col	A	GLNO	9	B	442	0.001180	73,226	0.373819	36.15%	37.80%	40.29%	39.94%
GLNO	Col	A	GLNO	12	B	612	0.001442	91,706	0.431556	29.00%	32.41%	34.19%	34.01%
GLNO	Col	A	GLNO	Col	B	838	0.001768	109,200	0.453607	23.42%	23.44%	24.98%	23.89%
GLNO	3	B	GL	6	250kB	18,825	0.075786	42,566	0.356278	37.48%	36.13%	38.48%	38.31%
GLNO	3	B	GL	9	250kB	18,737	0.072172	42,606	0.329172	42.13%	41.84%	44.54%	44.21%
GLNO	3	B	GL	3	A	18,663	0.331781	37,099	0.693634	30.85%	25.45%	27.22%	25.98%
GLNO	3	B	GL	6	A	18,664	0.175661	39,948	0.560359	35.29%	34.88%	37.00%	36.61%
GLNO	3	B	GL	9	A	18,664	0.119450	41,613	0.426721	40.18%	40.09%	42.66%	42.28%
GLNO	3	B	GL	12	A	18,664	0.090492	42,533	0.323947	44.84%	43.49%	46.64%	46.14%
GLNO	3	B	GL	Col	A	18,664	0.072835	42,883	0.243236	47.96%	50.31%	54.11%	53.85%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GLNO	3	B	GL	3	B	18,901	0.336964	37,220	0.699506	31.27%	25.45%	27.34%	26.05%
GLNO	3	B	GL	6	B	18,903	0.178179	40,084	0.563167	35.19%	34.85%	36.91%	36.55%
GLNO	3	B	GL	9	B	18,903	0.121103	41,698	0.428441	40.17%	40.05%	42.62%	42.40%
GLNO	3	B	GL	12	B	18,903	0.091722	42,611	0.323706	44.94%	43.39%	46.62%	46.15%
GLNO	3	B	GL	Col	B	18,903	0.073814	42,951	0.242694	47.96%	50.39%	54.07%	53.86%
GLNO	3	B	GL	3	C	18,779	0.335004	37,086	0.695641	31.68%	25.60%	27.48%	26.18%
GLNO	3	B	GL	6	C	18,779	0.177067	40,032	0.563014	35.65%	34.53%	36.66%	36.46%
GLNO	3	B	GL	9	C	18,779	0.120335	41,609	0.428826	40.62%	40.24%	42.78%	42.43%
GLNO	3	B	GL	12	C	18,779	0.091135	42,533	0.323748	45.20%	43.69%	46.73%	46.24%
GLNO	3	B	GL	Col	C	18,779	0.073339	42,892	0.242756	48.16%	50.51%	54.14%	54.02%
GLNO	3	B	GL	3	D	18,564	0.329272	37,158	0.695465	31.36%	25.78%	27.50%	26.13%
GLNO	3	B	GL	6	D	18,564	0.174508	40,121	0.562439	35.29%	34.66%	36.78%	36.40%
GLNO	3	B	GL	9	D	18,564	0.118712	41,659	0.427175	40.34%	40.11%	42.68%	42.31%
GLNO	3	B	GL	12	D	18,564	0.089951	42,521	0.323817	44.86%	43.34%	46.54%	45.97%
GLNO	3	B	GL	Col	D	18,564	0.072408	42,858	0.242923	47.72%	50.24%	54.07%	53.89%
GLNO	3	B	GLNO	6	B	37,468	0.428362	45,454	0.686980	31.89%	32.63%	34.69%	34.66%
GLNO	3	B	GLNO	9	B	37,468	0.272558	45,454	0.482860	38.82%	39.21%	41.71%	41.45%
GLNO	3	B	GLNO	12	B	37,468	0.199863	45,454	0.351729	43.92%	43.07%	46.12%	45.64%
GLNO	3	B	GLNO	Col	B	37,468	0.157781	45,454	0.259797	47.24%	49.82%	53.59%	53.46%
GLNO	6	B	GL	6	250kB	43,748	0.159972	59,363	0.481108	30.86%	28.06%	29.53%	28.26%
GLNO	6	B	GL	9	250kB	43,702	0.153529	59,830	0.450117	34.48%	31.74%	33.63%	32.73%
GLNO	6	B	GL	3	A	18,664	0.175661	39,945	0.559846	34.59%	34.57%	36.76%	36.42%
GLNO	6	B	GL	6	A	43,656	0.332597	54,377	0.700987	25.69%	21.58%	23.10%	21.88%
GLNO	6	B	GL	9	A	43,658	0.240864	57,831	0.566909	30.91%	27.60%	29.41%	28.16%
GLNO	6	B	GL	12	A	43,658	0.188786	59,519	0.440813	37.63%	33.28%	35.68%	34.49%
GLNO	6	B	GL	Col	A	43,658	0.155225	60,357	0.336178	42.06%	43.06%	46.11%	45.49%
GLNO	6	B	GL	3	B	18,902	0.178168	39,950	0.561174	35.02%	34.57%	36.72%	36.40%
GLNO	6	B	GL	6	B	43,980	0.335692	54,627	0.706286	25.63%	21.71%	23.22%	21.81%
GLNO	6	B	GL	9	B	43,981	0.242973	57,943	0.569235	31.02%	27.44%	29.30%	28.26%
GLNO	6	B	GL	12	B	43,981	0.190384	59,607	0.440392	37.81%	33.15%	35.67%	34.59%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GLNO	6	B	GL	Col	B	43,981	0.156509	60,403	0.335135	42.20%	43.23%	46.12%	45.60%
GLNO	6	B	GL	3	C	18,780	0.177078	39,895	0.560213	35.63%	34.47%	36.71%	36.51%
GLNO	6	B	GL	6	C	43,597	0.332198	54,476	0.704097	25.99%	21.77%	23.52%	22.18%
GLNO	6	B	GL	9	C	43,598	0.240558	57,770	0.568714	31.38%	27.61%	29.59%	28.27%
GLNO	6	B	GL	12	C	43,598	0.188542	59,454	0.439856	38.04%	33.41%	35.78%	34.57%
GLNO	6	B	GL	Col	C	43,598	0.155022	60,312	0.335106	42.45%	43.32%	46.24%	45.79%
GLNO	6	B	GL	3	D	18,565	0.174519	39,926	0.559407	35.25%	34.74%	36.87%	36.50%
GLNO	6	B	GL	6	D	43,755	0.333529	54,754	0.707306	25.70%	21.63%	23.16%	21.91%
GLNO	6	B	GL	9	D	43,756	0.241496	58,083	0.570509	31.17%	27.47%	29.23%	28.00%
GLNO	6	B	GL	12	D	43,757	0.189272	59,679	0.442509	37.75%	33.21%	35.66%	34.41%
GLNO	6	B	GL	Col	D	43,757	0.155616	60,475	0.336871	41.96%	43.02%	46.12%	45.59%
GLNO	6	B	GLNO	9	B	87,468	0.636279	66,165	0.702874	26.96%	24.08%	25.80%	24.76%
GLNO	6	B	GLNO	12	B	87,468	0.466576	66,165	0.511994	36.02%	31.95%	34.31%	33.18%
GLNO	6	B	GLNO	Col	B	87,468	0.368336	66,165	0.378172	41.03%	42.09%	45.18%	44.69%
GLNO	9	B	GL	6	250kB	43,750	0.135252	73,368	0.534157	29.63%	27.13%	28.65%	27.46%
GLNO	9	B	GL	9	250kB	68,615	0.221527	81,722	0.587932	25.91%	23.66%	24.98%	24.31%
GLNO	9	B	GL	3	A	18,664	0.119450	41,454	0.423817	39.92%	39.93%	42.56%	42.08%
GLNO	9	B	GL	6	A	43,659	0.240871	57,559	0.562319	30.61%	27.38%	29.29%	28.14%
GLNO	9	B	GL	9	A	68,797	0.333776	77,154	0.697229	22.79%	20.01%	21.27%	20.30%
GLNO	9	B	GL	12	A	68,797	0.268616	80,880	0.571065	28.44%	24.72%	26.49%	25.49%
GLNO	9	B	GL	Col	A	68,797	0.224741	82,841	0.447729	34.65%	36.33%	38.82%	38.53%
GLNO	9	B	GL	3	B	18,902	0.121096	41,439	0.424271	40.00%	39.89%	42.46%	42.01%
GLNO	9	B	GL	6	B	43,982	0.242980	57,745	0.565042	31.10%	27.30%	29.47%	28.13%
GLNO	9	B	GL	9	B	69,041	0.335229	77,191	0.698479	23.24%	20.09%	21.44%	20.55%
GLNO	9	B	GL	12	B	69,041	0.269742	80,922	0.569853	28.55%	24.38%	26.31%	25.46%
GLNO	9	B	GL	Col	B	69,041	0.225660	82,811	0.445707	34.67%	36.44%	38.84%	38.61%
GLNO	9	B	GL	3	C	18,780	0.120342	41,391	0.423706	40.80%	39.95%	42.54%	42.18%
GLNO	9	B	GL	6	C	43,599	0.240565	57,679	0.564722	31.07%	27.58%	29.65%	28.33%
GLNO	9	B	GL	9	C	68,602	0.332643	76,799	0.694882	23.58%	20.17%	21.60%	20.42%
GLNO	9	B	GL	12	C	68,602	0.267733	80,648	0.568172	28.95%	24.84%	26.52%	25.58%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GLNO	9	B	GL	Col	C	68,602	0.224019	82,637	0.445185	35.04%	36.64%	38.93%	38.84%
GLNO	9	B	GL	3	D	18,565	0.118719	41,485	0.424256	40.20%	40.09%	42.64%	42.08%
GLNO	9	B	GL	6	D	43,758	0.241510	57,915	0.566567	30.98%	27.54%	29.56%	28.39%
GLNO	9	B	GL	9	D	68,806	0.333788	77,433	0.701202	23.40%	20.12%	21.44%	20.41%
GLNO	9	B	GL	12	D	68,808	0.268640	81,162	0.574184	28.61%	24.74%	26.59%	25.60%
GLNO	9	B	GL	Col	D	68,808	0.224764	83,020	0.448890	34.53%	36.30%	38.85%	38.65%
GLNO	9	B	GLNO	12	B	137,468	0.733288	94,135	0.728430	24.27%	20.87%	22.54%	21.83%
GLNO	9	B	GLNO	Col	B	137,468	0.578891	94,135	0.538037	32.63%	34.65%	37.12%	37.07%
GLNO	12	B	GL	6	250kB	43,750	0.117144	82,214	0.502524	34.20%	30.60%	32.60%	31.54%
GLNO	12	B	GL	9	250kB	68,618	0.190746	91,389	0.555803	25.65%	23.99%	25.35%	24.78%
GLNO	12	B	GL	3	A	18,664	0.090492	42,274	0.320049	44.72%	43.59%	46.71%	46.02%
GLNO	12	B	GL	6	A	43,659	0.188792	59,174	0.435615	37.23%	33.44%	35.80%	34.62%
GLNO	12	B	GL	9	A	68,797	0.268616	80,845	0.569083	27.80%	25.20%	26.80%	25.78%
GLNO	12	B	GL	12	A	94,006	0.334650	104,238	0.679664	20.75%	19.03%	20.07%	19.24%
GLNO	12	B	GL	Col	A	94,007	0.284089	108,700	0.559556	26.21%	30.02%	31.68%	31.67%
GLNO	12	B	GL	3	B	18,902	0.091717	42,210	0.319785	44.54%	43.50%	46.54%	45.94%
GLNO	12	B	GL	6	B	43,982	0.190389	59,273	0.436592	37.56%	33.46%	36.00%	34.66%
GLNO	12	B	GL	9	B	69,041	0.269742	80,786	0.568863	28.44%	25.38%	27.00%	25.94%
GLNO	12	B	GL	12	B	94,168	0.335326	104,646	0.682284	20.76%	18.86%	20.12%	19.13%
GLNO	12	B	GL	Col	B	94,168	0.284646	108,984	0.559699	26.30%	30.16%	31.86%	31.87%
GLNO	12	B	GL	3	C	18,780	0.091141	42,207	0.319830	45.26%	43.48%	46.62%	46.02%
GLNO	12	B	GL	6	C	43,599	0.188548	59,293	0.437206	37.39%	33.47%	35.96%	34.60%
GLNO	12	B	GL	9	C	68,602	0.267733	80,468	0.566888	27.98%	25.17%	26.90%	25.79%
GLNO	12	B	GL	12	C	93,504	0.332363	104,194	0.678824	20.95%	18.95%	20.09%	19.20%
GLNO	12	B	GL	Col	C	93,504	0.282207	108,692	0.558357	26.63%	30.46%	31.90%	32.10%
GLNO	12	B	GL	3	D	18,565	0.089956	42,273	0.320032	44.80%	43.64%	46.68%	46.00%
GLNO	12	B	GL	6	D	43,758	0.189277	59,463	0.437975	37.69%	33.61%	35.98%	34.86%
GLNO	12	B	GL	9	D	68,806	0.268630	81,017	0.570783	28.41%	25.01%	26.85%	25.90%
GLNO	12	B	GL	12	D	93,752	0.333410	104,471	0.682202	20.81%	18.78%	20.05%	19.18%
GLNO	12	B	GL	Col	D	93,753	0.283079	108,775	0.559873	26.26%	29.91%	31.76%	31.91%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
GLNO	12	B	GLNO	Col	B	187,468	0.789445	129,230	0.738626	21.96%	26.84%	28.38%	28.81%
GLNO	Col	B	GL	6	250kB	43,750	0.103313	86,949	0.424977	39.38%	39.94%	42.66%	42.26%
GLNO	Col	B	GL	9	250kB	68,618	0.167470	96,495	0.470590	31.49%	33.34%	35.65%	35.36%
GLNO	Col	B	GL	3	A	18,664	0.072835	42,608	0.240069	47.70%	50.02%	53.78%	53.53%
GLNO	Col	B	GL	6	A	43,659	0.155229	59,978	0.331799	41.41%	42.69%	45.72%	45.29%
GLNO	Col	B	GL	9	A	68,797	0.224741	82,695	0.444735	33.82%	36.04%	38.53%	38.27%
GLNO	Col	B	GL	12	A	94,008	0.284093	108,366	0.555811	25.72%	29.50%	31.33%	31.39%
GLNO	Col	B	GL	Col	A	119,113	0.334774	138,294	0.657300	18.85%	18.49%	19.73%	18.96%
GLNO	Col	B	GL	3	B	18,902	0.073810	42,526	0.239706	47.84%	50.10%	53.80%	53.62%
GLNO	Col	B	GL	6	B	43,982	0.156513	60,052	0.332304	41.97%	42.69%	46.00%	45.36%
GLNO	Col	B	GL	9	B	69,041	0.225660	82,622	0.444427	34.30%	36.15%	38.63%	38.23%
GLNO	Col	B	GL	12	B	94,171	0.284658	108,752	0.557703	25.89%	29.54%	31.32%	31.19%
GLNO	Col	B	GL	Col	B	119,278	0.335319	138,992	0.660480	18.53%	18.71%	19.84%	18.97%
GLNO	Col	B	GL	3	C	18,780	0.073344	42,524	0.239734	47.94%	49.78%	53.62%	53.40%
GLNO	Col	B	GL	6	C	43,599	0.155026	60,069	0.332660	41.80%	42.75%	45.96%	45.24%
GLNO	Col	B	GL	9	C	68,602	0.224019	82,296	0.442811	33.79%	35.90%	38.53%	38.03%
GLNO	Col	B	GL	12	C	93,506	0.282215	108,378	0.555676	25.78%	29.32%	31.25%	31.17%
GLNO	Col	B	GL	Col	C	118,658	0.333143	138,563	0.658185	18.95%	18.88%	20.17%	19.31%
GLNO	Col	B	GL	3	D	18,565	0.072413	42,609	0.240072	47.67%	50.08%	53.80%	53.58%
GLNO	Col	B	GL	6	D	43,758	0.155620	60,283	0.333649	41.84%	42.96%	46.09%	45.66%
GLNO	Col	B	GL	9	D	68,806	0.224756	82,881	0.446062	33.82%	35.92%	38.68%	38.44%
GLNO	Col	B	GL	12	D	93,753	0.283079	108,611	0.557757	25.66%	29.44%	31.41%	31.40%
GLNO	Col	B	GL	Col	D	118,869	0.333832	138,712	0.660288	18.80%	18.55%	20.04%	19.12%
FixV	3	A	FixV	6	A	28,270	0.220398	33,069	1.000000	38.76%	36.11%	38.75%	N/A
FixV	3	A	FixV	12	A	28,270	0.123846	33,069	1.000000	45.82%	45.15%	48.83%	N/A
FixV	3	A	FixV	6	B	28,270	0.220391	33,069	1.000000	38.85%	36.25%	38.97%	N/A
FixV	3	A	FixV	12	B	28,270	0.123843	33,069	1.000000	45.91%	45.14%	48.90%	N/A
FixV	6	A	FixV	12	A	128,268	0.561918	33,069	1.000000	31.26%	30.20%	32.27%	N/A
FixV	6	A	FixV	6	B	31,001	0.137453	33,069	1.000000	26.17%	22.19%	23.67%	N/A
FixV	6	A	FixV	12	B	31,002	0.095233	33,069	1.000000	34.15%	32.33%	34.65%	N/A

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
FixV	12	A	FixV	12	B	31,211	0.073381	33,069	1.000000	23.30%	20.95%	22.48%	N/A
FixV	6	B	FixV	12	A	31,005	0.095243	33,069	1.000000	33.86%	31.85%	34.15%	N/A
FixV	6	B	FixV	12	B	128,272	0.561926	33,069	1.000000	31.45%	29.87%	32.12%	N/A
RTRC	GCat	1k	GL	3	A	0	0.000000	11,489	0.209868	74.27%	79.68%	80.39%	80.69%
RTRC	GCat	1k	GL	6	A	0	0.000000	13,361	0.181718	74.11%	79.10%	79.74%	80.26%
RTRC	GCat	1k	GL	9	A	0	0.000000	15,075	0.151196	73.79%	78.68%	79.26%	79.89%
RTRC	GCat	1k	GL	12	A	0	0.000000	16,166	0.121264	73.59%	78.26%	78.82%	79.53%
RTRC	GCat	1k	GL	Col	A	0	0.000000	17,055	0.095934	73.38%	77.63%	78.18%	78.87%
RTRC	GCat	1k	GL	3	B	0	0.000000	11,500	0.210708	74.37%	79.64%	80.38%	80.69%
RTRC	GCat	1k	GL	6	B	0	0.000000	13,349	0.181471	74.18%	79.13%	79.79%	80.27%
RTRC	GCat	1k	GL	9	B	0	0.000000	15,021	0.150736	73.80%	78.72%	79.28%	79.86%
RTRC	GCat	1k	GL	12	B	0	0.000000	16,172	0.120937	73.61%	78.29%	78.81%	79.51%
RTRC	GCat	1k	GL	Col	B	0	0.000000	17,084	0.095713	73.30%	77.71%	78.17%	78.85%
RTRC	GCat	1k	GL	3	C	0	0.000000	11,512	0.211094	74.46%	79.66%	80.32%	80.63%
RTRC	GCat	1k	GL	6	C	0	0.000000	13,421	0.182940	74.09%	79.18%	79.83%	80.28%
RTRC	GCat	1k	GL	9	C	0	0.000000	15,058	0.151748	73.70%	78.70%	79.31%	79.87%
RTRC	GCat	1k	GL	12	C	0	0.000000	16,206	0.121527	73.52%	78.22%	78.81%	79.50%
RTRC	GCat	1k	GL	Col	C	0	0.000000	17,093	0.095955	73.25%	77.66%	78.18%	78.85%
RTRC	GCat	1k	GL	3	D	0	0.000000	11,484	0.209746	74.35%	79.67%	80.34%	80.66%
RTRC	GCat	1k	GL	6	D	0	0.000000	13,362	0.181199	74.20%	79.08%	79.74%	80.27%
RTRC	GCat	1k	GL	9	D	0	0.000000	15,042	0.150740	73.84%	78.65%	79.28%	79.90%
RTRC	GCat	1k	GL	12	D	0	0.000000	16,193	0.121488	73.57%	78.27%	78.82%	79.53%
RTRC	GCat	1k	GL	Col	D	0	0.000000	17,068	0.095960	73.25%	77.67%	78.18%	78.86%
RTRC	GCat	1k	RTRC	GCat150k	B	763	0.005079	20,953	0.092752	72.44%	72.93%	73.72%	75.05%
RTRC	GCat	1k	RTRC	GCat		1,000	0.004258	21,103	0.077374	72.55%	72.93%	73.74%	75.07%
RTRC	GCat	1k	RTRC	GCat150k		1,000	0.006667	21,103	0.097804	72.37%	72.84%	73.65%	75.00%
RTRC	GCat150k		GL	3	A	0	0.000000	31,321	0.136429	71.84%	72.71%	76.12%	75.38%
RTRC	GCat150k		GL	6	A	0	0.000000	40,341	0.167243	71.06%	71.50%	74.54%	74.00%
RTRC	GCat150k		GL	9	A	0	0.000000	51,007	0.197366	70.46%	70.37%	73.25%	72.85%
RTRC	GCat150k		GL	12	A	0	0.000000	60,578	0.213629	70.02%	69.36%	72.04%	71.85%

Table A – 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
RTRC	GCat150k		GL	Col	A	0	0.000000	71,033	0.223047	69.55%	67.43%	69.86%	69.49%
RTRC	GCat150k		GL	3	B	0	0.000000	31,156	0.135704	71.92%	72.70%	76.10%	75.35%
RTRC	GCat150k		GL	6	B	0	0.000000	40,267	0.166870	71.12%	71.47%	74.66%	74.00%
RTRC	GCat150k		GL	9	B	0	0.000000	50,862	0.196776	70.51%	70.35%	73.32%	72.77%
RTRC	GCat150k		GL	12	B	0	0.000000	60,674	0.213726	70.13%	69.37%	72.06%	71.78%
RTRC	GCat150k		GL	Col	B	0	0.000000	71,056	0.222616	69.63%	67.44%	69.83%	69.44%
RTRC	GCat150k		GL	3	C	0	0.000000	31,256	0.136217	72.08%	72.61%	75.91%	75.17%
RTRC	GCat150k		GL	6	C	0	0.000000	40,461	0.167896	71.20%	71.67%	74.74%	74.02%
RTRC	GCat150k		GL	9	C	0	0.000000	50,783	0.196703	70.58%	70.46%	73.31%	72.80%
RTRC	GCat150k		GL	12	C	0	0.000000	60,696	0.214073	70.11%	69.40%	72.00%	71.77%
RTRC	GCat150k		GL	Col	C	0	0.000000	71,233	0.223538	69.68%	67.47%	69.82%	69.44%
RTRC	GCat150k		GL	3	D	0	0.000000	31,254	0.136095	72.06%	72.55%	75.98%	75.28%
RTRC	GCat150k		GL	6	D	0	0.000000	40,361	0.167189	71.26%	71.41%	74.64%	74.15%
RTRC	GCat150k		GL	9	D	0	0.000000	50,867	0.196679	70.58%	70.36%	73.34%	72.91%
RTRC	GCat150k		GL	12	D	0	0.000000	60,656	0.213960	70.02%	69.33%	72.10%	71.84%
RTRC	GCat150k		GL	Col	D	0	0.000000	71,088	0.223189	69.44%	67.49%	69.88%	69.47%
RTRC	GCat150k	B	GL	3	A	0	0.000000	31,530	0.131730	71.80%	72.79%	76.11%	75.49%
RTRC	GCat150k	B	GL	6	A	0	0.000000	40,735	0.162418	70.97%	71.59%	74.55%	74.13%
RTRC	GCat150k	B	GL	9	A	0	0.000000	51,572	0.192534	70.32%	70.42%	73.27%	72.99%
RTRC	GCat150k	B	GL	12	A	0	0.000000	61,376	0.209651	69.85%	69.44%	72.10%	72.03%
RTRC	GCat150k	B	GL	Col	A	0	0.000000	72,130	0.220342	69.36%	67.45%	69.88%	69.63%
RTRC	GCat150k	B	GL	3	B	0	0.000000	31,350	0.130964	71.82%	72.81%	76.10%	75.46%
RTRC	GCat150k	B	GL	6	B	0	0.000000	40,670	0.162103	71.00%	71.57%	74.68%	74.13%
RTRC	GCat150k	B	GL	9	B	0	0.000000	51,476	0.192184	70.36%	70.43%	73.34%	72.91%
RTRC	GCat150k	B	GL	12	B	0	0.000000	61,551	0.210075	69.95%	69.45%	72.11%	71.96%
RTRC	GCat150k	B	GL	Col	B	0	0.000000	72,286	0.220424	69.44%	67.46%	69.86%	69.58%
RTRC	GCat150k	B	GL	3	C	0	0.000000	31,482	0.131605	72.03%	72.69%	75.89%	75.29%
RTRC	GCat150k	B	GL	6	C	0	0.000000	40,873	0.163125	71.10%	71.76%	74.76%	74.16%
RTRC	GCat150k	B	GL	9	C	0	0.000000	51,383	0.192046	70.43%	70.52%	73.34%	72.94%
RTRC	GCat150k	B	GL	12	C	0	0.000000	61,560	0.210354	69.94%	69.49%	72.07%	71.95%

Table A - 3 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	Doc Overlap	DO Ratio	Term Overlap	TO Ratio	NICHD04 TC%	T-500 TC%	T-1000 TC%	T-5000 TC%
RTRC	GCat150k	B	GL	Col	C	0	0.000000	72,404	0.221097	69.49%	67.50%	69.84%	69.57%
RTRC	GCat150k	B	GL	3	D	0	0.000000	31,491	0.131544	71.97%	72.61%	75.95%	75.39%
RTRC	GCat150k	B	GL	6	D	0	0.000000	40,788	0.162523	71.16%	71.52%	74.68%	74.29%
RTRC	GCat150k	B	GL	9	D	0	0.000000	51,482	0.192098	70.46%	70.43%	73.37%	73.05%
RTRC	GCat150k	B	GL	12	D	0	0.000000	61,496	0.210144	69.87%	69.43%	72.17%	72.02%
RTRC	GCat150k	B	GL	Col	D	0	0.000000	72,291	0.220875	69.26%	67.50%	69.90%	69.60%
RTRC	GCat		GL	3	A	0	0.000000	32,641	0.114438	71.84%	72.83%	76.15%	75.47%
RTRC	GCat		GL	6	A	0	0.000000	42,502	0.143578	71.01%	71.62%	74.60%	74.12%
RTRC	GCat		GL	9	A	0	0.000000	54,252	0.173793	70.37%	70.48%	73.34%	73.01%
RTRC	GCat		GL	12	A	0	0.000000	65,091	0.193710	69.94%	69.52%	72.17%	72.05%
RTRC	GCat		GL	Col	A	0	0.000000	77,253	0.209235	69.45%	67.56%	69.97%	69.67%
RTRC	GCat		GL	3	B	0	0.000000	32,466	0.113817	71.88%	72.85%	76.14%	75.44%
RTRC	GCat		GL	6	B	0	0.000000	42,426	0.143273	71.06%	71.60%	74.73%	74.12%
RTRC	GCat		GL	9	B	0	0.000000	54,136	0.173417	70.40%	70.48%	73.41%	72.92%
RTRC	GCat		GL	12	B	0	0.000000	65,261	0.194073	70.03%	69.52%	72.18%	71.98%
RTRC	GCat		GL	Col	B	0	0.000000	77,355	0.209148	69.52%	67.55%	69.94%	69.62%
RTRC	GCat		GL	3	C	0	0.000000	32,606	0.114376	72.08%	72.74%	75.93%	75.28%
RTRC	GCat		GL	6	C	0	0.000000	42,636	0.144146	71.15%	71.79%	74.81%	74.15%
RTRC	GCat		GL	9	C	0	0.000000	54,042	0.173277	70.50%	70.59%	73.40%	72.96%
RTRC	GCat		GL	12	C	0	0.000000	65,303	0.194417	70.02%	69.57%	72.14%	71.97%
RTRC	GCat		GL	Col	C	0	0.000000	77,564	0.210029	69.58%	67.61%	69.92%	69.61%
RTRC	GCat		GL	3	D	0	0.000000	32,594	0.114253	72.02%	72.67%	75.99%	75.38%
RTRC	GCat		GL	6	D	0	0.000000	42,535	0.143600	71.22%	71.56%	74.72%	74.27%
RTRC	GCat		GL	9	D	0	0.000000	54,125	0.173288	70.53%	70.50%	73.43%	73.07%
RTRC	GCat		GL	12	D	0	0.000000	65,201	0.194099	69.95%	69.51%	72.23%	72.04%
RTRC	GCat		GL	Col	D	0	0.000000	77,383	0.209604	69.35%	67.61%	69.98%	69.64%

Table A - 4: Space comparison results for all experiments detailing OTV-Norm, document centroid comparison cosine, and the term centroid comparison cosine for each two of the four anchor sets used: NICHD04, and T-1000. Group column values: GL = grade level, GLNO = grade level non-overlapping, Large = large size 6th and 9th grade sets, FixV = fixed vocabulary sets. Set column indicates the grade level or specific RTRC set used in the comparison.

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
RTRC	GCat		RTRC	GCat	mod	0.0047	1.000000	1.000000	0.0047	1.000000	1.000000
RTRC	GCat150k	B	RTRC	GCat		12.6175	0.998052	0.987450	12.5118	0.998500	0.988970
RTRC	GCat150k		RTRC	GCat150k	B	13.4298	0.998601	0.988706	13.2160	0.997741	0.989958
RTRC	GCat150k		RTRC	GCat		13.8142	0.997246	0.989258	13.5557	0.996395	0.988422
GL	3	A	GL	6	A	14.5657	0.997645	0.990717	14.0020	0.998891	0.992751
GL	3	A	GL	9	A	15.4048	0.994341	0.990329	14.6754	0.998269	0.992701
GL	3	A	GL	12	A	15.8553	0.994413	0.985740	15.1638	0.995337	0.984257
GL	3	A	GL	Col	A	16.3239	0.991154	0.982967	15.6898	0.986270	0.963192
GL	3	A	GL	3	B	14.8668	0.998322	0.993368	14.0877	0.999559	0.995073
GL	3	A	GL	6	B	15.5298	0.996312	0.990861	14.5921	0.999183	0.994809
GL	3	A	GL	9	B	15.8783	0.992567	0.981781	14.9511	0.997035	0.989724
GL	3	A	GL	12	B	16.2008	0.994599	0.977427	15.3641	0.992587	0.969071
GL	3	A	GL	Col	B	16.5466	0.992433	0.981605	15.8488	0.986804	0.970415
GL	3	A	GL	3	C	14.9064	0.995453	0.988668	14.1409	0.999371	0.995084
GL	3	A	GL	6	C	15.5140	0.994903	0.991156	14.5711	0.999167	0.995593
GL	3	A	GL	9	C	15.8753	0.996046	0.985897	14.9874	0.997431	0.989659
GL	3	A	GL	12	C	16.2065	0.994464	0.975545	15.3570	0.993879	0.975284
GL	3	A	GL	Col	C	16.5631	0.994805	0.983155	15.8524	0.980036	0.956458
GL	3	A	GL	3	D	14.9263	0.998370	0.992051	14.1460	0.999389	0.992349
GL	3	A	GL	6	D	15.5655	0.997193	0.989212	14.6519	0.999185	0.991337
GL	3	A	GL	9	D	15.9441	0.993379	0.980898	15.0483	0.998241	0.990083
GL	3	A	GL	12	D	16.2551	0.995057	0.983128	15.4293	0.994885	0.979987
GL	3	A	GL	Col	D	16.5682	0.994264	0.980736	15.8312	0.980720	0.935850
GL	6	A	GL	9	A	13.6549	0.999298	0.992572	13.3263	0.999502	0.992658
GL	6	A	GL	12	A	14.4929	0.998413	0.985049	14.0261	0.998548	0.984967

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GL	6	A	GL	Col	A	15.2842	0.996482	0.988559	14.7682	0.993778	0.982728
GL	6	A	GL	3	B	15.5753	0.997940	0.992213	14.6185	0.998815	0.991750
GL	6	A	GL	6	B	13.8256	0.998715	0.992053	13.3240	0.999680	0.993299
GL	6	A	GL	9	B	14.3304	0.997881	0.992067	13.7081	0.998998	0.994312
GL	6	A	GL	12	B	14.9963	0.993525	0.981377	14.3227	0.997235	0.985369
GL	6	A	GL	Col	B	15.5358	0.992017	0.987708	14.9535	0.987491	0.982242
GL	6	A	GL	3	C	15.5324	0.995689	0.988967	14.6471	0.998550	0.993351
GL	6	A	GL	6	C	13.9527	0.999332	0.993114	13.4354	0.999749	0.993699
GL	6	A	GL	9	C	14.2924	0.997345	0.992597	13.6954	0.998559	0.994262
GL	6	A	GL	12	C	14.9504	0.995730	0.986688	14.2794	0.996661	0.986139
GL	6	A	GL	Col	C	15.5239	0.994971	0.989280	14.9223	0.992766	0.985178
GL	6	A	GL	3	D	15.5727	0.997516	0.990812	14.6458	0.998765	0.990688
GL	6	A	GL	6	D	13.9023	0.998382	0.989922	13.3850	0.999589	0.989783
GL	6	A	GL	9	D	14.3504	0.997508	0.988982	13.7737	0.999317	0.991482
GL	6	A	GL	12	D	15.0708	0.997052	0.988646	14.4712	0.997680	0.984893
GL	6	A	GL	Col	D	15.4478	0.992867	0.986225	14.8591	0.993388	0.978553
GL	9	A	GL	12	A	13.1684	0.999651	0.992535	12.9414	0.999422	0.992962
GL	9	A	GL	Col	A	14.0617	0.998353	0.990563	13.7416	0.996406	0.988217
GL	9	A	GL	3	B	15.9499	0.996209	0.982746	15.0557	0.995797	0.980144
GL	9	A	GL	6	B	14.3679	0.998611	0.985543	13.7244	0.998376	0.984267
GL	9	A	GL	9	B	13.4735	0.999269	0.982499	13.0719	0.999458	0.982957
GL	9	A	GL	12	B	13.9393	0.998386	0.978646	13.4783	0.998253	0.978270
GL	9	A	GL	Col	B	14.4653	0.995857	0.985178	14.0041	0.994160	0.982954
GL	9	A	GL	3	C	15.8921	0.996631	0.979615	15.0103	0.995448	0.974701
GL	9	A	GL	6	C	14.5342	0.997811	0.985504	13.8653	0.998630	0.984883
GL	9	A	GL	9	C	13.3455	0.999063	0.982642	12.9856	0.999407	0.983709
GL	9	A	GL	12	C	13.9556	0.998188	0.979715	13.5068	0.998457	0.982287
GL	9	A	GL	Col	C	14.5297	0.998592	0.987218	14.0666	0.993478	0.983585
GL	9	A	GL	3	D	15.9400	0.996375	0.991103	15.0113	0.996049	0.988965

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GL	9	A	GL	6	D	14.4148	0.999074	0.991870	13.7811	0.999141	0.991258
GL	9	A	GL	9	D	13.4060	0.999398	0.996170	13.0511	0.999835	0.996503
GL	9	A	GL	12	D	13.6957	0.998490	0.994735	13.2862	0.999039	0.995654
GL	9	A	GL	Col	D	14.3674	0.997675	0.989800	13.9042	0.994928	0.985164
GL	12	A	GL	Col	A	13.0769	0.999536	0.995998	12.9123	0.999439	0.995546
GL	12	A	GL	3	B	16.2604	0.993198	0.972478	15.4217	0.987601	0.960172
GL	12	A	GL	6	B	15.0582	0.998031	0.982224	14.4400	0.993372	0.972107
GL	12	A	GL	9	B	14.1334	0.998318	0.979009	13.6879	0.997787	0.977492
GL	12	A	GL	12	B	13.2788	0.998747	0.974842	12.9588	0.999127	0.976712
GL	12	A	GL	Col	B	13.6608	0.998190	0.978469	13.3173	0.998233	0.978618
GL	12	A	GL	3	C	16.2160	0.995924	0.969499	15.3871	0.987290	0.950797
GL	12	A	GL	6	C	15.0304	0.996312	0.981778	14.3712	0.994718	0.971249
GL	12	A	GL	9	C	13.7600	0.998054	0.978240	13.3345	0.998136	0.979674
GL	12	A	GL	12	C	13.1034	0.998629	0.978110	12.8117	0.999189	0.979348
GL	12	A	GL	Col	C	13.5292	0.998766	0.979777	13.1918	0.998759	0.979449
GL	12	A	GL	3	D	16.2597	0.992245	0.986010	15.4500	0.986886	0.972553
GL	12	A	GL	6	D	15.0882	0.998335	0.990594	14.4628	0.995920	0.986128
GL	12	A	GL	9	D	13.8163	0.999118	0.992731	13.3812	0.999062	0.992475
GL	12	A	GL	12	D	13.3431	0.998217	0.992622	13.0319	0.999832	0.993486
GL	12	A	GL	Col	D	13.5858	0.998645	0.990333	13.2545	0.999390	0.991979
GL	Col	A	GL	3	B	16.6208	0.985992	0.958465	15.9388	0.972180	0.935318
GL	Col	A	GL	6	B	15.5862	0.994005	0.975438	14.9770	0.985263	0.955198
GL	Col	A	GL	9	B	14.5674	0.996868	0.968149	14.0792	0.993258	0.963479
GL	Col	A	GL	12	B	13.5945	0.996904	0.960643	13.2199	0.996690	0.960618
GL	Col	A	GL	Col	B	13.1493	0.997793	0.964847	12.9074	0.998054	0.964447
GL	Col	A	GL	3	C	16.5459	0.992034	0.950377	15.8683	0.966680	0.921900
GL	Col	A	GL	6	C	15.5562	0.994199	0.965146	14.9294	0.985384	0.949559
GL	Col	A	GL	9	C	14.4642	0.996146	0.965509	13.9928	0.993889	0.965373
GL	Col	A	GL	12	C	13.4698	0.996945	0.965521	13.1398	0.996346	0.963940

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GL	Col	A	GL	Col	C	13.1143	0.998412	0.966233	12.8962	0.998421	0.967080
GL	Col	A	GL	3	D	16.5145	0.991312	0.974823	15.7997	0.972499	0.952910
GL	Col	A	GL	6	D	15.5665	0.994064	0.979278	14.9725	0.988532	0.964167
GL	Col	A	GL	9	D	14.4560	0.997473	0.989383	13.9590	0.995567	0.985160
GL	Col	A	GL	12	D	13.4511	0.998919	0.991789	13.1289	0.998553	0.989222
GL	Col	A	GL	Col	D	13.0121	0.999334	0.992237	12.7727	0.999826	0.993101
GL	3	B	GL	6	B	14.5837	0.997871	0.995083	13.9398	0.998948	0.995018
GL	3	B	GL	9	B	15.4659	0.995469	0.987271	14.7494	0.996516	0.988627
GL	3	B	GL	12	B	15.9952	0.996503	0.981933	15.3124	0.990242	0.966823
GL	3	B	GL	Col	B	16.3623	0.990282	0.981365	15.7221	0.982505	0.970012
GL	3	B	GL	3	C	14.7592	0.998485	0.993442	14.0368	0.998889	0.994355
GL	3	B	GL	6	C	15.5479	0.993097	0.992233	14.5998	0.998083	0.995468
GL	3	B	GL	9	C	15.9987	0.993219	0.981565	15.0607	0.995426	0.985130
GL	3	B	GL	12	C	16.2906	0.993849	0.976484	15.4679	0.991165	0.967870
GL	3	B	GL	Col	C	16.6034	0.988529	0.982036	15.9301	0.982741	0.966681
GL	3	B	GL	3	D	14.8140	0.998169	0.992839	13.9614	0.999171	0.994478
GL	3	B	GL	6	D	15.5274	0.997069	0.991438	14.5717	0.998945	0.993070
GL	3	B	GL	9	D	15.9693	0.996978	0.987938	15.0748	0.997289	0.990848
GL	3	B	GL	12	D	16.3011	0.996356	0.985784	15.5075	0.992869	0.979140
GL	3	B	GL	Col	D	16.5792	0.993505	0.980138	15.8809	0.983748	0.927863
GL	6	B	GL	9	B	13.6228	0.998823	0.992377	13.2941	0.998660	0.991985
GL	6	B	GL	12	B	14.5434	0.997166	0.982309	14.0833	0.994456	0.973429
GL	6	B	GL	Col	B	15.2228	0.994350	0.987533	14.7661	0.984053	0.968435
GL	6	B	GL	3	C	15.4784	0.997911	0.987981	14.5695	0.997957	0.986294
GL	6	B	GL	6	C	13.8606	0.998024	0.996482	13.2966	0.999312	0.996725
GL	6	B	GL	9	C	14.3478	0.997520	0.989384	13.7215	0.998318	0.990326
GL	6	B	GL	12	C	15.1356	0.995783	0.980569	14.4618	0.993254	0.975159
GL	6	B	GL	Col	C	15.5659	0.991324	0.983706	14.9490	0.985130	0.973768
GL	6	B	GL	3	D	15.5611	0.997542	0.990467	14.6300	0.998041	0.989148

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GL	6	B	GL	6	D	13.6641	0.999153	0.991329	13.1486	0.999554	0.992106
GL	6	B	GL	9	D	14.4694	0.998380	0.988117	13.8474	0.998979	0.990393
GL	6	B	GL	12	D	14.9677	0.997508	0.991779	14.3589	0.996223	0.988434
GL	6	B	GL	Col	D	15.5769	0.995456	0.986157	14.9701	0.987793	0.964864
GL	9	B	GL	12	B	13.2954	0.999165	0.991215	13.0583	0.999241	0.991459
GL	9	B	GL	Col	B	14.1527	0.997773	0.991786	13.8232	0.993712	0.985700
GL	9	B	GL	3	C	15.8503	0.995676	0.983224	14.9650	0.996492	0.982938
GL	9	B	GL	6	C	14.4026	0.998894	0.994085	13.7696	0.998919	0.994229
GL	9	B	GL	9	C	13.3316	0.999522	0.996492	12.9471	0.999693	0.996791
GL	9	B	GL	12	C	13.8681	0.997669	0.988884	13.4039	0.998676	0.991820
GL	9	B	GL	Col	C	14.4189	0.997502	0.992138	13.9304	0.995330	0.989213
GL	9	B	GL	3	D	15.9371	0.993596	0.985457	15.0513	0.995830	0.983924
GL	9	B	GL	6	D	14.2785	0.998688	0.985915	13.6360	0.999016	0.985834
GL	9	B	GL	9	D	13.6697	0.999196	0.983471	13.2602	0.999460	0.983718
GL	9	B	GL	12	D	13.9748	0.998524	0.986361	13.5602	0.999491	0.986363
GL	9	B	GL	Col	D	14.4897	0.996985	0.985632	14.0124	0.996045	0.988388
GL	12	B	GL	Col	B	13.1709	0.998026	0.992264	12.9781	0.997944	0.991526
GL	12	B	GL	3	C	16.2002	0.994655	0.985964	15.3783	0.996540	0.987783
GL	12	B	GL	6	C	15.0576	0.997114	0.990162	14.3848	0.998102	0.990723
GL	12	B	GL	9	C	13.8928	0.998311	0.992963	13.3974	0.999262	0.994693
GL	12	B	GL	12	C	13.0189	0.999698	0.994167	12.7555	0.999152	0.993417
GL	12	B	GL	Col	C	13.5871	0.999381	0.994258	13.2532	0.997832	0.992462
GL	12	B	GL	3	D	16.2956	0.995844	0.977618	15.4644	0.995395	0.979135
GL	12	B	GL	6	D	15.0509	0.997243	0.977307	14.4102	0.997877	0.976113
GL	12	B	GL	9	D	13.6662	0.998836	0.980407	13.2039	0.998987	0.980538
GL	12	B	GL	12	D	13.1506	0.998928	0.980652	12.8705	0.999267	0.979349
GL	12	B	GL	Col	D	13.6094	0.999065	0.981461	13.2587	0.998953	0.985562
GL	Col	B	GL	3	C	16.5398	0.989893	0.965288	15.8371	0.987431	0.969503
GL	Col	B	GL	6	C	15.4937	0.996738	0.978281	14.8592	0.996239	0.981895

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GL	Col	B	GL	9	C	14.4055	0.998010	0.985539	13.9353	0.997031	0.986633
GL	Col	B	GL	12	C	13.5599	0.999240	0.988190	13.2120	0.999106	0.988787
GL	Col	B	GL	Col	C	12.7452	0.999697	0.994117	12.5709	0.999869	0.994329
GL	Col	B	GL	3	D	16.4892	0.995017	0.966208	15.8019	0.987786	0.967478
GL	Col	B	GL	6	D	15.5524	0.994472	0.975025	14.9462	0.995823	0.969643
GL	Col	B	GL	9	D	14.4348	0.995987	0.974105	13.9626	0.996624	0.975002
GL	Col	B	GL	12	D	13.6445	0.997034	0.972638	13.3392	0.997951	0.970194
GL	Col	B	GL	Col	D	12.8980	0.998437	0.964906	12.7061	0.998829	0.968613
GL	3	C	GL	6	C	14.6697	0.997591	0.992665	14.0691	0.999027	0.994200
GL	3	C	GL	9	C	15.3603	0.995756	0.989646	14.6425	0.997520	0.991116
GL	3	C	GL	12	C	15.8561	0.992426	0.976131	15.1675	0.992344	0.967446
GL	3	C	GL	Col	C	16.3593	0.994344	0.982701	15.7546	0.973596	0.939073
GL	3	C	GL	3	D	14.8092	0.998618	0.992597	14.0581	0.999408	0.993839
GL	3	C	GL	6	D	15.5478	0.997685	0.989796	14.6360	0.999043	0.990427
GL	3	C	GL	9	D	15.8413	0.998377	0.988403	14.9540	0.998301	0.989454
GL	3	C	GL	12	D	16.3212	0.988846	0.976290	15.4954	0.996180	0.981078
GL	3	C	GL	Col	D	16.5247	0.990554	0.973301	15.8300	0.984682	0.920878
GL	6	C	GL	9	C	13.6659	0.998318	0.992333	13.3255	0.998428	0.992205
GL	6	C	GL	12	C	14.5533	0.997776	0.985166	14.0639	0.995308	0.980593
GL	6	C	GL	Col	C	15.1969	0.992737	0.987022	14.6987	0.988062	0.979797
GL	6	C	GL	3	D	15.5776	0.997995	0.990837	14.6628	0.998567	0.989694
GL	6	C	GL	6	D	13.8043	0.999023	0.991889	13.3127	0.999558	0.992501
GL	6	C	GL	9	D	14.4178	0.998373	0.989513	13.7670	0.999212	0.991735
GL	6	C	GL	12	D	15.0314	0.993259	0.988012	14.3712	0.996768	0.989356
GL	6	C	GL	Col	D	15.5997	0.994279	0.980984	14.9564	0.988759	0.969370
GL	9	C	GL	12	C	13.1535	0.999448	0.992670	12.9456	0.998865	0.992387
GL	9	C	GL	Col	C	13.8423	0.998838	0.993742	13.5502	0.993330	0.986159
GL	9	C	GL	3	D	15.9486	0.998475	0.985618	15.0473	0.996349	0.983931
GL	9	C	GL	6	D	14.4518	0.998974	0.985582	13.8583	0.999018	0.984996

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GL	9	C	GL	9	D	13.6542	0.999434	0.982207	13.3082	0.999563	0.983008
GL	9	C	GL	12	D	13.9242	0.999229	0.986623	13.4729	0.999360	0.987610
GL	9	C	GL	Col	D	14.4487	0.996211	0.988621	13.9664	0.995654	0.988422
GL	12	C	GL	Col	C	12.8873	0.999402	0.993990	12.7154	0.998284	0.992682
GL	12	C	GL	3	D	16.2362	0.996320	0.980519	15.4198	0.995109	0.977922
GL	12	C	GL	6	D	15.1191	0.998295	0.980839	14.4247	0.997610	0.979025
GL	12	C	GL	9	D	13.8764	0.998879	0.978889	13.4237	0.998871	0.980047
GL	12	C	GL	12	D	13.2683	0.998538	0.976807	12.9881	0.999372	0.977278
GL	12	C	GL	Col	D	13.5436	0.998758	0.978992	13.2077	0.999090	0.984744
GL	Col	C	GL	3	D	16.5340	0.994728	0.963130	15.8647	0.987839	0.966236
GL	Col	C	GL	6	D	15.5157	0.995075	0.972044	14.8908	0.995173	0.970550
GL	Col	C	GL	9	D	14.4452	0.996107	0.973830	13.9360	0.996277	0.974679
GL	Col	C	GL	12	D	13.6034	0.997996	0.972997	13.2928	0.998048	0.971494
GL	Col	C	GL	Col	D	12.9193	0.998205	0.964540	12.6697	0.998743	0.966170
GL	3	D	GL	6	D	14.5959	0.999066	0.993486	13.9928	0.999209	0.993135
GL	3	D	GL	9	D	15.3859	0.997928	0.991939	14.6647	0.997344	0.990958
GL	3	D	GL	12	D	15.8899	0.994056	0.981420	15.2082	0.993089	0.969595
GL	3	D	GL	Col	D	16.2868	0.995204	0.981937	15.6961	0.984345	0.917759
GL	6	D	GL	9	D	13.5191	0.999268	0.994337	13.1913	0.999276	0.994676
GL	6	D	GL	12	D	14.4754	0.998004	0.994426	14.0240	0.996784	0.992231
GL	6	D	GL	Col	D	15.2224	0.993609	0.983083	14.7304	0.989736	0.969010
GL	9	D	GL	12	D	13.2184	0.999552	0.994475	13.0165	0.999397	0.995733
GL	9	D	GL	Col	D	14.0695	0.997851	0.990104	13.7231	0.994993	0.986003
GL	12	D	GL	Col	D	12.9511	0.999247	0.993445	12.7663	0.998790	0.993207
Large	6	250kB	GL	3	A	15.9235	0.994329	0.988556	15.0214	0.998507	0.993319
Large	6	250kB	GL	6	A	14.9181	0.997989	0.991210	14.2742	0.998529	0.991988
Large	6	250kB	GL	9	A	14.3214	0.999080	0.990945	13.6974	0.999183	0.993400
Large	6	250kB	GL	12	A	14.5018	0.998452	0.988850	13.9009	0.997761	0.989859
Large	6	250kB	GL	Col	A	15.0202	0.997139	0.990882	14.4429	0.993530	0.983862

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
Large	6	250kB	GL	3	B	15.8936	0.995632	0.990228	15.0162	0.997881	0.992065
Large	6	250kB	GL	6	B	15.1250	0.998785	0.993527	14.4085	0.998930	0.994305
Large	6	250kB	GL	9	B	14.3377	0.997543	0.990658	13.7021	0.998714	0.992896
Large	6	250kB	GL	12	B	14.5375	0.996242	0.983679	13.9053	0.996289	0.982438
Large	6	250kB	GL	Col	B	14.9695	0.994924	0.990402	14.3717	0.991176	0.983501
Large	6	250kB	GL	3	C	16.0146	0.995353	0.987222	15.0673	0.998460	0.992433
Large	6	250kB	GL	6	C	15.0078	0.996959	0.992503	14.2803	0.999281	0.994982
Large	6	250kB	GL	9	C	14.4742	0.998935	0.993490	13.8209	0.998552	0.993556
Large	6	250kB	GL	12	C	14.5167	0.998399	0.989721	13.8805	0.995793	0.983802
Large	6	250kB	GL	Col	C	15.0113	0.996245	0.991327	14.4410	0.991496	0.981600
Large	6	250kB	GL	3	D	15.8685	0.996976	0.989309	14.9861	0.998343	0.991214
Large	6	250kB	GL	6	D	15.0147	0.998160	0.987401	14.3287	0.999243	0.990639
Large	6	250kB	GL	9	D	14.5110	0.997790	0.987694	13.8535	0.999137	0.991089
Large	6	250kB	GL	12	D	14.6094	0.997702	0.989854	14.0001	0.998053	0.990433
Large	6	250kB	GL	Col	D	14.9616	0.997177	0.990808	14.3909	0.989024	0.971785
Large	9	250kB	GL	3	A	16.0463	0.995138	0.986926	15.2475	0.997533	0.990122
Large	9	250kB	GL	6	A	14.9169	0.998174	0.993017	14.2722	0.997659	0.991027
Large	9	250kB	GL	9	A	13.8381	0.999544	0.987935	13.4193	0.999399	0.987790
Large	9	250kB	GL	12	A	13.7690	0.999515	0.981972	13.3010	0.999313	0.984977
Large	9	250kB	GL	Col	A	14.1153	0.998746	0.984759	13.6739	0.994491	0.987220
Large	9	250kB	GL	3	B	16.1841	0.995925	0.988716	15.3571	0.997729	0.989436
Large	9	250kB	GL	6	B	15.1006	0.998401	0.991322	14.4141	0.998592	0.991513
Large	9	250kB	GL	9	B	13.8877	0.999151	0.991974	13.4337	0.999477	0.992794
Large	9	250kB	GL	12	B	13.6904	0.999231	0.993583	13.2407	0.998116	0.991839
Large	9	250kB	GL	Col	B	14.1681	0.998938	0.994790	13.7137	0.992365	0.983678
Large	9	250kB	GL	3	C	16.1241	0.997996	0.990941	15.2674	0.997154	0.989224
Large	9	250kB	GL	6	C	15.0853	0.996397	0.990296	14.3942	0.998981	0.992891
Large	9	250kB	GL	9	C	13.8746	0.999180	0.992531	13.3913	0.998819	0.992559
Large	9	250kB	GL	12	C	13.5742	0.999176	0.993587	13.1317	0.998135	0.992267

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
Large	9	250kB	GL	Col	C	14.0983	0.998481	0.994088	13.6900	0.993690	0.988395
Large	9	250kB	GL	3	D	16.1121	0.996196	0.985648	15.2608	0.996400	0.985463
Large	9	250kB	GL	6	D	14.9481	0.998136	0.983248	14.2958	0.998538	0.983410
Large	9	250kB	GL	9	D	13.8177	0.998443	0.985745	13.3404	0.999354	0.985763
Large	9	250kB	GL	12	D	13.7718	0.998388	0.983182	13.3328	0.999489	0.987010
Large	9	250kB	GL	Col	D	14.2202	0.998569	0.988906	13.7918	0.994791	0.989447
GLNO	3	A	GL	6	250kB	15.9242	0.994882	0.989508	15.0299	0.998407	0.994083
GLNO	3	A	GL	9	250kB	16.1210	0.993356	0.980779	15.2452	0.996158	0.983119
GLNO	3	A	GL	3	A	14.5741	0.998096	0.993388	13.7909	0.999612	0.996251
GLNO	3	A	GL	6	A	15.4639	0.998007	0.992129	14.5507	0.998543	0.992308
GLNO	3	A	GL	9	A	15.9207	0.996713	0.991919	14.9882	0.998241	0.993353
GLNO	3	A	GL	12	A	16.2441	0.994754	0.984850	15.3647	0.995434	0.981076
GLNO	3	A	GL	Col	A	16.5436	0.995512	0.987491	15.8344	0.988193	0.956690
GLNO	3	A	GL	3	B	14.7730	0.998416	0.993809	14.0203	0.999401	0.994634
GLNO	3	A	GL	6	B	15.5193	0.998208	0.992733	14.5931	0.999002	0.993474
GLNO	3	A	GL	9	B	15.8116	0.996528	0.990167	14.9258	0.997095	0.991029
GLNO	3	A	GL	12	B	16.1991	0.994564	0.983481	15.3195	0.993645	0.974170
GLNO	3	A	GL	Col	B	16.5019	0.992903	0.984852	15.7792	0.987647	0.973577
GLNO	3	A	GL	3	C	14.7465	0.995772	0.990955	13.9132	0.999542	0.995467
GLNO	3	A	GL	6	C	15.4685	0.996627	0.992937	14.5177	0.999195	0.994459
GLNO	3	A	GL	9	C	15.8036	0.995966	0.988528	14.8781	0.997336	0.990982
GLNO	3	A	GL	12	C	16.2587	0.993438	0.979127	15.3916	0.993997	0.973206
GLNO	3	A	GL	Col	C	16.5795	0.985128	0.976463	15.8258	0.989703	0.976085
GLNO	3	A	GL	3	D	14.8012	0.997646	0.991249	14.0177	0.999451	0.993005
GLNO	3	A	GL	6	D	15.4919	0.998386	0.991127	14.5552	0.999245	0.991917
GLNO	3	A	GL	9	D	15.8819	0.997237	0.989155	14.9582	0.998059	0.991057
GLNO	3	A	GL	12	D	16.2037	0.996408	0.988108	15.3988	0.996193	0.983208
GLNO	3	A	GL	Col	D	16.5298	0.995218	0.984852	15.7942	0.987838	0.940399
GLNO	3	A	GLNO	6	A	14.6372	0.998548	0.991331	14.0373	0.998949	0.992202

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GLNO	3	A	GLNO	9	A	15.3349	0.996936	0.990856	14.6049	0.997771	0.992131
GLNO	3	A	GLNO	12	A	15.8772	0.995235	0.989572	15.1500	0.994386	0.987281
GLNO	3	A	GLNO	Col	A	16.2762	0.990825	0.973003	15.5956	0.985445	0.960208
GLNO	3	A	GLNO	3	B	16.3901	0.992733	0.988049	15.0212	0.999282	0.993157
GLNO	3	A	GLNO	6	B	16.1916	0.996846	0.991466	14.9290	0.999111	0.994316
GLNO	3	A	GLNO	9	B	16.3112	0.996229	0.987942	15.2415	0.998269	0.991182
GLNO	3	A	GLNO	12	B	16.5137	0.996863	0.988423	15.5698	0.996200	0.986367
GLNO	3	A	GLNO	Col	B	16.7426	0.996166	0.974951	15.9329	0.989762	0.913267
GLNO	6	A	GL	6	250kB	14.9816	0.996957	0.990910	14.2824	0.997651	0.991941
GLNO	6	A	GL	9	250kB	14.9756	0.998083	0.986663	14.3199	0.996697	0.983396
GLNO	6	A	GL	3	A	15.5021	0.996612	0.989308	14.5680	0.997937	0.991453
GLNO	6	A	GL	6	A	14.0304	0.998494	0.989173	13.5473	0.999475	0.992155
GLNO	6	A	GL	9	A	14.3912	0.998172	0.991813	13.7598	0.998974	0.992983
GLNO	6	A	GL	12	A	15.1090	0.996628	0.986362	14.4425	0.997940	0.986975
GLNO	6	A	GL	Col	A	15.5707	0.995104	0.987686	14.9525	0.992353	0.979657
GLNO	6	A	GL	3	B	15.4343	0.997633	0.993984	14.4842	0.997985	0.993339
GLNO	6	A	GL	6	B	13.5425	0.999337	0.996818	13.0144	0.999734	0.997385
GLNO	6	A	GL	9	B	14.1926	0.996608	0.989856	13.5855	0.998526	0.992583
GLNO	6	A	GL	12	B	15.0712	0.996085	0.979678	14.3897	0.996418	0.978881
GLNO	6	A	GL	Col	B	15.4928	0.993007	0.983762	14.8847	0.991630	0.981901
GLNO	6	A	GL	3	C	15.4895	0.995483	0.985163	14.5647	0.997709	0.985758
GLNO	6	A	GL	6	C	13.8276	0.998915	0.996199	13.2649	0.999657	0.996666
GLNO	6	A	GL	9	C	14.4767	0.998030	0.990061	13.8687	0.998611	0.991443
GLNO	6	A	GL	12	C	15.0486	0.995895	0.981837	14.3883	0.995901	0.982452
GLNO	6	A	GL	Col	C	15.5404	0.995156	0.987469	14.9091	0.990510	0.981603
GLNO	6	A	GL	3	D	15.6015	0.996929	0.989246	14.6713	0.998053	0.990003
GLNO	6	A	GL	6	D	13.6713	0.998684	0.991001	13.1627	0.999582	0.992263
GLNO	6	A	GL	9	D	14.3008	0.997410	0.990397	13.6614	0.999307	0.991122
GLNO	6	A	GL	12	D	15.0268	0.996342	0.987664	14.3425	0.997186	0.989195

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GLNO	6	A	GL	Col	D	15.5652	0.993439	0.982993	14.9375	0.989986	0.971110
GLNO	6	A	GLNO	9	A	13.5072	0.999325	0.996492	13.2165	0.998522	0.996179
GLNO	6	A	GLNO	12	A	14.5869	0.996224	0.990737	14.0972	0.996825	0.991222
GLNO	6	A	GLNO	Col	A	15.2493	0.994369	0.979684	14.7546	0.990048	0.969146
GLNO	6	A	GLNO	3	B	16.2687	0.995171	0.987446	15.0390	0.998182	0.989410
GLNO	6	A	GLNO	6	B	15.0457	0.998161	0.990449	13.9966	0.999478	0.991335
GLNO	6	A	GLNO	9	B	15.1975	0.995713	0.988532	14.2390	0.999158	0.991128
GLNO	6	A	GLNO	12	B	15.5141	0.994963	0.983701	14.6565	0.997006	0.985646
GLNO	6	A	GLNO	Col	B	15.8497	0.994799	0.976855	15.1000	0.990571	0.956512
GLNO	9	A	GL	6	250kB	14.3086	0.997197	0.990159	13.6654	0.998370	0.991597
GLNO	9	A	GL	9	250kB	13.8791	0.998858	0.990074	13.4518	0.999210	0.991332
GLNO	9	A	GL	3	A	15.9156	0.991431	0.980197	15.0126	0.997082	0.986943
GLNO	9	A	GL	6	A	14.3834	0.998103	0.987958	13.7928	0.998983	0.989232
GLNO	9	A	GL	9	A	13.3286	0.999351	0.985064	13.0169	0.999558	0.985289
GLNO	9	A	GL	12	A	13.9257	0.999192	0.983683	13.5141	0.999366	0.984412
GLNO	9	A	GL	Col	A	14.4625	0.998369	0.986926	14.0053	0.997630	0.986042
GLNO	9	A	GL	3	B	15.8972	0.995986	0.991715	14.9849	0.997020	0.990855
GLNO	9	A	GL	6	B	14.3612	0.999120	0.993950	13.7567	0.998918	0.993489
GLNO	9	A	GL	9	B	13.3860	0.999537	0.994780	13.0241	0.999698	0.995343
GLNO	9	A	GL	12	B	13.7752	0.998076	0.986058	13.3009	0.998886	0.988558
GLNO	9	A	GL	Col	B	14.4082	0.996976	0.990770	13.9597	0.994795	0.987954
GLNO	9	A	GL	3	C	15.8939	0.995325	0.982968	14.9761	0.996507	0.982387
GLNO	9	A	GL	6	C	14.3129	0.997593	0.994593	13.6751	0.999312	0.994260
GLNO	9	A	GL	9	C	13.2743	0.999061	0.994025	12.9010	0.999616	0.994938
GLNO	9	A	GL	12	C	13.8568	0.998640	0.990863	13.3954	0.998825	0.991470
GLNO	9	A	GL	Col	C	14.4464	0.998093	0.991939	13.9527	0.995777	0.989873
GLNO	9	A	GL	3	D	15.9569	0.996650	0.985273	15.0479	0.996646	0.984038
GLNO	9	A	GL	6	D	14.3433	0.997591	0.986258	13.7288	0.998959	0.987173
GLNO	9	A	GL	9	D	13.3846	0.998917	0.985394	12.9752	0.999530	0.985342

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GLNO	9	A	GL	12	D	13.8419	0.998622	0.987426	13.3736	0.999451	0.987962
GLNO	9	A	GL	Col	D	14.4056	0.998301	0.986111	13.8995	0.997169	0.988340
GLNO	9	A	GLNO	12	A	13.2222	0.999428	0.995324	12.9936	0.999151	0.995104
GLNO	9	A	GLNO	Col	A	14.0501	0.997392	0.986520	13.7112	0.992871	0.978433
GLNO	9	A	GLNO	3	B	16.4382	0.993976	0.980631	15.3680	0.996394	0.982722
GLNO	9	A	GLNO	6	B	15.1311	0.997233	0.987092	14.1499	0.998790	0.988742
GLNO	9	A	GLNO	9	B	14.2186	0.997836	0.987049	13.4206	0.999409	0.988452
GLNO	9	A	GLNO	12	B	14.6092	0.997948	0.981629	13.8378	0.998811	0.983773
GLNO	9	A	GLNO	Col	B	14.8230	0.997418	0.978128	14.1495	0.996457	0.974306
GLNO	12	A	GL	3	A	16.2270	0.994628	0.982300	15.3999	0.995302	0.983097
GLNO	12	A	GL	6	A	15.0605	0.997673	0.984509	14.4205	0.997169	0.984910
GLNO	12	A	GL	9	A	13.7493	0.998663	0.983481	13.3136	0.998600	0.981902
GLNO	12	A	GL	12	A	13.2776	0.999068	0.978060	12.9988	0.998891	0.976474
GLNO	12	A	GL	Col	A	13.6585	0.999185	0.977986	13.2897	0.999284	0.983138
GLNO	12	A	GL	3	B	16.2902	0.995631	0.984475	15.4647	0.995914	0.984060
GLNO	12	A	GL	6	B	14.9651	0.998560	0.988768	14.3145	0.997767	0.988834
GLNO	12	A	GL	9	B	13.8108	0.997888	0.992577	13.3640	0.999204	0.993857
GLNO	12	A	GL	12	B	13.3208	0.999294	0.994681	12.9999	0.999806	0.995903
GLNO	12	A	GL	Col	B	13.6623	0.998867	0.992574	13.3317	0.997996	0.990000
GLNO	12	A	GL	3	C	16.1904	0.996266	0.984203	15.3734	0.995502	0.979853
GLNO	12	A	GL	6	C	15.0428	0.998415	0.988425	14.3622	0.997667	0.989263
GLNO	12	A	GL	9	C	13.8389	0.998656	0.992069	13.3657	0.998987	0.993257
GLNO	12	A	GL	12	C	13.1423	0.999713	0.994399	12.8285	0.999477	0.994351
GLNO	12	A	GL	Col	C	13.6589	0.999168	0.993198	13.2985	0.998547	0.992665
GLNO	12	A	GL	3	D	16.2789	0.997022	0.979645	15.4530	0.995482	0.980464
GLNO	12	A	GL	6	D	14.9785	0.996422	0.980704	14.3111	0.997361	0.978903
GLNO	12	A	GL	9	D	13.9243	0.998933	0.979729	13.4693	0.998695	0.980659
GLNO	12	A	GL	12	D	13.1663	0.999216	0.978667	12.8666	0.999213	0.978988
GLNO	12	A	GL	Col	D	13.5458	0.998604	0.980290	13.1988	0.999137	0.987038

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GLNO	12	A	GLNO	Col	A	13.0152	0.999554	0.995062	12.8637	0.998083	0.992964
GLNO	12	A	GLNO	3	B	16.5664	0.993576	0.978118	15.6549	0.995078	0.979733
GLNO	12	A	GLNO	6	B	15.6379	0.996558	0.980272	14.7620	0.997632	0.981576
GLNO	12	A	GLNO	9	B	14.5960	0.998384	0.978392	13.8445	0.998633	0.978356
GLNO	12	A	GLNO	12	B	14.0285	0.998516	0.978324	13.3924	0.999075	0.980195
GLNO	12	A	GLNO	Col	B	14.1117	0.998452	0.973187	13.5302	0.999325	0.978044
GLNO	12	A	Large	6	250kB	14.5156	0.998537	0.988358	13.8514	0.998381	0.988240
GLNO	12	A	Large	9	250kB	13.8367	0.998651	0.988385	13.3604	0.999224	0.990003
GLNO	Col	A	GL	6	250kB	14.9709	0.996242	0.977775	14.3945	0.992776	0.980818
GLNO	Col	A	GL	9	250kB	14.1017	0.997392	0.984085	13.6609	0.996792	0.984529
GLNO	Col	A	GL	3	A	16.5374	0.991735	0.973942	15.8108	0.991613	0.971457
GLNO	Col	A	GL	6	A	15.5620	0.995347	0.981159	14.9588	0.995062	0.977389
GLNO	Col	A	GL	9	A	14.4037	0.998457	0.979250	13.9559	0.996134	0.975758
GLNO	Col	A	GL	12	A	13.5529	0.998322	0.966198	13.2380	0.997904	0.965942
GLNO	Col	A	GL	Col	A	13.0990	0.998689	0.964538	12.8660	0.998688	0.964023
GLNO	Col	A	GL	3	B	16.5427	0.997200	0.972814	15.8660	0.993221	0.973428
GLNO	Col	A	GL	6	B	15.5400	0.996183	0.975707	14.9856	0.994697	0.978826
GLNO	Col	A	GL	9	B	14.5188	0.998274	0.986250	14.0879	0.997006	0.987356
GLNO	Col	A	GL	12	B	13.5692	0.998809	0.990968	13.2604	0.998989	0.991072
GLNO	Col	A	GL	Col	B	12.7351	0.999575	0.991399	12.5295	0.999774	0.991547
GLNO	Col	A	GL	3	C	16.5240	0.992799	0.971751	15.8437	0.991216	0.974496
GLNO	Col	A	GL	6	C	15.5595	0.995453	0.977555	14.9345	0.995099	0.981663
GLNO	Col	A	GL	9	C	14.2833	0.998286	0.987124	13.8463	0.996494	0.988710
GLNO	Col	A	GL	12	C	13.5534	0.999154	0.990739	13.2542	0.999101	0.991347
GLNO	Col	A	GL	Col	C	13.0385	0.999779	0.992875	12.8147	0.999668	0.993257
GLNO	Col	A	GL	3	D	16.5372	0.994398	0.970545	15.8346	0.986108	0.966897
GLNO	Col	A	GL	6	D	15.5193	0.994451	0.972652	14.9022	0.993414	0.966049
GLNO	Col	A	GL	9	D	14.5214	0.997846	0.971920	14.0877	0.995749	0.972186
GLNO	Col	A	GL	12	D	13.4601	0.998163	0.970308	13.2051	0.997792	0.969623

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GLNO	Col	A	GL	Col	D	12.9209	0.998553	0.968398	12.7062	0.998702	0.968777
GLNO	Col	A	GLNO	3	B	16.7275	0.987346	0.963090	15.9772	0.991850	0.969527
GLNO	Col	A	GLNO	6	B	15.8696	0.996129	0.971616	15.1637	0.994049	0.972348
GLNO	Col	A	GLNO	9	B	14.8687	0.995025	0.962994	14.2550	0.995242	0.964599
GLNO	Col	A	GLNO	12	B	14.1223	0.998142	0.964869	13.6075	0.997746	0.967492
GLNO	Col	A	GLNO	Col	B	13.6962	0.998196	0.963078	13.2114	0.998681	0.966666
GLNO	3	B	GL	6	250kB	15.9714	0.997145	0.989567	15.1199	0.997924	0.990809
GLNO	3	B	GL	9	250kB	16.1768	0.994559	0.975849	15.3420	0.997313	0.982283
GLNO	3	B	GL	3	A	14.6854	0.997723	0.991598	13.9161	0.999454	0.993580
GLNO	3	B	GL	6	A	15.5587	0.995074	0.983868	14.6542	0.998686	0.989807
GLNO	3	B	GL	9	A	15.9212	0.997589	0.991836	15.0531	0.997983	0.994042
GLNO	3	B	GL	12	A	16.2093	0.995898	0.984761	15.4103	0.995734	0.979143
GLNO	3	B	GL	Col	A	16.4984	0.989271	0.976845	15.8480	0.987525	0.943715
GLNO	3	B	GL	3	B	14.7910	0.998926	0.991697	14.0436	0.999342	0.991953
GLNO	3	B	GL	6	B	15.5050	0.994459	0.989542	14.5603	0.998654	0.992347
GLNO	3	B	GL	9	B	15.9359	0.994592	0.981180	15.0298	0.997593	0.987538
GLNO	3	B	GL	12	B	16.2653	0.990193	0.953255	15.4508	0.995494	0.972238
GLNO	3	B	GL	Col	B	16.5186	0.985134	0.952718	15.8277	0.984763	0.962221
GLNO	3	B	GL	3	C	14.8705	0.997342	0.990796	14.1001	0.999455	0.992512
GLNO	3	B	GL	6	C	15.4809	0.998145	0.990140	14.5647	0.999212	0.992210
GLNO	3	B	GL	9	C	15.8904	0.996426	0.982896	15.0168	0.997544	0.987426
GLNO	3	B	GL	12	C	16.2998	0.993561	0.970104	15.4619	0.994731	0.972417
GLNO	3	B	GL	Col	C	16.5288	0.986474	0.966387	15.8646	0.979336	0.955008
GLNO	3	B	GL	3	D	14.6988	0.998112	0.995429	13.9270	0.999651	0.996745
GLNO	3	B	GL	6	D	15.4964	0.997680	0.991015	14.5404	0.999146	0.992767
GLNO	3	B	GL	9	D	15.9210	0.996248	0.984534	15.0192	0.998051	0.990715
GLNO	3	B	GL	12	D	16.2572	0.996166	0.986819	15.4280	0.995738	0.980580
GLNO	3	B	GL	Col	D	16.5659	0.973633	0.945098	15.8511	0.984314	0.920809
GLNO	3	B	GLNO	6	B	14.6404	0.998286	0.994467	13.9883	0.999376	0.995666

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GLNO	3	B	GLNO	9	B	15.4730	0.996305	0.989685	14.7664	0.998127	0.991966
GLNO	3	B	GLNO	12	B	15.8327	0.994263	0.987471	15.1616	0.993416	0.972418
GLNO	3	B	GLNO	Col	B	16.2847	0.986893	0.943472	15.6980	0.972962	0.822135
GLNO	6	B	GL	6	250kB	15.0456	0.998046	0.989105	14.2828	0.998660	0.989185
GLNO	6	B	GL	9	250kB	15.1008	0.996933	0.979822	14.3884	0.998214	0.982716
GLNO	6	B	GL	3	A	15.5764	0.996575	0.985539	14.6180	0.998686	0.989217
GLNO	6	B	GL	6	A	13.9084	0.998930	0.989723	13.4135	0.999543	0.990664
GLNO	6	B	GL	9	A	14.5285	0.999269	0.995406	13.8843	0.999315	0.995792
GLNO	6	B	GL	12	A	15.0523	0.996221	0.988398	14.3824	0.997310	0.986418
GLNO	6	B	GL	Col	A	15.5674	0.995044	0.982596	14.9341	0.992262	0.973011
GLNO	6	B	GL	3	B	15.5158	0.997935	0.988394	14.5629	0.998409	0.988028
GLNO	6	B	GL	6	B	13.9729	0.999433	0.989596	13.4699	0.999547	0.990865
GLNO	6	B	GL	9	B	14.3364	0.998810	0.988205	13.6677	0.999290	0.990236
GLNO	6	B	GL	12	B	15.1114	0.997479	0.978232	14.4248	0.996131	0.972330
GLNO	6	B	GL	Col	B	15.6706	0.992806	0.982100	15.0214	0.991278	0.981295
GLNO	6	B	GL	3	C	15.4847	0.996669	0.984486	14.5636	0.998278	0.984831
GLNO	6	B	GL	6	C	13.6611	0.998746	0.992837	13.1430	0.999605	0.992832
GLNO	6	B	GL	9	C	14.4398	0.998550	0.987883	13.8102	0.998975	0.990030
GLNO	6	B	GL	12	C	15.2126	0.996509	0.979719	14.5295	0.995835	0.977589
GLNO	6	B	GL	Col	C	15.5860	0.995273	0.986654	14.9664	0.990456	0.980841
GLNO	6	B	GL	3	D	15.5066	0.998598	0.994022	14.5633	0.998640	0.993176
GLNO	6	B	GL	6	D	13.6992	0.999185	0.995552	13.1661	0.999589	0.996202
GLNO	6	B	GL	9	D	14.4360	0.998719	0.993820	13.7883	0.999430	0.995162
GLNO	6	B	GL	12	D	15.0729	0.996052	0.990427	14.3905	0.997558	0.990470
GLNO	6	B	GL	Col	D	15.6230	0.994173	0.985835	15.0164	0.991057	0.972848
GLNO	6	B	GLNO	9	B	13.4981	0.999477	0.994589	13.1551	0.999565	0.994425
GLNO	6	B	GLNO	12	B	14.6039	0.997559	0.989248	14.0645	0.996290	0.983768
GLNO	6	B	GLNO	Col	B	15.3229	0.993378	0.976026	14.8314	0.993770	0.954697
GLNO	9	B	GL	6	250kB	14.3978	0.997711	0.985078	13.7506	0.997250	0.981027

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GLNO	9	B	GL	9	250kB	13.8250	0.998131	0.980975	13.3619	0.998860	0.982008
GLNO	9	B	GL	3	A	15.8925	0.995078	0.981083	14.9806	0.994918	0.976355
GLNO	9	B	GL	6	A	14.3545	0.997962	0.981193	13.7263	0.998044	0.979392
GLNO	9	B	GL	9	A	13.4482	0.999288	0.992012	13.0591	0.999755	0.993114
GLNO	9	B	GL	12	A	13.6998	0.997867	0.992513	13.2590	0.999079	0.993889
GLNO	9	B	GL	Col	A	14.3847	0.994831	0.986442	13.9362	0.995741	0.987290
GLNO	9	B	GL	3	B	15.9148	0.996233	0.981986	15.0008	0.995441	0.979309
GLNO	9	B	GL	6	B	14.5184	0.998266	0.987645	13.8588	0.997474	0.984055
GLNO	9	B	GL	9	B	13.5097	0.998674	0.984859	13.1340	0.999234	0.985092
GLNO	9	B	GL	12	B	13.7705	0.998156	0.975648	13.3049	0.998791	0.976524
GLNO	9	B	GL	Col	B	14.3957	0.995248	0.984717	13.9312	0.995962	0.983924
GLNO	9	B	GL	3	C	15.9001	0.993577	0.972393	15.0526	0.994620	0.970626
GLNO	9	B	GL	6	C	14.5096	0.996439	0.986281	13.8840	0.998261	0.984077
GLNO	9	B	GL	9	C	13.3489	0.998512	0.982872	12.9427	0.998993	0.985492
GLNO	9	B	GL	12	C	13.7564	0.997733	0.979836	13.2813	0.998620	0.981898
GLNO	9	B	GL	Col	C	14.4178	0.994982	0.985378	13.9750	0.995432	0.985612
GLNO	9	B	GL	3	D	15.8355	0.996906	0.990592	14.9413	0.995910	0.986939
GLNO	9	B	GL	6	D	14.4470	0.999289	0.994078	13.8337	0.998949	0.992677
GLNO	9	B	GL	9	D	13.3213	0.999545	0.993119	12.9850	0.999593	0.993060
GLNO	9	B	GL	12	D	13.8726	0.997993	0.994531	13.4396	0.999170	0.994494
GLNO	9	B	GL	Col	D	14.3690	0.995177	0.984705	13.9197	0.994244	0.983510
GLNO	9	B	GLNO	12	B	13.1163	0.999577	0.994530	12.8981	0.999320	0.994685
GLNO	9	B	GLNO	Col	B	14.0675	0.997805	0.989150	13.7588	0.996904	0.982608
GLNO	12	B	GL	6	250kB	14.5028	0.997910	0.980203	13.8392	0.994257	0.972737
GLNO	12	B	GL	9	250kB	13.8118	0.998432	0.971475	13.3612	0.997467	0.970455
GLNO	12	B	GL	3	A	16.2659	0.993387	0.974831	15.4455	0.987166	0.958042
GLNO	12	B	GL	6	A	15.0754	0.996955	0.971761	14.4014	0.994438	0.965406
GLNO	12	B	GL	9	A	13.7983	0.999204	0.992157	13.3625	0.998769	0.991952
GLNO	12	B	GL	12	A	13.1121	0.999714	0.995602	12.8181	0.999753	0.995694

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GLNO	12	B	GL	Col	A	13.6324	0.999226	0.995482	13.3222	0.999440	0.995741
GLNO	12	B	GL	3	B	16.2941	0.993512	0.972259	15.4880	0.986611	0.960638
GLNO	12	B	GL	6	B	15.1220	0.996038	0.981693	14.4651	0.992662	0.971603
GLNO	12	B	GL	9	B	13.9585	0.997215	0.975774	13.4627	0.997730	0.977115
GLNO	12	B	GL	12	B	13.2373	0.998798	0.975999	12.9049	0.999199	0.976923
GLNO	12	B	GL	Col	B	13.4801	0.998752	0.981231	13.1094	0.998688	0.978956
GLNO	12	B	GL	3	C	16.3080	0.993610	0.962916	15.4744	0.985656	0.946387
GLNO	12	B	GL	6	C	15.0460	0.992495	0.977245	14.3694	0.993812	0.970566
GLNO	12	B	GL	9	C	13.7665	0.997277	0.977579	13.3097	0.997802	0.978010
GLNO	12	B	GL	12	C	13.3028	0.998685	0.976067	13.0191	0.998871	0.977115
GLNO	12	B	GL	Col	C	13.5735	0.998351	0.979265	13.2488	0.998370	0.979439
GLNO	12	B	GL	3	D	16.1841	0.997070	0.986734	15.4135	0.985602	0.971180
GLNO	12	B	GL	6	D	15.0841	0.997813	0.989546	14.4180	0.995700	0.984102
GLNO	12	B	GL	9	D	13.9545	0.998122	0.991652	13.4661	0.998917	0.990855
GLNO	12	B	GL	12	D	13.1002	0.999308	0.994991	12.8079	0.999889	0.994523
GLNO	12	B	GL	Col	D	13.4949	0.999171	0.992473	13.1726	0.999403	0.992705
GLNO	12	B	GLNO	Col	B	13.0075	0.999722	0.994642	12.8446	0.999635	0.993944
GLNO	Col	B	GL	6	250kB	14.9848	0.996715	0.970635	14.3741	0.982691	0.948289
GLNO	Col	B	GL	9	250kB	14.2574	0.994595	0.951887	13.8420	0.991955	0.946074
GLNO	Col	B	GL	3	A	16.4906	0.993518	0.962126	15.8096	0.971453	0.928582
GLNO	Col	B	GL	6	A	15.5207	0.994928	0.956957	14.9030	0.983775	0.936373
GLNO	Col	B	GL	9	A	14.4525	0.996229	0.987333	13.9569	0.994273	0.984345
GLNO	Col	B	GL	12	A	13.5850	0.999010	0.993100	13.2630	0.998598	0.993292
GLNO	Col	B	GL	Col	A	13.1060	0.999656	0.995901	12.8853	0.999693	0.996001
GLNO	Col	B	GL	3	B	16.5376	0.992568	0.963798	15.8180	0.969873	0.933004
GLNO	Col	B	GL	6	B	15.5100	0.992984	0.980603	14.8910	0.983434	0.954531
GLNO	Col	B	GL	9	B	14.5185	0.995654	0.966269	14.0377	0.991890	0.960755
GLNO	Col	B	GL	12	B	13.6843	0.996974	0.958151	13.2942	0.996174	0.954219
GLNO	Col	B	GL	Col	B	13.0497	0.998139	0.967500	12.8138	0.998216	0.966913

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
GLNO	Col	B	GL	3	C	16.4944	0.990246	0.944218	15.8280	0.965924	0.908719
GLNO	Col	B	GL	6	C	15.5418	0.986335	0.962078	14.9428	0.985751	0.949214
GLNO	Col	B	GL	9	C	14.4248	0.994177	0.962870	13.9669	0.992459	0.961634
GLNO	Col	B	GL	12	C	13.5299	0.996608	0.963477	13.1785	0.995865	0.960994
GLNO	Col	B	GL	Col	C	12.9504	0.997948	0.965563	12.7116	0.998159	0.965447
GLNO	Col	B	GL	3	D	16.5408	0.994004	0.978402	15.8315	0.968305	0.947456
GLNO	Col	B	GL	6	D	15.5203	0.994990	0.982904	14.8960	0.986467	0.967954
GLNO	Col	B	GL	9	D	14.3306	0.995596	0.988080	13.8359	0.994631	0.985142
GLNO	Col	B	GL	12	D	13.5714	0.999019	0.991885	13.2230	0.998429	0.991919
GLNO	Col	B	GL	Col	D	12.9177	0.999694	0.991310	12.7083	0.999847	0.991805
FixV	3	A	FixV	6	A	16.1095	0.996311	0.989675	15.4018	0.997811	0.992077
FixV	3	A	FixV	12	A	16.8972	0.995154	0.985755	16.2237	0.994032	0.987968
FixV	3	A	FixV	6	B	16.0807	0.994871	0.981410	15.3796	0.996804	0.987615
FixV	3	A	FixV	12	B	16.8662	0.993704	0.985328	16.2211	0.994318	0.987803
FixV	6	A	FixV	12	A	13.9433	0.998487	0.994702	13.5609	0.998669	0.994271
FixV	6	A	FixV	6	B	13.5520	0.998985	0.996745	12.8729	0.999803	0.998168
FixV	6	A	FixV	12	B	14.6294	0.997285	0.990450	13.9997	0.998679	0.993871
FixV	12	A	FixV	12	B	13.1413	0.999661	0.996931	12.7936	0.999857	0.997292
FixV	6	B	FixV	12	A	14.5943	0.997444	0.993752	13.9744	0.998571	0.994974
FixV	6	B	FixV	12	B	13.9193	0.997689	0.992755	13.5523	0.998508	0.993691
RTRC	GCat	1k	GL	3	A	17.7485	0.012554	0.355610	17.6366	0.268745	0.513275
RTRC	GCat	1k	GL	6	A	18.3241	0.057062	0.435285	18.2031	0.487731	0.660265
RTRC	GCat	1k	GL	9	A	18.8792	0.625279	0.789977	18.7735	0.604277	0.787189
RTRC	GCat	1k	GL	12	A	19.1668	0.471550	0.645542	19.0068	0.255475	0.541827
RTRC	GCat	1k	GL	Col	A	19.4271	0.151786	0.273684	19.3266	0.336159	0.633381
RTRC	GCat	1k	GL	3	B	17.7377	-0.017060	0.220582	17.6288	0.248687	0.500129
RTRC	GCat	1k	GL	6	B	18.2954	0.720602	0.834180	18.1840	0.375665	0.634567
RTRC	GCat	1k	GL	9	B	18.8529	0.575123	0.756583	18.7588	0.647436	0.792256
RTRC	GCat	1k	GL	12	B	19.1866	0.183958	0.583008	19.0770	0.544133	0.674793

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
RTRC	GCat	1k	GL	Col	B	19.4355	0.707615	0.867253	19.3287	0.596127	0.793408
RTRC	GCat	1k	GL	3	C	17.7248	0.731519	0.820424	17.6403	0.012071	0.267550
RTRC	GCat	1k	GL	6	C	18.3247	0.059891	0.320869	18.1969	0.511435	0.758093
RTRC	GCat	1k	GL	9	C	18.8724	0.054110	0.465738	18.7499	0.546852	0.710212
RTRC	GCat	1k	GL	12	C	19.1843	0.676898	0.759587	19.0797	0.314897	0.585093
RTRC	GCat	1k	GL	Col	C	19.4001	0.717445	0.872638	19.2843	0.300226	0.633797
RTRC	GCat	1k	GL	3	D	17.7148	0.601578	0.774145	17.6476	0.555730	0.814777
RTRC	GCat	1k	GL	6	D	18.2621	0.560554	0.735858	18.1735	0.567455	0.780984
RTRC	GCat	1k	GL	9	D	18.8891	0.184021	0.540269	18.7954	0.506925	0.703100
RTRC	GCat	1k	GL	12	D	19.2117	0.605586	0.792609	19.1442	0.530190	0.770191
RTRC	GCat	1k	GL	Col	D	19.4655	0.769650	0.825171	19.3626	0.213189	0.606985
RTRC	GCat	1k	RTRC	GCat150k	B	19.8142	0.802022	0.812383	19.4715	0.871342	0.939801
RTRC	GCat	1k	RTRC	GCat		19.8572	0.838871	0.837949	19.5317	0.837371	0.890459
RTRC	GCat	1k	RTRC	GCat150k		19.8000	0.762248	0.850934	19.4862	0.811108	0.916361
RTRC	GCat150k		GL	3	A	18.2579	0.820906	0.794011	18.1628	0.732114	0.796978
RTRC	GCat150k		GL	6	A	18.6341	0.829764	0.807228	18.5233	0.683501	0.709054
RTRC	GCat150k		GL	9	A	19.0188	0.782695	0.790965	18.9432	0.712869	0.804517
RTRC	GCat150k		GL	12	A	19.2601	0.758290	0.826252	19.1709	0.477040	0.764307
RTRC	GCat150k		GL	Col	A	19.4900	0.416183	0.652349	19.3436	0.710180	0.832856
RTRC	GCat150k		GL	3	B	18.2796	0.883297	0.888109	18.1431	0.671332	0.728236
RTRC	GCat150k		GL	6	B	18.6008	0.640704	0.612652	18.4568	0.502611	0.681594
RTRC	GCat150k		GL	9	B	19.0556	0.379830	0.617841	18.9109	0.632508	0.747157
RTRC	GCat150k		GL	12	B	19.3095	0.745154	0.834191	19.1823	0.739211	0.790956
RTRC	GCat150k		GL	Col	B	19.4705	0.886657	0.914597	19.3415	0.794750	0.862163
RTRC	GCat150k		GL	3	C	18.2011	0.871677	0.847196	18.1042	0.638429	0.787070
RTRC	GCat150k		GL	6	C	18.6942	0.590877	0.746536	18.5634	0.764213	0.825102
RTRC	GCat150k		GL	9	C	19.0121	0.757559	0.814845	18.9340	0.369190	0.564930
RTRC	GCat150k		GL	12	C	19.3091	0.874110	0.837038	19.1841	0.689577	0.738858
RTRC	GCat150k		GL	Col	C	19.4454	0.620883	0.822840	19.3448	0.912573	0.908145

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
RTRC	GCat150k		GL	3	D	18.2315	0.948420	0.928154	18.1258	0.768229	0.849989
RTRC	GCat150k		GL	6	D	18.6585	0.725096	0.752871	18.5435	0.821747	0.856052
RTRC	GCat150k		GL	9	D	19.0492	0.530191	0.638718	18.9054	0.659436	0.718251
RTRC	GCat150k		GL	12	D	19.3232	0.853401	0.842640	19.1728	0.548559	0.697185
RTRC	GCat150k		GL	Col	D	19.4702	0.688027	0.824057	19.3254	0.663816	0.780323
RTRC	GCat150k	B	GL	3	A	18.2999	0.640539	0.614068	18.2027	0.054937	0.434178
RTRC	GCat150k	B	GL	6	A	18.7149	0.838154	0.841270	18.6055	0.561494	0.680348
RTRC	GCat150k	B	GL	9	A	19.1129	0.866817	0.891774	18.9761	0.826633	0.882561
RTRC	GCat150k	B	GL	12	A	19.3261	0.886525	0.890779	19.2104	0.899700	0.901011
RTRC	GCat150k	B	GL	Col	A	19.5167	0.661218	0.773692	19.3928	0.802317	0.845898
RTRC	GCat150k	B	GL	3	B	18.2865	0.919611	0.901446	18.2228	0.472723	0.704098
RTRC	GCat150k	B	GL	6	B	18.6902	0.934631	0.935482	18.6046	0.620256	0.740607
RTRC	GCat150k	B	GL	9	B	19.0867	0.845086	0.881749	18.9876	0.786892	0.813354
RTRC	GCat150k	B	GL	12	B	19.3270	0.805921	0.855910	19.2358	0.847156	0.844927
RTRC	GCat150k	B	GL	Col	B	19.5028	0.854765	0.916050	19.4091	0.765257	0.851421
RTRC	GCat150k	B	GL	3	C	18.2925	0.841218	0.852950	18.1885	0.440468	0.681468
RTRC	GCat150k	B	GL	6	C	18.7231	0.825212	0.820835	18.6060	0.610918	0.745222
RTRC	GCat150k	B	GL	9	C	19.0930	0.755169	0.844095	18.9619	0.757277	0.780420
RTRC	GCat150k	B	GL	12	C	19.3533	0.930510	0.895857	19.2605	0.795150	0.797889
RTRC	GCat150k	B	GL	Col	C	19.5148	0.956453	0.952486	19.3883	0.827898	0.851872
RTRC	GCat150k	B	GL	3	D	18.2860	0.950228	0.928163	18.1891	0.514666	0.747652
RTRC	GCat150k	B	GL	6	D	18.6852	0.882408	0.872829	18.5942	0.883096	0.884491
RTRC	GCat150k	B	GL	9	D	19.1045	0.853805	0.861477	18.9620	0.817111	0.858663
RTRC	GCat150k	B	GL	12	D	19.3678	0.589241	0.717396	19.2183	0.900934	0.895666
RTRC	GCat150k	B	GL	Col	D	19.5171	0.842162	0.837758	19.4182	0.759864	0.830605
RTRC	GCat		GL	3	A	18.3162	0.821974	0.836979	18.1926	0.330324	0.620010
RTRC	GCat		GL	6	A	18.7333	0.855088	0.869959	18.6293	0.438022	0.578377
RTRC	GCat		GL	9	A	19.1309	0.849533	0.883069	19.0196	0.697871	0.833898
RTRC	GCat		GL	12	A	19.3433	0.780605	0.767254	19.2421	0.735627	0.793468

Table A – 4 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	NICH04 OTV-Norm	NICHD04 Doc Cent	NICHD04 Term Cent	T-1000 OTV-Norm	T-1000 Doc Cent	T-1000 Term Cent
RTRC	GCat		GL	Col	A	19.5177	0.665445	0.812218	19.4039	0.859303	0.901388
RTRC	GCat		GL	3	B	18.3058	0.892677	0.886412	18.1966	0.452075	0.673847
RTRC	GCat		GL	6	B	18.7033	0.922608	0.917905	18.5860	0.860739	0.898180
RTRC	GCat		GL	9	B	19.0858	0.830309	0.850838	18.9539	0.673224	0.767073
RTRC	GCat		GL	12	B	19.3656	0.728508	0.805165	19.2586	0.754902	0.789567
RTRC	GCat		GL	Col	B	19.5121	0.768427	0.889877	19.4165	0.634103	0.746186
RTRC	GCat		GL	3	C	18.2918	0.876361	0.870612	18.2172	0.388505	0.602175
RTRC	GCat		GL	6	C	18.7392	0.635215	0.715244	18.6476	0.700464	0.808352
RTRC	GCat		GL	9	C	19.1051	0.827247	0.859919	18.9982	0.587908	0.659908
RTRC	GCat		GL	12	C	19.3806	0.856096	0.818564	19.2504	0.518759	0.672235
RTRC	GCat		GL	Col	C	19.5171	0.774932	0.840016	19.3417	0.677696	0.771454
RTRC	GCat		GL	3	D	18.2621	0.903924	0.899095	18.1786	0.622823	0.776022
RTRC	GCat		GL	6	D	18.7041	0.738424	0.766910	18.5812	0.802520	0.851542
RTRC	GCat		GL	9	D	19.1068	0.011420	0.236841	18.9981	0.826483	0.859742
RTRC	GCat		GL	12	D	19.3391	0.693764	0.779864	19.2617	0.783171	0.819476
RTRC	GCat		GL	Col	D	19.5061	0.711730	0.726569	19.4307	0.590984	0.733745

Table A - 5: Space comparison results for all experiments detailing OTV-Norm, document centroid comparison cosine, and the term centroid comparison cosine for two of the four anchor sets used: T-500, and T-5000. Group column values: GL = grade level, GLNO = grade level non-overlapping, Large = large size 6th and 9th grade sets, FixV = fixed vocabulary sets. Set column indicates the grade level or specific RTRC set used in the comparison.

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
RTRC	GCat		RTRC	GCat	mod	0.0047	1.000000	1.000000	0.0047	1.000000	1.000000
RTRC	GCat150k	B	RTRC	GCat		12.6335	0.998507	0.988187	12.4489	0.998456	0.989067
RTRC	GCat150k		RTRC	GCat150k	B	13.4731	0.996463	0.987587	13.0320	0.998610	0.988759
RTRC	GCat150k		RTRC	GCat		13.8665	0.997205	0.989829	13.3301	0.997830	0.990624
GL	3	A	GL	6	A	14.4074	0.998071	0.991359	13.7214	0.999119	0.992887
GL	3	A	GL	9	A	15.2201	0.997281	0.992325	14.3545	0.998850	0.993510
GL	3	A	GL	12	A	15.7207	0.994301	0.980068	14.7820	0.997716	0.987981
GL	3	A	GL	Col	A	16.2366	0.986922	0.955372	15.2741	0.992453	0.976631
GL	3	A	GL	3	B	14.5898	0.998724	0.994638	13.8210	0.999836	0.995554
GL	3	A	GL	6	B	15.1431	0.997920	0.993832	14.2626	0.999241	0.994890
GL	3	A	GL	9	B	15.5618	0.995362	0.987039	14.6166	0.997921	0.991451
GL	3	A	GL	12	B	15.9459	0.992369	0.972210	14.9643	0.995123	0.977862
GL	3	A	GL	Col	B	16.3792	0.983858	0.962523	15.4253	0.988753	0.976475
GL	3	A	GL	3	C	14.6175	0.998628	0.994270	13.8813	0.999546	0.995010
GL	3	A	GL	6	C	15.1391	0.997561	0.994146	14.2553	0.999297	0.995613
GL	3	A	GL	9	C	15.5753	0.996250	0.988755	14.6197	0.997776	0.991161
GL	3	A	GL	12	C	15.9528	0.992374	0.972609	14.9957	0.995123	0.980351
GL	3	A	GL	Col	C	16.4062	0.979673	0.958220	15.4338	0.987474	0.975711
GL	3	A	GL	3	D	14.6482	0.998593	0.992172	13.8756	0.999596	0.992619
GL	3	A	GL	6	D	15.2105	0.997842	0.990715	14.3130	0.999522	0.990997
GL	3	A	GL	9	D	15.6217	0.996713	0.986939	14.6920	0.998799	0.990083
GL	3	A	GL	12	D	15.9932	0.993114	0.979857	15.0476	0.996621	0.985557
GL	3	A	GL	Col	D	16.3772	0.986447	0.951483	15.4172	0.989815	0.964095
GL	6	A	GL	9	A	13.5869	0.999110	0.992963	13.1785	0.999586	0.992886
GL	6	A	GL	12	A	14.4164	0.995044	0.981743	13.7701	0.998822	0.986418

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GL	6	A	GL	Col	A	15.2388	0.984764	0.973555	14.4673	0.994197	0.983853
GL	6	A	GL	3	B	15.1620	0.998299	0.991146	14.2901	0.999365	0.992098
GL	6	A	GL	6	B	13.6488	0.999596	0.993330	13.1658	0.999824	0.992971
GL	6	A	GL	9	B	14.0720	0.997989	0.992472	13.5083	0.999405	0.994330
GL	6	A	GL	12	B	14.7703	0.995754	0.982280	14.0417	0.996846	0.985341
GL	6	A	GL	Col	B	15.4392	0.982337	0.965898	14.6208	0.991970	0.986470
GL	6	A	GL	3	C	15.2050	0.997393	0.992408	14.3168	0.999066	0.993836
GL	6	A	GL	6	C	13.7509	0.999386	0.993899	13.2734	0.999846	0.993829
GL	6	A	GL	9	C	14.0582	0.998040	0.993072	13.4746	0.999274	0.995316
GL	6	A	GL	12	C	14.7494	0.993539	0.981782	14.0138	0.997653	0.988711
GL	6	A	GL	Col	C	15.4162	0.980873	0.963109	14.6009	0.991209	0.984662
GL	6	A	GL	3	D	15.1882	0.997901	0.990162	14.3158	0.998955	0.990518
GL	6	A	GL	6	D	13.6888	0.999506	0.989855	13.2293	0.999677	0.989454
GL	6	A	GL	9	D	14.1751	0.998649	0.990555	13.5532	0.999572	0.991614
GL	6	A	GL	12	D	14.9012	0.995228	0.982429	14.1732	0.998398	0.985738
GL	6	A	GL	Col	D	15.3436	0.987633	0.970792	14.5317	0.992631	0.975377
GL	9	A	GL	12	A	13.1514	0.998849	0.992206	12.8404	0.999308	0.992733
GL	9	A	GL	Col	A	14.0676	0.993435	0.985450	13.5242	0.996675	0.989454
GL	9	A	GL	3	B	15.6311	0.990702	0.973711	14.6827	0.995940	0.979360
GL	9	A	GL	6	B	14.1354	0.997637	0.984255	13.5036	0.998352	0.983922
GL	9	A	GL	9	B	13.3214	0.999167	0.982474	12.9570	0.999518	0.983048
GL	9	A	GL	12	B	13.7798	0.998124	0.979434	13.3039	0.998901	0.979998
GL	9	A	GL	Col	B	14.3812	0.992120	0.982674	13.7697	0.995495	0.985104
GL	9	A	GL	3	C	15.5916	0.992958	0.974197	14.6484	0.995551	0.974138
GL	9	A	GL	6	C	14.2549	0.997970	0.984972	13.6408	0.998264	0.983308
GL	9	A	GL	9	C	13.2041	0.999150	0.983067	12.8641	0.999496	0.983859
GL	9	A	GL	12	C	13.8151	0.998340	0.982295	13.3379	0.999063	0.983601
GL	9	A	GL	Col	C	14.4453	0.993910	0.982803	13.8220	0.995498	0.986288
GL	9	A	GL	3	D	15.5923	0.992662	0.987820	14.6346	0.996697	0.989017

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GL	9	A	GL	6	D	14.1772	0.998736	0.991632	13.5567	0.999051	0.991050
GL	9	A	GL	9	D	13.2663	0.999540	0.996160	12.9324	0.999888	0.996647
GL	9	A	GL	12	D	13.5924	0.998765	0.995181	13.1203	0.999536	0.995746
GL	9	A	GL	Col	D	14.2714	0.995951	0.986616	13.6658	0.996126	0.987341
GL	12	A	GL	Col	A	13.0959	0.999060	0.994847	12.8139	0.999541	0.995690
GL	12	A	GL	3	B	15.9697	0.984876	0.959271	15.0449	0.986943	0.960076
GL	12	A	GL	6	B	14.8709	0.992736	0.972753	14.1458	0.993844	0.971030
GL	12	A	GL	9	B	13.9820	0.997153	0.974956	13.5190	0.997427	0.976068
GL	12	A	GL	12	B	13.1639	0.998499	0.975648	12.8718	0.999233	0.976658
GL	12	A	GL	Col	B	13.5626	0.997345	0.976545	13.1782	0.998746	0.978800
GL	12	A	GL	3	C	15.9724	0.986309	0.953328	14.9942	0.985875	0.946415
GL	12	A	GL	6	C	14.8179	0.994612	0.973671	14.0744	0.993558	0.967646
GL	12	A	GL	9	C	13.6236	0.997537	0.977873	13.1616	0.997606	0.976660
GL	12	A	GL	12	C	13.0073	0.998709	0.978493	12.7190	0.999280	0.979784
GL	12	A	GL	Col	C	13.4710	0.997803	0.978557	13.0510	0.998681	0.978962
GL	12	A	GL	3	D	16.0171	0.985024	0.975715	15.0590	0.989035	0.973700
GL	12	A	GL	6	D	14.8974	0.994154	0.984125	14.1682	0.995754	0.984318
GL	12	A	GL	9	D	13.6725	0.998874	0.992330	13.2263	0.998900	0.991735
GL	12	A	GL	12	D	13.2245	0.999054	0.993569	12.9446	0.999895	0.993438
GL	12	A	GL	Col	D	13.5185	0.997520	0.989325	13.1143	0.999379	0.992355
GL	Col	A	GL	3	B	16.5037	0.967221	0.939431	15.5332	0.966512	0.933344
GL	Col	A	GL	6	B	15.4477	0.984037	0.971421	14.6322	0.983682	0.951362
GL	Col	A	GL	9	B	14.4318	0.993139	0.964518	13.8244	0.991917	0.960326
GL	Col	A	GL	12	B	13.4655	0.995840	0.958498	13.0752	0.996384	0.959987
GL	Col	A	GL	Col	B	13.0478	0.998022	0.965033	12.8337	0.998542	0.964810
GL	Col	A	GL	3	C	16.4385	0.968100	0.928650	15.4462	0.964200	0.912795
GL	Col	A	GL	6	C	15.4477	0.985514	0.958353	14.6065	0.981642	0.944858
GL	Col	A	GL	9	C	14.3651	0.992157	0.965464	13.7479	0.992147	0.959971
GL	Col	A	GL	12	C	13.3899	0.996046	0.962983	13.0020	0.996017	0.963400

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GL	Col	A	GL	Col	C	13.0413	0.998335	0.966892	12.8242	0.998545	0.966856
GL	Col	A	GL	3	D	16.3480	0.958383	0.946236	15.3737	0.967312	0.938483
GL	Col	A	GL	6	D	15.4682	0.984320	0.966262	14.6371	0.984334	0.959684
GL	Col	A	GL	9	D	14.3437	0.994383	0.985986	13.7402	0.994062	0.983602
GL	Col	A	GL	12	D	13.3645	0.998696	0.990274	12.9958	0.998431	0.989797
GL	Col	A	GL	Col	D	12.9322	0.999703	0.992762	12.7027	0.999860	0.993172
GL	3	B	GL	6	B	14.4086	0.998318	0.994914	13.6741	0.999380	0.995689
GL	3	B	GL	9	B	15.2855	0.992677	0.982131	14.4205	0.997059	0.989731
GL	3	B	GL	12	B	15.8251	0.991775	0.973393	14.9379	0.992226	0.970364
GL	3	B	GL	Col	B	16.2744	0.981027	0.966050	15.3287	0.987431	0.977905
GL	3	B	GL	3	C	14.5137	0.998336	0.993271	13.7618	0.999412	0.994945
GL	3	B	GL	6	C	15.1441	0.997508	0.994886	14.2719	0.999141	0.996668
GL	3	B	GL	9	C	15.6440	0.993310	0.981163	14.6959	0.997665	0.990169
GL	3	B	GL	12	C	16.0381	0.989562	0.968776	15.0936	0.992461	0.972124
GL	3	B	GL	Col	C	16.5000	0.984381	0.968943	15.5340	0.986797	0.974027
GL	3	B	GL	3	D	14.4668	0.997975	0.993622	13.7064	0.999716	0.994409
GL	3	B	GL	6	D	15.1406	0.997434	0.990670	14.2470	0.999562	0.993028
GL	3	B	GL	9	D	15.6232	0.996612	0.989722	14.6943	0.998708	0.992256
GL	3	B	GL	12	D	16.0645	0.993804	0.980780	15.1381	0.995289	0.982482
GL	3	B	GL	Col	D	16.4172	0.977395	0.929428	15.4906	0.984951	0.937211
GL	6	B	GL	9	B	13.5889	0.998061	0.991066	13.1340	0.999093	0.992928
GL	6	B	GL	12	B	14.4813	0.995343	0.976748	13.8148	0.996185	0.979082
GL	6	B	GL	Col	B	15.2251	0.979626	0.961960	14.4254	0.989207	0.980726
GL	6	B	GL	3	C	15.1115	0.997128	0.985417	14.2484	0.998173	0.985805
GL	6	B	GL	6	C	13.6009	0.999068	0.996348	13.1304	0.999506	0.996702
GL	6	B	GL	9	C	14.1282	0.997915	0.989967	13.4978	0.998808	0.991892
GL	6	B	GL	12	C	14.9263	0.994487	0.979529	14.1591	0.995486	0.981015
GL	6	B	GL	Col	C	15.4354	0.977766	0.953436	14.5978	0.987408	0.978240
GL	6	B	GL	3	D	15.1949	0.996082	0.987455	14.3093	0.998810	0.990134

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GL	6	B	GL	6	D	13.4930	0.998971	0.991142	12.9785	0.999689	0.991716
GL	6	B	GL	9	D	14.2656	0.997833	0.989003	13.6266	0.999426	0.990910
GL	6	B	GL	12	D	14.7973	0.995430	0.989198	14.0642	0.997780	0.990110
GL	6	B	GL	Col	D	15.4361	0.981726	0.957163	14.6209	0.990680	0.970587
GL	9	B	GL	12	B	13.2546	0.998848	0.991107	12.9399	0.999190	0.991362
GL	9	B	GL	Col	B	14.1399	0.989082	0.979401	13.6000	0.996093	0.990423
GL	9	B	GL	3	C	15.5653	0.994308	0.982837	14.6002	0.997568	0.984345
GL	9	B	GL	6	C	14.1822	0.998033	0.994262	13.5413	0.999449	0.993994
GL	9	B	GL	9	C	13.1668	0.999651	0.996553	12.8174	0.999920	0.997126
GL	9	B	GL	12	C	13.7082	0.997775	0.990694	13.2209	0.999040	0.993051
GL	9	B	GL	Col	C	14.3358	0.988994	0.977247	13.6851	0.996736	0.991861
GL	9	B	GL	3	D	15.5927	0.994419	0.983984	14.6709	0.997574	0.984280
GL	9	B	GL	6	D	14.0474	0.998281	0.986306	13.4248	0.999331	0.986395
GL	9	B	GL	9	D	13.4956	0.999509	0.984194	13.1559	0.999556	0.983798
GL	9	B	GL	12	D	13.8554	0.998724	0.986283	13.3769	0.999640	0.986610
GL	9	B	GL	Col	D	14.3842	0.994223	0.986123	13.7596	0.997856	0.990027
GL	12	B	GL	Col	B	13.1427	0.997492	0.991152	12.8808	0.997832	0.991374
GL	12	B	GL	3	C	15.9751	0.993070	0.985244	14.9841	0.997721	0.989012
GL	12	B	GL	6	C	14.8421	0.996783	0.991120	14.0985	0.999093	0.990786
GL	12	B	GL	9	C	13.6951	0.999003	0.994558	13.2335	0.999520	0.994831
GL	12	B	GL	12	C	12.9279	0.999269	0.993564	12.6566	0.999856	0.994628
GL	12	B	GL	Col	C	13.5250	0.990754	0.981617	13.1162	0.996712	0.990708
GL	12	B	GL	3	D	16.0605	0.991840	0.977923	15.0813	0.997116	0.980564
GL	12	B	GL	6	D	14.8864	0.996719	0.977759	14.1232	0.998313	0.976213
GL	12	B	GL	9	D	13.5314	0.998526	0.979836	13.0480	0.999044	0.981044
GL	12	B	GL	12	D	13.0539	0.999146	0.979108	12.7787	0.999219	0.979083
GL	12	B	GL	Col	D	13.4992	0.998554	0.985169	13.1259	0.999350	0.985658
GL	Col	B	GL	3	C	16.3966	0.981320	0.965565	15.4278	0.991613	0.969750
GL	Col	B	GL	6	C	15.3630	0.994009	0.981838	14.5428	0.995454	0.982531

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GL	Col	B	GL	9	C	14.2984	0.995496	0.984671	13.6926	0.997704	0.986947
GL	Col	B	GL	12	C	13.4935	0.997974	0.986973	13.0789	0.999316	0.989300
GL	Col	B	GL	Col	C	12.7105	0.999761	0.994106	12.5021	0.999913	0.994287
GL	Col	B	GL	3	D	16.3680	0.977016	0.964484	15.3782	0.992164	0.972397
GL	Col	B	GL	6	D	15.4243	0.989830	0.967105	14.6141	0.994134	0.967940
GL	Col	B	GL	9	D	14.3490	0.993711	0.975778	13.7253	0.996720	0.975705
GL	Col	B	GL	12	D	13.5761	0.997872	0.969571	13.1934	0.997982	0.971431
GL	Col	B	GL	Col	D	12.8335	0.998636	0.964556	12.6235	0.998576	0.964204
GL	3	C	GL	6	C	14.5240	0.998243	0.992909	13.8129	0.999421	0.994573
GL	3	C	GL	9	C	15.1967	0.994588	0.986052	14.3182	0.998309	0.994102
GL	3	C	GL	12	C	15.7247	0.991211	0.972889	14.7801	0.994848	0.980248
GL	3	C	GL	Col	C	16.3198	0.975900	0.949946	15.3549	0.986961	0.974662
GL	3	C	GL	3	D	14.5486	0.998875	0.994074	13.7875	0.999575	0.993781
GL	3	C	GL	6	D	15.2118	0.998284	0.989242	14.2893	0.999540	0.989953
GL	3	C	GL	9	D	15.5462	0.997055	0.986988	14.5801	0.998926	0.990531
GL	3	C	GL	12	D	16.0665	0.991807	0.979289	15.0871	0.996274	0.983187
GL	3	C	GL	Col	D	16.4050	0.972698	0.906453	15.4185	0.986447	0.944046
GL	6	C	GL	9	C	13.6075	0.998209	0.991600	13.1571	0.999182	0.994127
GL	6	C	GL	12	C	14.4989	0.993458	0.978371	13.8239	0.996556	0.985014
GL	6	C	GL	Col	C	15.1808	0.975719	0.952150	14.3859	0.989076	0.982996
GL	6	C	GL	3	D	15.2427	0.997575	0.989679	14.3346	0.998850	0.990082
GL	6	C	GL	6	D	13.6265	0.999361	0.991850	13.1474	0.999637	0.992349
GL	6	C	GL	9	D	14.1873	0.998400	0.990645	13.5489	0.999519	0.991965
GL	6	C	GL	12	D	14.8313	0.996347	0.991122	14.1045	0.997953	0.990575
GL	6	C	GL	Col	D	15.4434	0.985687	0.965571	14.6309	0.991489	0.973043
GL	9	C	GL	12	C	13.1509	0.997923	0.990917	12.8320	0.999000	0.992981
GL	9	C	GL	Col	C	13.8743	0.987407	0.972155	13.3271	0.995182	0.989932
GL	9	C	GL	3	D	15.6338	0.995561	0.984067	14.6853	0.997655	0.984768
GL	9	C	GL	6	D	14.2747	0.998459	0.984390	13.6416	0.999224	0.985009

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GL	9	C	GL	9	D	13.5299	0.999337	0.982492	13.1955	0.999588	0.982614
GL	9	C	GL	12	D	13.7676	0.998008	0.987499	13.3172	0.999600	0.987740
GL	9	C	GL	Col	D	14.3344	0.992965	0.986670	13.7341	0.997832	0.992529
GL	12	C	GL	Col	C	12.9071	0.997780	0.992323	12.6224	0.998895	0.994035
GL	12	C	GL	3	D	15.9919	0.993488	0.979798	15.0173	0.995280	0.980426
GL	12	C	GL	6	D	14.9054	0.996863	0.980199	14.1555	0.998031	0.979179
GL	12	C	GL	9	D	13.7369	0.998294	0.980086	13.2687	0.999071	0.980001
GL	12	C	GL	12	D	13.1590	0.999331	0.977233	12.9002	0.999295	0.976624
GL	12	C	GL	Col	D	13.4676	0.997618	0.985692	13.0762	0.999468	0.983905
GL	Col	C	GL	3	D	16.4263	0.985361	0.969786	15.4322	0.990949	0.970379
GL	Col	C	GL	6	D	15.4144	0.986928	0.965186	14.5613	0.995046	0.970390
GL	Col	C	GL	9	D	14.3464	0.994156	0.973818	13.7157	0.995777	0.974714
GL	Col	C	GL	12	D	13.5566	0.997618	0.970506	13.1589	0.997878	0.971144
GL	Col	C	GL	Col	D	12.8218	0.998446	0.965016	12.6022	0.998620	0.965059
GL	3	D	GL	6	D	14.4634	0.997971	0.991701	13.7363	0.999395	0.993230
GL	3	D	GL	9	D	15.1827	0.996033	0.989026	14.3375	0.998438	0.992925
GL	3	D	GL	12	D	15.7701	0.989883	0.969430	14.8339	0.994414	0.978030
GL	3	D	GL	Col	D	16.2467	0.969474	0.860672	15.2866	0.990144	0.947362
GL	6	D	GL	9	D	13.4691	0.999081	0.994613	13.0300	0.999507	0.995137
GL	6	D	GL	12	D	14.4534	0.994480	0.988944	13.7745	0.997820	0.994328
GL	6	D	GL	Col	D	15.2162	0.979272	0.936983	14.4259	0.989892	0.970937
GL	9	D	GL	12	D	13.1990	0.998512	0.994990	12.9022	0.999566	0.995558
GL	9	D	GL	Col	D	14.0666	0.994156	0.985251	13.5309	0.995642	0.988328
GL	12	D	GL	Col	D	12.9374	0.998778	0.993099	12.6765	0.999054	0.994165
Large	6	250kB	GL	3	A	15.5371	0.998431	0.992954	14.6433	0.999107	0.994041
Large	6	250kB	GL	6	A	14.6704	0.997636	0.990708	13.9844	0.999224	0.993311
Large	6	250kB	GL	9	A	14.0648	0.998472	0.994357	13.4788	0.999403	0.993223
Large	6	250kB	GL	12	A	14.3089	0.996571	0.989771	13.6761	0.998847	0.990275
Large	6	250kB	GL	Col	A	14.8894	0.989900	0.976891	14.1698	0.995305	0.985868

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
Large	6	250kB	GL	3	B	15.5386	0.997340	0.992601	14.6374	0.999254	0.993828
Large	6	250kB	GL	6	B	14.8498	0.997820	0.994193	14.1267	0.999449	0.995017
Large	6	250kB	GL	9	B	14.0856	0.997749	0.991176	13.4884	0.998977	0.993334
Large	6	250kB	GL	12	B	14.3150	0.994194	0.977267	13.6782	0.997023	0.984572
Large	6	250kB	GL	Col	B	14.8152	0.989914	0.981448	14.0869	0.993623	0.988426
Large	6	250kB	GL	3	C	15.6267	0.997116	0.991298	14.7210	0.999053	0.992602
Large	6	250kB	GL	6	C	14.6999	0.998413	0.994549	14.0089	0.999510	0.994736
Large	6	250kB	GL	9	C	14.1935	0.997875	0.991918	13.6073	0.998993	0.994136
Large	6	250kB	GL	12	C	14.2659	0.995252	0.981842	13.6468	0.996897	0.986546
Large	6	250kB	GL	Col	C	14.8881	0.989546	0.978479	14.1530	0.992007	0.985074
Large	6	250kB	GL	3	D	15.5476	0.997890	0.991330	14.6100	0.998904	0.991450
Large	6	250kB	GL	6	D	14.7699	0.998380	0.991797	14.0461	0.999565	0.990760
Large	6	250kB	GL	9	D	14.2312	0.998890	0.991881	13.6410	0.999419	0.991162
Large	6	250kB	GL	12	D	14.4130	0.997055	0.990265	13.7739	0.998658	0.991049
Large	6	250kB	GL	Col	D	14.8352	0.991279	0.980240	14.1131	0.994522	0.985261
Large	9	250kB	GL	3	A	15.8102	0.996352	0.989167	14.8642	0.998412	0.990875
Large	9	250kB	GL	6	A	14.7138	0.996843	0.990831	14.0042	0.998672	0.992460
Large	9	250kB	GL	9	A	13.6881	0.998982	0.987607	13.2450	0.999460	0.987284
Large	9	250kB	GL	12	A	13.5677	0.999370	0.986189	13.1737	0.999615	0.983962
Large	9	250kB	GL	Col	A	13.9963	0.995422	0.988879	13.4625	0.996205	0.986863
Large	9	250kB	GL	3	B	15.8687	0.995937	0.988818	14.9838	0.998182	0.989564
Large	9	250kB	GL	6	B	14.8643	0.998134	0.992375	14.1521	0.998733	0.991442
Large	9	250kB	GL	9	B	13.7300	0.999397	0.992875	13.2628	0.999615	0.992725
Large	9	250kB	GL	12	B	13.5142	0.997774	0.990086	13.1036	0.998647	0.992796
Large	9	250kB	GL	Col	B	14.0350	0.991917	0.986396	13.5032	0.995184	0.990936
Large	9	250kB	GL	3	C	15.8651	0.995527	0.988707	14.9019	0.997854	0.990135
Large	9	250kB	GL	6	C	14.8386	0.997616	0.993038	14.1322	0.999051	0.992685
Large	9	250kB	GL	9	C	13.6880	0.999076	0.993303	13.2256	0.999422	0.993830
Large	9	250kB	GL	12	C	13.4047	0.998063	0.992114	12.9890	0.999005	0.994066

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
Large	9	250kB	GL	Col	C	14.0222	0.991598	0.985288	13.4760	0.995634	0.992230
Large	9	250kB	GL	3	D	15.8452	0.995832	0.986070	14.8846	0.997555	0.986263
Large	9	250kB	GL	6	D	14.7303	0.996820	0.983860	14.0159	0.998868	0.983973
Large	9	250kB	GL	9	D	13.6433	0.999142	0.985692	13.1800	0.999470	0.985278
Large	9	250kB	GL	12	D	13.5699	0.999269	0.986838	13.2002	0.999579	0.986370
Large	9	250kB	GL	Col	D	14.1073	0.996187	0.990368	13.5955	0.996009	0.991504
GLNO	3	A	GL	6	250kB	15.5342	0.997960	0.992985	14.6701	0.999070	0.995287
GLNO	3	A	GL	9	250kB	15.8267	0.994706	0.981353	14.8732	0.997315	0.987482
GLNO	3	A	GL	3	A	14.2887	0.997076	0.992069	13.5201	0.999802	0.996467
GLNO	3	A	GL	6	A	15.1016	0.997323	0.989670	14.2193	0.998950	0.993309
GLNO	3	A	GL	9	A	15.5689	0.996182	0.989960	14.6314	0.998781	0.993934
GLNO	3	A	GL	12	A	15.9222	0.991521	0.975121	14.9694	0.997471	0.984699
GLNO	3	A	GL	Col	A	16.3858	0.980604	0.914487	15.4324	0.991292	0.969904
GLNO	3	A	GL	3	B	14.4946	0.999036	0.994044	13.7403	0.999531	0.994853
GLNO	3	A	GL	6	B	15.1452	0.997862	0.991306	14.2742	0.999307	0.993464
GLNO	3	A	GL	9	B	15.4776	0.994761	0.985122	14.5780	0.997899	0.992247
GLNO	3	A	GL	12	B	15.9079	0.989117	0.962787	14.9374	0.995787	0.981151
GLNO	3	A	GL	Col	B	16.3464	0.982459	0.968916	15.3654	0.991393	0.982972
GLNO	3	A	GL	3	C	14.4273	0.998872	0.994748	13.6454	0.999682	0.995616
GLNO	3	A	GL	6	C	15.0504	0.998142	0.993022	14.1747	0.999239	0.994281
GLNO	3	A	GL	9	C	15.4685	0.995977	0.987134	14.5177	0.998143	0.992710
GLNO	3	A	GL	12	C	15.9433	0.990849	0.966295	15.0027	0.996224	0.982429
GLNO	3	A	GL	Col	C	16.3854	0.975294	0.951271	15.4216	0.987966	0.977631
GLNO	3	A	GL	3	D	14.5367	0.998332	0.993281	13.7642	0.999688	0.993266
GLNO	3	A	GL	6	D	15.1183	0.997575	0.990309	14.2293	0.999588	0.991115
GLNO	3	A	GL	9	D	15.5451	0.996873	0.986616	14.6176	0.998785	0.991408
GLNO	3	A	GL	12	D	15.9847	0.991797	0.980145	15.0343	0.996614	0.983739
GLNO	3	A	GL	Col	D	16.3729	0.980404	0.912623	15.3904	0.991584	0.955532
GLNO	3	A	GLNO	6	A	14.5069	0.997471	0.989567	13.7838	0.999391	0.991976

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GLNO	3	A	GLNO	9	A	15.1306	0.996201	0.989722	14.2774	0.998159	0.992676
GLNO	3	A	GLNO	12	A	15.7213	0.992247	0.982406	14.7962	0.996671	0.990544
GLNO	3	A	GLNO	Col	A	16.1691	0.981100	0.953690	15.2138	0.989715	0.974425
GLNO	3	A	GLNO	3	B	15.7083	0.997484	0.992688	14.6341	0.999509	0.993683
GLNO	3	A	GLNO	6	B	15.5435	0.997468	0.993345	14.5526	0.999183	0.994400
GLNO	3	A	GLNO	9	B	15.8766	0.996461	0.988485	14.8691	0.998600	0.990851
GLNO	3	A	GLNO	12	B	16.1590	0.991387	0.983737	15.1804	0.996849	0.987895
GLNO	3	A	GLNO	Col	B	16.5092	0.980774	0.907501	15.5239	0.991777	0.931697
GLNO	6	A	GL	6	250kB	14.7187	0.996646	0.990282	13.9975	0.998866	0.993459
GLNO	6	A	GL	9	250kB	14.7681	0.994282	0.978790	14.0485	0.998365	0.987472
GLNO	6	A	GL	3	A	15.1371	0.997714	0.990517	14.2223	0.999077	0.992646
GLNO	6	A	GL	6	A	13.8731	0.998976	0.991294	13.3718	0.999800	0.992604
GLNO	6	A	GL	9	A	14.1584	0.998241	0.991820	13.5333	0.999539	0.993017
GLNO	6	A	GL	12	A	14.9203	0.994203	0.987962	14.1667	0.998495	0.988842
GLNO	6	A	GL	Col	A	15.4592	0.987406	0.974235	14.6201	0.992271	0.982191
GLNO	6	A	GL	3	B	15.0453	0.998018	0.993595	14.1454	0.999071	0.994536
GLNO	6	A	GL	6	B	13.3441	0.999525	0.996960	12.8495	0.999882	0.997437
GLNO	6	A	GL	9	B	13.9948	0.997708	0.991768	13.3777	0.999187	0.993905
GLNO	6	A	GL	12	B	14.8720	0.993527	0.972428	14.1093	0.996393	0.978777
GLNO	6	A	GL	Col	B	15.3918	0.981219	0.967018	14.5544	0.989307	0.980066
GLNO	6	A	GL	3	C	15.1152	0.997735	0.986860	14.2440	0.998812	0.987870
GLNO	6	A	GL	6	C	13.5923	0.999424	0.996343	13.1111	0.999895	0.996801
GLNO	6	A	GL	9	C	14.3003	0.997126	0.988704	13.6647	0.999199	0.992809
GLNO	6	A	GL	12	C	14.8611	0.991980	0.973818	14.1118	0.996814	0.985561
GLNO	6	A	GL	Col	C	15.4122	0.985989	0.976786	14.5628	0.987167	0.977374
GLNO	6	A	GL	3	D	15.2311	0.997155	0.989705	14.3267	0.998832	0.990742
GLNO	6	A	GL	6	D	13.5212	0.998976	0.991351	13.0019	0.999701	0.992193
GLNO	6	A	GL	9	D	14.1062	0.997736	0.990153	13.4433	0.999612	0.991562
GLNO	6	A	GL	12	D	14.8337	0.995249	0.988812	14.0637	0.997859	0.989819

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GLNO	6	A	GL	Col	D	15.4280	0.984027	0.956570	14.6128	0.991838	0.977312
GLNO	6	A	GLNO	9	A	13.4990	0.998336	0.995856	13.0609	0.999359	0.996681
GLNO	6	A	GLNO	12	A	14.5237	0.993423	0.986831	13.8450	0.997153	0.992129
GLNO	6	A	GLNO	Col	A	15.2341	0.980487	0.951332	14.4439	0.991094	0.973493
GLNO	6	A	GLNO	3	B	15.6878	0.997658	0.989064	14.6552	0.998847	0.990257
GLNO	6	A	GLNO	6	B	14.5361	0.998758	0.990761	13.7450	0.999445	0.991483
GLNO	6	A	GLNO	9	B	14.7463	0.998061	0.990211	13.9615	0.999564	0.991353
GLNO	6	A	GLNO	12	B	15.1714	0.996322	0.988120	14.3418	0.997695	0.986431
GLNO	6	A	GLNO	Col	B	15.5967	0.984501	0.944856	14.7458	0.992949	0.964085
GLNO	9	A	GL	6	250kB	14.0368	0.998219	0.991102	13.4542	0.999082	0.991626
GLNO	9	A	GL	9	250kB	13.7102	0.998891	0.990466	13.2882	0.999573	0.991971
GLNO	9	A	GL	3	A	15.5969	0.995861	0.986575	14.6490	0.997755	0.987406
GLNO	9	A	GL	6	A	14.1792	0.997941	0.988637	13.5754	0.999101	0.989303
GLNO	9	A	GL	9	A	13.2568	0.999399	0.985490	12.8979	0.999649	0.985604
GLNO	9	A	GL	12	A	13.8076	0.999260	0.985697	13.3512	0.999613	0.985702
GLNO	9	A	GL	Col	A	14.3918	0.993016	0.985507	13.7675	0.996022	0.986811
GLNO	9	A	GL	3	B	15.5627	0.996099	0.990696	14.6309	0.997750	0.991439
GLNO	9	A	GL	6	B	14.1702	0.998134	0.993355	13.5296	0.999355	0.993961
GLNO	9	A	GL	9	B	13.2403	0.999642	0.995220	12.9121	0.999879	0.995596
GLNO	9	A	GL	12	B	13.5934	0.998527	0.988417	13.1491	0.999027	0.989085
GLNO	9	A	GL	Col	B	14.3160	0.990952	0.981214	13.7159	0.992435	0.984759
GLNO	9	A	GL	3	C	15.5662	0.995422	0.982518	14.5883	0.997502	0.982876
GLNO	9	A	GL	6	C	14.0583	0.998551	0.993844	13.4648	0.999310	0.994078
GLNO	9	A	GL	9	C	13.1107	0.999560	0.994027	12.7824	0.999904	0.995285
GLNO	9	A	GL	12	C	13.7019	0.997851	0.989706	13.2393	0.999205	0.992841
GLNO	9	A	GL	Col	C	14.3536	0.988008	0.974359	13.7220	0.993010	0.986248
GLNO	9	A	GL	3	D	15.6264	0.995246	0.985109	14.6698	0.997358	0.984761
GLNO	9	A	GL	6	D	14.1526	0.997643	0.986747	13.5117	0.999184	0.987036
GLNO	9	A	GL	9	D	13.2163	0.999406	0.986091	12.8597	0.999609	0.985507

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GLNO	9	A	GL	12	D	13.6842	0.998810	0.987454	13.2173	0.999404	0.988540
GLNO	9	A	GL	Col	D	14.2826	0.992180	0.984042	13.6704	0.996349	0.990324
GLNO	9	A	GLNO	12	A	13.1879	0.998981	0.994791	12.8822	0.999230	0.995141
GLNO	9	A	GLNO	Col	A	14.0304	0.988830	0.970249	13.5094	0.994532	0.984169
GLNO	9	A	GLNO	3	B	15.9916	0.995637	0.983178	14.9737	0.997906	0.984226
GLNO	9	A	GLNO	6	B	14.6445	0.997878	0.987357	13.8624	0.999088	0.988782
GLNO	9	A	GLNO	9	B	13.7910	0.998823	0.988067	13.2180	0.999595	0.988938
GLNO	9	A	GLNO	12	B	14.2255	0.997790	0.985222	13.6292	0.999447	0.985831
GLNO	9	A	GLNO	Col	B	14.5683	0.992581	0.972211	13.8936	0.997080	0.980297
GLNO	12	A	GL	3	A	15.9890	0.992568	0.981638	15.0135	0.996792	0.984211
GLNO	12	A	GL	6	A	14.8879	0.994701	0.982132	14.1356	0.997960	0.985013
GLNO	12	A	GL	9	A	13.5992	0.998386	0.981637	13.1471	0.998916	0.981695
GLNO	12	A	GL	12	A	13.1663	0.999153	0.977740	12.8997	0.999413	0.978292
GLNO	12	A	GL	Col	A	13.5650	0.998935	0.983844	13.1433	0.999354	0.984151
GLNO	12	A	GL	3	B	16.0327	0.993912	0.984414	15.0790	0.997119	0.985384
GLNO	12	A	GL	6	B	14.7876	0.996268	0.989725	14.0230	0.998383	0.989599
GLNO	12	A	GL	9	B	13.6587	0.998562	0.993001	13.1912	0.999479	0.994020
GLNO	12	A	GL	12	B	13.1672	0.999596	0.995584	12.8999	0.999899	0.996011
GLNO	12	A	GL	Col	B	13.5863	0.996961	0.988381	13.1865	0.997369	0.989544
GLNO	12	A	GL	3	C	15.9600	0.993149	0.981164	14.9999	0.996052	0.980609
GLNO	12	A	GL	6	C	14.7968	0.995643	0.989901	14.0653	0.998633	0.988910
GLNO	12	A	GL	9	C	13.6737	0.998445	0.992428	13.1940	0.999424	0.993820
GLNO	12	A	GL	12	C	13.0328	0.999644	0.994295	12.7337	0.999885	0.994867
GLNO	12	A	GL	Col	C	13.5815	0.996942	0.989838	13.1607	0.998047	0.992162
GLNO	12	A	GL	3	D	16.0260	0.991502	0.978217	15.0685	0.995764	0.980920
GLNO	12	A	GL	6	D	14.7598	0.996230	0.978801	14.0044	0.998264	0.980031
GLNO	12	A	GL	9	D	13.7514	0.998430	0.980457	13.3088	0.999044	0.980616
GLNO	12	A	GL	12	D	13.0539	0.999148	0.978381	12.7727	0.999301	0.979104
GLNO	12	A	GL	Col	D	13.4695	0.998062	0.986725	13.0485	0.999592	0.986227

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GLNO	12	A	GLNO	Col	A	13.0274	0.997693	0.992135	12.7531	0.998744	0.994405
GLNO	12	A	GLNO	3	B	16.2493	0.994969	0.978866	15.2456	0.997127	0.980397
GLNO	12	A	GLNO	6	B	15.2586	0.996610	0.981719	14.4342	0.997881	0.981933
GLNO	12	A	GLNO	9	B	14.2267	0.998094	0.977898	13.6380	0.998924	0.978665
GLNO	12	A	GLNO	12	B	13.6760	0.999057	0.980374	13.2326	0.999202	0.978747
GLNO	12	A	GLNO	Col	B	13.8689	0.998683	0.979422	13.3530	0.999480	0.977711
GLNO	12	A	Large	6	250kB	14.2511	0.998033	0.988217	13.6192	0.998618	0.988281
GLNO	12	A	Large	9	250kB	13.6067	0.999079	0.990034	13.2242	0.999314	0.990067
GLNO	Col	A	GL	6	250kB	14.8308	0.992313	0.979738	14.0881	0.996145	0.982336
GLNO	Col	A	GL	9	250kB	13.9956	0.996986	0.984592	13.4678	0.998128	0.986549
GLNO	Col	A	GL	3	A	16.3607	0.987416	0.969284	15.3901	0.993912	0.973394
GLNO	Col	A	GL	6	A	15.4253	0.991185	0.974890	14.6301	0.996081	0.978509
GLNO	Col	A	GL	9	A	14.3344	0.995738	0.975611	13.7131	0.997261	0.976158
GLNO	Col	A	GL	12	A	13.4916	0.997380	0.966410	13.0967	0.998299	0.966439
GLNO	Col	A	GL	Col	A	13.0323	0.998537	0.963075	12.7962	0.998714	0.964433
GLNO	Col	A	GL	3	B	16.4199	0.989710	0.973888	15.4512	0.992517	0.972306
GLNO	Col	A	GL	6	B	15.4556	0.990997	0.979231	14.6331	0.996185	0.979759
GLNO	Col	A	GL	9	B	14.4420	0.995734	0.987055	13.8311	0.998248	0.988304
GLNO	Col	A	GL	12	B	13.5121	0.997901	0.991364	13.1153	0.999281	0.991486
GLNO	Col	A	GL	Col	B	12.6732	0.999525	0.991329	12.4595	0.999859	0.991650
GLNO	Col	A	GL	3	C	16.4168	0.989577	0.974755	15.4274	0.991470	0.972758
GLNO	Col	A	GL	6	C	15.4268	0.994026	0.983153	14.5994	0.996322	0.982875
GLNO	Col	A	GL	9	C	14.2145	0.992871	0.986444	13.5867	0.997739	0.989453
GLNO	Col	A	GL	12	C	13.5323	0.998265	0.990192	13.1165	0.999372	0.991722
GLNO	Col	A	GL	Col	C	12.9495	0.999626	0.993163	12.7412	0.999839	0.993009
GLNO	Col	A	GL	3	D	16.4156	0.985721	0.970197	15.4108	0.991668	0.973108
GLNO	Col	A	GL	6	D	15.3794	0.989037	0.964734	14.5636	0.996091	0.969848
GLNO	Col	A	GL	9	D	14.4525	0.993541	0.970963	13.8412	0.996895	0.973910
GLNO	Col	A	GL	12	D	13.4605	0.997462	0.967861	13.0545	0.998197	0.970638

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GLNO	Col	A	GL	Col	D	12.8487	0.998698	0.968602	12.6294	0.998720	0.968459
GLNO	Col	A	GLNO	3	B	16.5432	0.988212	0.969243	15.5505	0.993817	0.971641
GLNO	Col	A	GLNO	6	B	15.6598	0.993530	0.971662	14.8178	0.995774	0.973613
GLNO	Col	A	GLNO	9	B	14.6648	0.994899	0.965526	13.9724	0.997036	0.966390
GLNO	Col	A	GLNO	12	B	13.9147	0.997533	0.966400	13.4226	0.998325	0.967331
GLNO	Col	A	GLNO	Col	B	13.4498	0.998541	0.965495	13.1009	0.998593	0.964687
GLNO	3	B	GL	6	250kB	15.6477	0.996975	0.990089	14.7389	0.998769	0.992939
GLNO	3	B	GL	9	250kB	15.9030	0.995000	0.976052	14.9683	0.997852	0.983779
GLNO	3	B	GL	3	A	14.4364	0.998263	0.991692	13.6533	0.999472	0.993415
GLNO	3	B	GL	6	A	15.1911	0.996365	0.985217	14.3303	0.999112	0.990591
GLNO	3	B	GL	9	A	15.6247	0.996705	0.992026	14.6959	0.998952	0.995285
GLNO	3	B	GL	12	A	15.9663	0.993361	0.975676	15.0145	0.997521	0.984652
GLNO	3	B	GL	Col	A	16.3939	0.975978	0.884621	15.4317	0.990840	0.960387
GLNO	3	B	GL	3	B	14.5250	0.998333	0.990973	13.7748	0.999590	0.992100
GLNO	3	B	GL	6	B	15.1402	0.997738	0.992027	14.2402	0.999333	0.993030
GLNO	3	B	GL	9	B	15.6049	0.996587	0.986387	14.6553	0.998524	0.989526
GLNO	3	B	GL	12	B	16.0113	0.991794	0.965720	15.0486	0.996723	0.977243
GLNO	3	B	GL	Col	B	16.4026	0.982827	0.962674	15.4250	0.987136	0.971873
GLNO	3	B	GL	3	C	14.6058	0.998548	0.991060	13.8409	0.999604	0.992949
GLNO	3	B	GL	6	C	15.1530	0.997784	0.990449	14.2377	0.999460	0.992209
GLNO	3	B	GL	9	C	15.6107	0.994518	0.981971	14.6632	0.998609	0.989454
GLNO	3	B	GL	12	C	16.0340	0.992998	0.968619	15.0833	0.996430	0.978694
GLNO	3	B	GL	Col	C	16.4185	0.982021	0.963280	15.4415	0.987704	0.972291
GLNO	3	B	GL	3	D	14.4050	0.998730	0.996181	13.6591	0.999491	0.996452
GLNO	3	B	GL	6	D	15.1386	0.997115	0.989902	14.2287	0.999370	0.992843
GLNO	3	B	GL	9	D	15.5969	0.996356	0.988647	14.6485	0.998957	0.992573
GLNO	3	B	GL	12	D	16.0034	0.990974	0.981158	15.0524	0.997288	0.985202
GLNO	3	B	GL	Col	D	16.4300	0.978998	0.925540	15.4435	0.990386	0.951974
GLNO	3	B	GLNO	6	B	14.4160	0.998600	0.995390	13.7341	0.999328	0.995859

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GLNO	3	B	GLNO	9	B	15.2996	0.997029	0.990765	14.4279	0.999107	0.993526
GLNO	3	B	GLNO	12	B	15.7120	0.990245	0.971776	14.7759	0.997330	0.988040
GLNO	3	B	GLNO	Col	B	16.2589	0.977303	0.843553	15.2928	0.990718	0.915179
GLNO	6	B	GL	6	250kB	14.7505	0.998005	0.989108	14.0082	0.999049	0.989791
GLNO	6	B	GL	9	250kB	14.8332	0.997963	0.981577	14.1188	0.998611	0.983592
GLNO	6	B	GL	3	A	15.1377	0.997449	0.986957	14.3001	0.998934	0.989350
GLNO	6	B	GL	6	A	13.7401	0.999219	0.989953	13.2750	0.999646	0.990914
GLNO	6	B	GL	9	A	14.3039	0.998972	0.995500	13.6715	0.999532	0.996174
GLNO	6	B	GL	12	A	14.8284	0.997110	0.984930	14.0899	0.998352	0.988664
GLNO	6	B	GL	Col	A	15.4124	0.988560	0.958083	14.5964	0.994614	0.979178
GLNO	6	B	GL	3	B	15.1429	0.996206	0.984082	14.2547	0.998663	0.987732
GLNO	6	B	GL	6	B	13.7645	0.999230	0.990094	13.3075	0.999739	0.990506
GLNO	6	B	GL	9	B	14.1176	0.998043	0.987991	13.4671	0.999392	0.990552
GLNO	6	B	GL	12	B	14.8805	0.996845	0.976070	14.1343	0.998079	0.977868
GLNO	6	B	GL	Col	B	15.5092	0.986771	0.975312	14.6821	0.991136	0.983485
GLNO	6	B	GL	3	C	15.1522	0.996191	0.982175	14.2624	0.998334	0.984285
GLNO	6	B	GL	6	C	13.4570	0.998516	0.990871	12.9563	0.999715	0.992916
GLNO	6	B	GL	9	C	14.2021	0.998600	0.988106	13.5899	0.999241	0.990360
GLNO	6	B	GL	12	C	15.0000	0.995780	0.976823	14.2450	0.997512	0.981645
GLNO	6	B	GL	Col	C	15.4635	0.986028	0.971380	14.6349	0.989594	0.979754
GLNO	6	B	GL	3	D	15.1139	0.996866	0.992049	14.2434	0.998630	0.993459
GLNO	6	B	GL	6	D	13.5059	0.999246	0.995893	13.0064	0.999832	0.996337
GLNO	6	B	GL	9	D	14.1865	0.998962	0.994761	13.5775	0.999593	0.995483
GLNO	6	B	GL	12	D	14.8583	0.996507	0.988956	14.1006	0.997995	0.991169
GLNO	6	B	GL	Col	D	15.4940	0.991389	0.970776	14.6731	0.991862	0.975237
GLNO	6	B	GLNO	9	B	13.4489	0.999241	0.994768	13.0026	0.999556	0.995141
GLNO	6	B	GLNO	12	B	14.5241	0.996984	0.985117	13.8141	0.997803	0.987450
GLNO	6	B	GLNO	Col	B	15.2953	0.990929	0.949693	14.5049	0.993336	0.947468
GLNO	9	B	GL	6	250kB	14.1283	0.996615	0.981750	13.5392	0.997823	0.982127

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GLNO	9	B	GL	9	250kB	13.6689	0.998216	0.979851	13.2120	0.999021	0.981998
GLNO	9	B	GL	3	A	15.5511	0.994152	0.975087	14.6272	0.995606	0.976723
GLNO	9	B	GL	6	A	14.1333	0.997616	0.979776	13.5290	0.997829	0.978768
GLNO	9	B	GL	9	A	13.2852	0.999476	0.993104	12.9506	0.999811	0.993152
GLNO	9	B	GL	12	A	13.5710	0.998041	0.992610	13.0911	0.999564	0.994692
GLNO	9	B	GL	Col	A	14.3278	0.994618	0.983737	13.6890	0.997563	0.990483
GLNO	9	B	GL	3	B	15.5794	0.995355	0.980835	14.6372	0.995594	0.978409
GLNO	9	B	GL	6	B	14.2679	0.997791	0.985468	13.6278	0.997824	0.984058
GLNO	9	B	GL	9	B	13.3636	0.999070	0.985173	13.0129	0.999581	0.985973
GLNO	9	B	GL	12	B	13.5955	0.998287	0.975910	13.1370	0.998926	0.976259
GLNO	9	B	GL	Col	B	14.2813	0.995230	0.983705	13.6783	0.997230	0.985265
GLNO	9	B	GL	3	C	15.6212	0.992759	0.969844	14.6608	0.995445	0.971035
GLNO	9	B	GL	6	C	14.2784	0.997566	0.985007	13.6679	0.998289	0.983544
GLNO	9	B	GL	9	C	13.1861	0.999108	0.985150	12.8095	0.999575	0.986603
GLNO	9	B	GL	12	C	13.5959	0.997655	0.979147	13.1170	0.998995	0.982444
GLNO	9	B	GL	Col	C	14.3457	0.993192	0.982493	13.7233	0.997292	0.986926
GLNO	9	B	GL	3	D	15.5375	0.993736	0.986820	14.5802	0.996213	0.987502
GLNO	9	B	GL	6	D	14.2360	0.998724	0.993219	13.6113	0.999045	0.992439
GLNO	9	B	GL	9	D	13.2144	0.999669	0.993115	12.8570	0.999889	0.993340
GLNO	9	B	GL	12	D	13.7446	0.998465	0.994214	13.2721	0.999338	0.994784
GLNO	9	B	GL	Col	D	14.2840	0.994071	0.981061	13.6774	0.996833	0.988450
GLNO	9	B	GLNO	12	B	13.0933	0.998943	0.994264	12.7942	0.998761	0.993231
GLNO	9	B	GLNO	Col	B	14.0871	0.993509	0.974453	13.5556	0.997865	0.984618
GLNO	12	B	GL	6	250kB	14.2541	0.991955	0.974189	13.6114	0.994204	0.971560
GLNO	12	B	GL	9	250kB	13.6234	0.997774	0.971114	13.2140	0.997318	0.968686
GLNO	12	B	GL	3	A	16.0250	0.984930	0.956982	15.0641	0.987969	0.960480
GLNO	12	B	GL	6	A	14.8622	0.991963	0.963624	14.1197	0.993148	0.962098
GLNO	12	B	GL	9	A	13.6687	0.998271	0.992230	13.1965	0.998841	0.991829
GLNO	12	B	GL	12	A	13.0355	0.999524	0.995401	12.7201	0.999883	0.995774

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GLNO	12	B	GL	Col	A	13.5768	0.998872	0.995163	13.1794	0.999513	0.995740
GLNO	12	B	GL	3	B	16.0464	0.986149	0.961738	15.1200	0.986303	0.958863
GLNO	12	B	GL	6	B	14.9204	0.993662	0.975231	14.1740	0.993643	0.971852
GLNO	12	B	GL	9	B	13.7610	0.997143	0.975412	13.2724	0.997725	0.976665
GLNO	12	B	GL	12	B	13.1102	0.998911	0.975741	12.8125	0.999262	0.976868
GLNO	12	B	GL	Col	B	13.3722	0.998376	0.979033	12.9674	0.998834	0.979133
GLNO	12	B	GL	3	C	16.0388	0.985678	0.955099	15.0789	0.985910	0.946760
GLNO	12	B	GL	6	C	14.8458	0.994327	0.973127	14.0855	0.992853	0.966040
GLNO	12	B	GL	9	C	13.5763	0.997203	0.977677	13.1377	0.997444	0.976780
GLNO	12	B	GL	12	C	13.1995	0.998807	0.976491	12.9199	0.999218	0.977571
GLNO	12	B	GL	Col	C	13.5076	0.998008	0.978406	13.1051	0.998856	0.979096
GLNO	12	B	GL	3	D	15.9930	0.984235	0.974615	15.0141	0.987743	0.972383
GLNO	12	B	GL	6	D	14.8849	0.994507	0.983743	14.1390	0.995325	0.982615
GLNO	12	B	GL	9	D	13.7956	0.998352	0.991320	13.3032	0.998696	0.991178
GLNO	12	B	GL	12	D	13.0057	0.999641	0.994403	12.7217	0.999909	0.994527
GLNO	12	B	GL	Col	D	13.4390	0.999154	0.992516	13.0277	0.999594	0.993497
GLNO	12	B	GLNO	Col	B	13.0134	0.999376	0.993583	12.7479	0.999559	0.993389
GLNO	Col	B	GL	6	250kB	14.8086	0.982501	0.949528	14.0929	0.983606	0.947300
GLNO	Col	B	GL	9	250kB	14.1652	0.992391	0.944330	13.6279	0.991242	0.941375
GLNO	Col	B	GL	3	A	16.3787	0.968934	0.925215	15.3912	0.971734	0.931375
GLNO	Col	B	GL	6	A	15.3949	0.976450	0.925974	14.5850	0.981062	0.929759
GLNO	Col	B	GL	9	A	14.3200	0.993975	0.985426	13.7115	0.993936	0.984374
GLNO	Col	B	GL	12	A	13.5373	0.998240	0.992899	13.1246	0.998459	0.993363
GLNO	Col	B	GL	Col	A	13.0421	0.998917	0.995128	12.8146	0.999922	0.996160
GLNO	Col	B	GL	3	B	16.3916	0.971701	0.932659	15.4407	0.966408	0.929618
GLNO	Col	B	GL	6	B	15.3683	0.980002	0.956616	14.5405	0.982609	0.952733
GLNO	Col	B	GL	9	B	14.4148	0.992025	0.959756	13.7855	0.991838	0.959119
GLNO	Col	B	GL	12	B	13.5462	0.995991	0.955314	13.1606	0.995715	0.952800
GLNO	Col	B	GL	Col	B	12.9509	0.998403	0.966606	12.7383	0.998617	0.966499

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
GLNO	Col	B	GL	3	C	16.3801	0.964280	0.916046	15.4269	0.962579	0.900599
GLNO	Col	B	GL	6	C	15.4338	0.983560	0.955756	14.6202	0.979379	0.944704
GLNO	Col	B	GL	9	C	14.3323	0.991285	0.960468	13.7210	0.991135	0.955266
GLNO	Col	B	GL	12	C	13.4258	0.996216	0.962249	13.0356	0.995719	0.960579
GLNO	Col	B	GL	Col	C	12.8592	0.998151	0.965812	12.6428	0.998586	0.966811
GLNO	Col	B	GL	3	D	16.4202	0.942597	0.940644	15.4115	0.965553	0.941263
GLNO	Col	B	GL	6	D	15.4170	0.978243	0.965415	14.5689	0.985052	0.968694
GLNO	Col	B	GL	9	D	14.2375	0.992865	0.985106	13.6092	0.993457	0.983913
GLNO	Col	B	GL	12	D	13.4782	0.998206	0.991788	13.0883	0.998337	0.991406
GLNO	Col	B	GL	Col	D	12.8470	0.999699	0.991779	12.6347	0.999891	0.992047
FixV	3	A	FixV	6	A	15.9573	0.994568	0.988792	N/A	N/A	N/A
FixV	3	A	FixV	12	A	16.7767	0.989260	0.982618	N/A	N/A	N/A
FixV	3	A	FixV	6	B	15.9519	0.992927	0.981460	N/A	N/A	N/A
FixV	3	A	FixV	12	B	16.7669	0.984439	0.981575	N/A	N/A	N/A
FixV	6	A	FixV	12	A	13.9117	0.997971	0.993954	N/A	N/A	N/A
FixV	6	A	FixV	6	B	13.1995	0.999369	0.997597	N/A	N/A	N/A
FixV	6	A	FixV	12	B	14.4291	0.997236	0.992603	N/A	N/A	N/A
FixV	12	A	FixV	12	B	13.0249	0.999629	0.997035	N/A	N/A	N/A
FixV	6	B	FixV	12	A	14.4067	0.997736	0.994143	N/A	N/A	N/A
FixV	6	B	FixV	12	B	13.8834	0.997622	0.993148	N/A	N/A	N/A
RTRC	GCat	1k	GL	3	A	17.7923	0.001970	0.207000	17.4166	0.459534	0.585353
RTRC	GCat	1k	GL	6	A	18.3745	0.593325	0.675849	17.9919	0.608744	0.700515
RTRC	GCat	1k	GL	9	A	18.9009	0.598590	0.787523	18.5255	0.367305	0.702310
RTRC	GCat	1k	GL	12	A	19.2099	0.196921	0.497905	18.8249	0.322876	0.589140
RTRC	GCat	1k	GL	Col	A	19.5199	0.218587	0.551068	19.0928	0.034436	0.303836
RTRC	GCat	1k	GL	3	B	17.7874	0.365823	0.486795	17.4094	0.555814	0.722989
RTRC	GCat	1k	GL	6	B	18.3759	0.608059	0.758579	17.9520	0.629720	0.814611
RTRC	GCat	1k	GL	9	B	18.9215	0.438454	0.596784	18.5248	0.564673	0.722304
RTRC	GCat	1k	GL	12	B	19.2422	0.435718	0.635893	18.8460	0.258562	0.531902

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
RTRC	GCat	1k	GL	Col	B	19.5041	0.336115	0.600942	19.0812	0.217867	0.585493
RTRC	GCat	1k	GL	3	C	17.7960	0.007239	0.303833	17.4216	0.528445	0.705470
RTRC	GCat	1k	GL	6	C	18.3927	0.029330	0.311750	17.9812	0.598825	0.771328
RTRC	GCat	1k	GL	9	C	18.9374	0.301564	0.520302	18.5194	0.602344	0.750360
RTRC	GCat	1k	GL	12	C	19.2830	0.480861	0.638949	18.8726	0.205925	0.567888
RTRC	GCat	1k	GL	Col	C	19.4671	0.106179	0.359290	19.0560	0.285393	0.721695
RTRC	GCat	1k	GL	3	D	17.7905	-0.000266	0.219128	17.4505	0.580529	0.764302
RTRC	GCat	1k	GL	6	D	18.3398	0.397395	0.564172	17.9671	0.581987	0.797480
RTRC	GCat	1k	GL	9	D	18.9405	0.593063	0.783905	18.5788	0.582476	0.778116
RTRC	GCat	1k	GL	12	D	19.2844	0.440872	0.675955	18.9001	0.357081	0.652829
RTRC	GCat	1k	GL	Col	D	19.5404	0.080866	0.420164	19.1223	0.020076	0.395871
RTRC	GCat	1k	RTRC	GCat150k	B	19.6901	0.771990	0.862867	19.0174	0.856335	0.929790
RTRC	GCat	1k	RTRC	GCat		19.7782	0.740204	0.728679	19.1197	0.852692	0.896005
RTRC	GCat	1k	RTRC	GCat150k		19.7246	0.702293	0.852331	19.0544	0.800315	0.915971
RTRC	GCat150k		GL	3	A	18.2960	0.804889	0.865202	17.9654	0.753782	0.824922
RTRC	GCat150k		GL	6	A	18.6820	0.227745	0.663044	18.3454	0.782034	0.796104
RTRC	GCat150k		GL	9	A	19.0578	0.795388	0.870461	18.7453	0.615526	0.785710
RTRC	GCat150k		GL	12	A	19.3428	0.819251	0.878901	19.0069	0.432634	0.664481
RTRC	GCat150k		GL	Col	A	19.4761	0.867355	0.907166	19.1137	0.649582	0.813927
RTRC	GCat150k		GL	3	B	18.2882	0.704325	0.854902	17.9720	0.333095	0.655845
RTRC	GCat150k		GL	6	B	18.6652	0.495320	0.822659	18.2892	0.529538	0.659899
RTRC	GCat150k		GL	9	B	19.0682	0.719055	0.829818	18.7143	0.649865	0.737378
RTRC	GCat150k		GL	12	B	19.3302	0.813159	0.851318	18.9921	0.634419	0.731966
RTRC	GCat150k		GL	Col	B	19.5388	0.922218	0.932210	19.1504	0.580807	0.768174
RTRC	GCat150k		GL	3	C	18.2672	0.629041	0.781824	17.9367	0.451314	0.653610
RTRC	GCat150k		GL	6	C	18.7331	0.437945	0.743245	18.3374	0.696602	0.833691
RTRC	GCat150k		GL	9	C	19.0660	0.747969	0.825464	18.7145	0.714219	0.747563
RTRC	GCat150k		GL	12	C	19.3859	0.777122	0.793290	18.9906	0.478200	0.681029
RTRC	GCat150k		GL	Col	C	19.4833	0.765623	0.864091	19.1358	0.641762	0.774929

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
RTRC	GCat150k		GL	3	D	18.2799	0.633571	0.799676	17.9301	0.593934	0.653837
RTRC	GCat150k		GL	6	D	18.6782	0.587590	0.821502	18.3389	0.688793	0.725307
RTRC	GCat150k		GL	9	D	19.0167	0.768728	0.842289	18.7156	0.539463	0.667261
RTRC	GCat150k		GL	12	D	19.3151	0.715542	0.789171	18.9820	0.565975	0.710178
RTRC	GCat150k		GL	Col	D	19.4727	0.830623	0.869745	19.1342	0.598719	0.782225
RTRC	GCat150k	B	GL	3	A	18.3235	0.589229	0.686260	18.0520	0.471040	0.669409
RTRC	GCat150k	B	GL	6	A	18.7459	0.708096	0.815489	18.3984	0.768055	0.799589
RTRC	GCat150k	B	GL	9	A	19.1104	0.721502	0.789938	18.8008	0.856421	0.900333
RTRC	GCat150k	B	GL	12	A	19.3987	0.864200	0.901154	19.0491	0.788261	0.836346
RTRC	GCat150k	B	GL	Col	A	19.5439	0.782975	0.869371	19.1888	0.382074	0.567877
RTRC	GCat150k	B	GL	3	B	18.3622	0.829409	0.865957	18.0405	-0.013146	0.426777
RTRC	GCat150k	B	GL	6	B	18.7356	0.850727	0.930601	18.3813	0.322941	0.445513
RTRC	GCat150k	B	GL	9	B	19.0890	0.892645	0.890352	18.7636	0.781267	0.822856
RTRC	GCat150k	B	GL	12	B	19.4128	0.869458	0.855281	19.0467	0.823263	0.835727
RTRC	GCat150k	B	GL	Col	B	19.5361	0.877456	0.882581	19.1955	0.895964	0.910838
RTRC	GCat150k	B	GL	3	C	18.3351	0.704663	0.801669	18.0085	0.560161	0.689228
RTRC	GCat150k	B	GL	6	C	18.7572	0.705056	0.832770	18.4247	0.247785	0.626471
RTRC	GCat150k	B	GL	9	C	19.1029	0.796615	0.843873	18.7387	0.777767	0.824255
RTRC	GCat150k	B	GL	12	C	19.4026	0.835790	0.850977	19.0436	0.779346	0.784196
RTRC	GCat150k	B	GL	Col	C	19.5367	0.551229	0.741756	19.2066	0.836161	0.880135
RTRC	GCat150k	B	GL	3	D	18.3199	0.714201	0.833818	17.9928	0.605754	0.705772
RTRC	GCat150k	B	GL	6	D	18.7296	0.782760	0.845991	18.3802	0.684498	0.747774
RTRC	GCat150k	B	GL	9	D	19.1370	0.915144	0.902905	18.7877	0.737036	0.778766
RTRC	GCat150k	B	GL	12	D	19.3638	0.751128	0.810527	19.0160	0.900040	0.922723
RTRC	GCat150k	B	GL	Col	D	19.5504	0.548902	0.722004	19.1922	0.858733	0.877239
RTRC	GCat		GL	3	A	18.3395	0.501347	0.662082	18.0185	0.489157	0.635362
RTRC	GCat		GL	6	A	18.7781	0.671410	0.840503	18.4175	0.683338	0.744005
RTRC	GCat		GL	9	A	19.1962	0.631085	0.773505	18.8361	0.731997	0.846672
RTRC	GCat		GL	12	A	19.3956	0.854858	0.897210	19.0618	0.543503	0.664382

Table A – 5 Continued

Group 1	Set 1	Series 1	Group 2	Set 2	Series 2	T-500 OTV-Norm	T-500 Doc Cent	T-500 Term Cent	T-5000 OTV-Norm	T-5000 Doc Cent	T-5000 Term Cent
RTRC	GCat		GL	Col	A	19.5829	0.817592	0.904360	19.2084	0.262108	0.607550
RTRC	GCat		GL	3	B	18.3321	0.750002	0.832112	18.0110	0.055601	0.515326
RTRC	GCat		GL	6	B	18.7194	0.770137	0.911551	18.3616	0.348752	0.470232
RTRC	GCat		GL	9	B	19.1218	0.892039	0.913650	18.7398	0.427623	0.620624
RTRC	GCat		GL	12	B	19.4179	0.889590	0.888828	19.0836	0.700374	0.784630
RTRC	GCat		GL	Col	B	19.5253	0.843179	0.914780	19.2194	0.745884	0.846572
RTRC	GCat		GL	3	C	18.3676	0.282378	0.607673	18.0187	0.365573	0.549749
RTRC	GCat		GL	6	C	18.7768	0.670435	0.815474	18.4287	0.240289	0.600926
RTRC	GCat		GL	9	C	19.1558	0.760565	0.826587	18.8017	0.635166	0.765753
RTRC	GCat		GL	12	C	19.3762	0.807430	0.839079	19.0192	0.703040	0.759252
RTRC	GCat		GL	Col	C	19.5664	0.815969	0.861761	19.1726	0.620503	0.804209
RTRC	GCat		GL	3	D	18.2980	0.524464	0.741682	18.0256	0.575540	0.742742
RTRC	GCat		GL	6	D	18.7333	0.652138	0.814389	18.4090	0.634960	0.771911
RTRC	GCat		GL	9	D	19.1565	0.782998	0.829375	18.8071	0.311245	0.561562
RTRC	GCat		GL	12	D	19.4024	0.659325	0.841116	19.0466	0.390103	0.685346
RTRC	GCat		GL	Col	D	19.5510	0.709884	0.807961	19.2233	0.395317	0.683025

VITA

John Martin received his Bachelor of Science degree in Computer Science from Milligan College in 1990, and after working in industry for a while pursued graduate studies at the University of Tennessee where, in 1997, he completed a Master of Science degree in Computer Science. John has been a principal owner of technology firms since 1998 and is a founder and CTO/CFO at Small Bear Technologies, Inc. a firm that specializes in large scale Latent Semantic Analysis. He returned to the University of Tennessee in 2011 to continue pursuit of a PhD in Computer Science with research interests in large scale text mining specifically related to Latent Semantic Analysis.

John has over twenty-five years of practical software design and development experience in various environments covering numerous software development and testing/certification projects spanning several different fields of application. This work has included business systems, weapons systems, air traffic control, and medical devices as well as involvement in the research and application of formal methods software engineering practices. He has been responsible for the development and support of large-scale on-line transaction processing systems, cross-platform system interfaces, embedded software systems, software system architectures, and application development standards and policy.