Journal of MultiDisciplinary Evaluation
Volume 8, Issue 19

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Using Stylometric Techniques to Evaluate New Testament Authorship

Kenneth D. Royal
University of Kentucky

Background: Bible scholars often debate the authorship of certain books appearing in the New Testament.

Purpose: The purpose of the present study was to evaluate New Testament authorship by using stylometric analytical techniques

Setting: This research focuses on texts appearing in the New Testament.

Intervention: This was an exploratory research on evaluation study with no intervention.

Research Design: A powerful, state-of-the-art psychometric model was applied to Biblical text in an effort to identify correlations among word usage and writing style among each of the New Testament books.

Data Collection and Analysis: Strong's Concordance was used to provide original Greek text. Computer programming was necessary to create a worksheet that contained a list of New Testament books, each Greek word appearing in the New Testament, and a count of each word's appearance relative to each book. Rasch-based Principal Components Analysis of standardized residual correlations was used to map stylistic similarities and differences.

Findings: With regard to substantive findings, the gospels (Matthew, Luke, Mark, and John) and the narrative book of Acts were closely correlated. Other texts presented a mix of expected and unexpected findings. With regard to other findings, the technique presented in this study offers a great deal of promise to various research and evaluation practices.

Keywords: *authorship; Bible; measurement; New Testament; psychometrics; stylometrics*

Stylometrics is the study of linguistic style. Typically, stylometric techniques are used to discern authorship of anonymous or disputed texts. Of course, the term "metric" in stylometrics implies the measurement of style. Similar to how measurement in various social and behavioral sciences has been rather slow to adopt sound objective measurement practices, the study of linguistic style is certainly no different. Despite the many advantages of computers and modern technology that now make stylometric studies much more feasible and less time-consuming than ever before, many scholars are still utilizing quantitative methods for data analysis that limit the inferential value of their research findings. The purpose of the present study was to (a) provide a demonstration of how objective measurement can be applied to stylometric studies, (b) provide substantive findings toward the highly disputed and long-standing debate of Biblical authorship, and (c) introduce this methodology to the mainstream research methods and evaluation literature as a viable approach for some evaluation studies.

Background and Overview

When attempting a stylometric analysis of the Bible, a number of important issues need to be considered. Some studies have attempted to use all words appearing in the Bible as the basis for such an analysis. Others have opted to use only some of the more substantive words. Some researchers have opted to use words from the original texts, whereas others have used translated versions. Readers are encouraged to see Barr (2003), Kenny (1986), Linacre (2001), and Whissell (2006) for some notable examples of related studies. Naturally, there are advantages and disadvantages to each of these approaches. Regardless of the aforementioned decision processes used by various researchers, it is important that researchers attempt to analyze data with state-of-the-art techniques, especially those that can be rather easily reproduced and results that can be cleanly compared.

Psychometrics, the study of mental measurement, commonly uses techniques from the item response theory perspective to analyze quantitative data. One form of item response theory, namely the Rasch family of models, is considered by many to be the "gold standard" for measurement analyses (Bond and Fox, 2007). This is largely because Rasch models are invariant and overcome erroneous assumptions typical of traditional statistical methods, such as treating raw scores as measures, ordinal data as interval, assuming all items are of equal difficulty, etc. Typically, Rasch analyses involve a comparison of people versus items. In this stylometric study, Rasch analyses will instead involve a comparison of Biblical texts versus the frequency of word choice.

One technique in particular will be used to evaluate the findings of this study. That is, the Rasch-based Principal Components Analysis (PCA) of standardized residual correlations. Rasch-based PCAs are routinely performed to investigate the dimensionality of a dataset. Typically, one hopes to explain as much variance as possible, while at the same time detecting a significant amount of the primary latent trait. Additional dimensions are evaluated by both the size of their Eigenvalues and the ratio for which they make up the explained variance, especially the item variance explained. However, this useful technique can be used for other purposes as well. For instance, the correlations produced from the analysis can be easily mapped for visual interpretation of the relationships present within the data structure. This results in essentially a single snapshot of data that even those unfamiliar with the technique (or Rasch measurement in general) can easily and accurately interpret.

The Present Study

In 2001, Rasch measurement pioneer Mike Linacre introduced a novel way to use objective measurement in stylometric analyses. While Linacre provided a very valuable contribution to the literature, much of the details of the analysis were not explained due to the space restrictions of the publication venue. As such, it is important that his work is followed-up with a bit more detail so that others can more easily replicate his ingenious methods. While the present study will attempt to essentially answer the same question of Linacre's study (who wrote various texts of the New Testament?), the decision processes leading up to the production of the final data file will differ in a number of important ways. The author does not contend the methods used in the present study are improved over those of Linacre's, but the author simply uses this as an opportunity to again investigate the topic of New Testament authorship with a different decision process, but same objective measurement analytics. Thus, the results of the two studies can be easily compared to determine the extent to which results appear stable, even when author decision processes differ when constructing the dataset.

Methods

Overview of Data and Design

The New Testament was written in Greek and consists of 27 books. Naturally, for those that do not speak Greek this poses a significant problem with regard to data interpretation and translation. Fortunately, James Strong developed the "Strong's Concordance of the Bible" that essentially provides an exhaustive cross-reference of every word in the King James Version (KJV) of the Bible with the original word used in the Greek New Testament. Strong's Concordance also includes an index of Hebrew words for cross-referencing the Old Testament. Strong's Concordance was first published in 1890, but is publically (and freely) available today on many websites. In fact, many websites have versions of Strong's Concordance that are linked to the KJV of the New Testament. Computer programmers can readily produce code to create a dataset much like the one used in this study.

With regard to the dataset, each of the 27 books was treated as column variables and each word appearing in Strong's Concordance constituted a row of data. Cells were populated with the frequency that each word was used in each respective book. Strong's Concordance consists of 5,624 different Greek words, each assigned a reference number. For the present study, each book was treated as a "person," and each Greek word was treated as an "item."

Difference in Decision-Tree Process Approach

As mentioned previously, Linacre's study evaluating New Testament authorship provided one roadmap for a stylometric study. The present study veers a slightly different course, thus this study differs from Linacre's in at least three key ways. First, in Linacre's study, words that appeared more than nine times in any book were truncated to a frequency of nine. However, when investigating the counts of word frequencies used in each book, some words appear at a considerably higher rate. In fact, approximately 5% of the words used in most books in the New Testaments appeared more than nine times. The author of the present study believes it is important to take advantage of that information and maximize its inferential value. Therefore, the present study deviated from Linacre's in that additional categories were used (n = 20) in hopes of increasing precision to the measures. The schema used in this analysis involved recoding various ranges into new categories. For instance, words that appeared 11-19 times were recoded to "11;" words that appeared 20-29 times were recoded to "12;" this process was repeated until words appearing 100 times or more were recoded into "20."

Second, it appears some non-substantive words may have been deleted from Linacre's dataset. While on one hand removing non-substantive words (such as articles, conjunctions, etc.) could potentially reduce noise from the measurement system, the author of the present study elected to keep all words appearing in the dataset. The rationale for this decision is due to the argument that one could perceive all words as being valuable. For instance, it may be the case that an author used some non-substantive word excessively, thus essentially being a trademark of sorts for his vernacular.

Third, Linacre utilized a rather novel method in his analysis in which he evaluated the correlation between word usage and book length. Of course, some books are rather small and others quite large, therefore if particular words are used a great deal in a very short book it could reveal quite a bit of information about authorship. Linacre essentially identified all the books that were negatively correlated in this manner and recoded the values via several iterations until no negative correlations remained. In Linacre's study, a total of 735 words were reversed. For the present study, the researcher elected not to evaluate the correlations between word usage and book length, as the directionality of the correlations was of less concern than their magnitudes. As such, the present study left word frequencies unaltered.

Data Analysis

The Rasch Rating Scale Model (Andrich, 1978) was selected as the measurement model, as the RRSM is well-suited for polytomous data that contain the same number of possible response options. According to the RRSM model, the probability of a person n responding in category x to item i, is given by:

$$P_{xni} = \frac{\exp \sum_{j=0}^{x} [\beta_n - (\delta_i + \tau_j)]}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} [\beta_n - (\delta_i + \tau_j)]} \qquad x = 0,1,...,m$$

where $\tau_0 = 0$ so that $\exp \sum_{j=0}^{0} [\beta_n - (\delta_i + \tau_j)] = 1$ and where β_n is the person's position on the variable, δ_i is the scale value (difficulty to endorse)

estimated for each item i and $\tau_1, \tau_2, \ldots, \tau_m$ are the m response thresholds estimated for the m + 1 rating categories. More specifically, a Principal Components Analysis (PCA) of standardized residual correlations (Linacre, 2011) was employed as the primary analytic technique to evaluate Biblical authorship. Winsteps measurement software (Linacre, 2010) was used to perform the data analysis.

As noted previously, the Rasch-based PCA can be particularly powerful and useful for stylometric studies, as the technique not only identifies similarities and contrasts in the data structure, but its results also can be presented in a single graphic. Rasch-based PCA techniques extract the primary (Rasch) dimension from the data prior to the first contrast. The observations that are not explained by the primary dimension are the residuals. According to Rasch measurement theory, when attempting to establish a unidimensional measurement system any residuals should be random in nature. If residuals are not random, it could indicate the presence of multidimensionality. In the present study, the presence of multidimensionality is somewhat irrelevant, as we are not attempting to construct a unidimensional measurement system, do not have the freedom to remove misfitting persons and items to accord to the desirable unidimensional structure, nor are we attempting to produce a linear continuum to discern the difficulty of various items. Instead, we are simply looking for similarities and differences based upon what the messy data give us. In some ways, this use of the Rasch model is contrary to its philosophy and typical application. It should be noted that multidimensionality may be more important in other stylometric studies, as this is largely governed by the decision analytic process of the researcher.

With traditional Rasch-based PCA interpretation, the books appearing at the polar ends of the vertical continuum would indicate those that contrast most sharply. In stylometric studies, the notion of multidimensionality may be somewhat muted. At best, the information can only indicate that the author's writing styles differed most dramatically. Such information can be very useful as it would quickly identify which texts were most unlikely to have been written by the same author. Previous research indicates that useful information can be obtained when using as few as 20 persons for a PCA of items, and 20 items for a PCA of persons (Arrindell & van der Ende, 1985), although many guidelines for Rasch-based PCA interpretation have yet to yield a consensus. The data employed in the present study exceed the minimum requirements for person and item volume, thus should yield useful findings.

Results and Discussion

Results of this study yield a number of interesting findings (see Figure 1 and Table 1). First, the six narrative books (Matthew, Mark, Luke, John, Acts and Revelation) are clustered relatively close together. Most Bible scholars are in agreement that Luke (or at the very least, the same author of Luke) also wrote Acts, so there is some stylometric evidence in support of this. Although Matthew, Mark and Luke are believed to have been penned by different authors, these books are no doubt closely correlated due to their similar accounts of the life of Jesus Christ. Many Bible scholars believe the book of John was written approximately 20 years after the other gospels, which could also speak to some of the noticeable differences between this text and the other gospels. Although there appears to be a number of stylometric similarities between the book of John (E) and Revelation (F), both written by someone named John, the book of John is generally believed to have been penned by the Apostle John, whereas Revelation is generally believed to have been penned by John of Patmos. Empirical stylometric evidence, however, might suggest the same author wrote both texts.

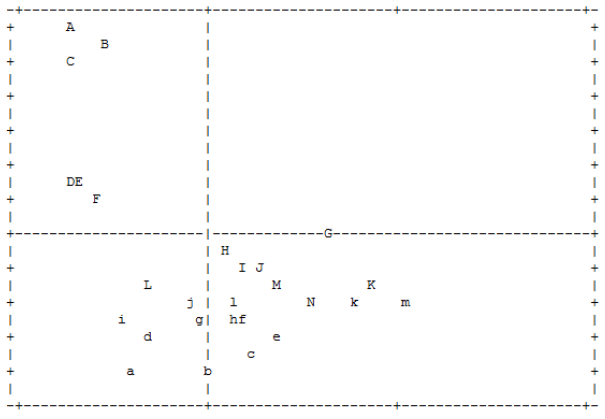

Figure 1. Residual Correlation Map of Books Appearing in the New Testament

Table 1
Legend for Interpreting Standardized Residual Correlation Map

Positive Loadings	Negative Loadings
A - Matthew	a - Romans
B - Mark	b - Ephesians
C - Luke	c - Colossians
D - Acts	d – 2 Corinthians
E - John	e – 2 Thessalonians
F - Revelation	f – 1 Thessalonians
G - Philemon	g - Galatians
	h - Philippians
	i – 1 Corinthians
	j – 1 John
	k - Jude
	l – 1 Peter
	m – 3 John
	N - Titus
	M – 2 Peter
	L - Hebrews
	K – 2 John
	J – 2 Timothy
	I – James
	H – 1 Timothy

Second, John the Evangelist is typically credited for 1 2 and 3 John (represented j, K, and m, respectively), yet only 2 and 3 John have strong empirical evidence to suggest they were written by the same author. 1 John appears some distance away from 2 and 3 John on the correlation map. James (represented by I) and Jude (represented by k) appear somewhat close to one another on the map, although most scholars are in agreement that each text was written by a different author. Both authors identified themselves as "a slave of Jesus Christ", but Jude introduced himself as "a brother of James" (Jas 1:1; Jude 1).

Third, the epistles that are generally thought to be written by Paul (i.e., Colossians, Philippians, Ephesians, Romans, Titus, 1 and 2 Thessalonians, 1 and 2 Corinthians, 1 and 2 Timothy, Galatians, and Philemon) are also clustered relatively close together. Titus (N) and 1 Corinthians (i) appear to serve as the lateral bookends for most of Paul's letters, with Romans (a) and Ephesians (b) and 1 Timothy (H) as the vertical bookends. Interestingly, 1 and 2 Peter also fall within the spectrum of Pauline authored letters. Although Paul is not considered a potential author of 1 and 2 Peter (l, M, respectively), this analysis reveals the choice of words and writing style used in 1 and 2 Peter may have been quite similar to many of Paul's letters.

Finally, Hebrews appears to be the text that is the most often disputed with regard to authorship. Many attribute this text to Paul, others attribute it to Luke, or perhaps Barnabus (Paul's assistant). Based on this stylometric analysis, Hebrews appears fairly closely related to 1 and 2 Corinthians, which provides some empirical evidence of Pauline authorship, but it should be noted that most other letters typically attributed to Paul appear even more distant from Hebrews. Collectively, this might provide mixed evidence of Pauline authorship. Interestingly, scholars who claim Luke may have written Hebrews also have some empirical evidence to support that assertion, as Hebrews location on the map is not particularly far from Acts (D). In fact, one could possibly make the argument that, empirically speaking, Hebrews is as closely correlated to Acts as Hebrews is to Titus (typically thought to be written by Paul). Of course, there is also a possibility that an entirely different author (such as Barnabas) penned this text. Unfortunately, this analysis failed to reveal a sufficiently strong correlation to make any real empirical claim of authorship for this particular text.

It should be noted that many, if not most, Christians believe all Biblical texts were inspired by the Holy Spirit (2 Tim. 3.16-17). That is, regardless of the human author used to write the text, all Biblical texts have a purpose, message, and special significance to readers. While the question of who authored Hebrews or any other text is a bit irrelevant for some, others may find comfort in knowing as it could potentially provide additional assurance of the text's validity. Clearly, however, Hebrews belongs in the New Testament as its message is in alignment with many of Paul's other writings.

Limitations and Future Research

It should be noted that this study possessed a number of very important limitations. Some limitations pertain to shortcomings of stylometric studies in general, and others specific to the present study. In any instance, future researchers may wish to do some things differently should they attempt to replicate this study or conduct a stylometric study of their own.

With regard to some of the more generic shortcomings of stylometric analyses, issues such as single word selection versus phrases are paramount. When attempting to investigate an author's writing style, it is helpful to look at not only the choice of individual words, but also complete phrases. An author's written work is much like a fingerprint. Various clues of the author's identity are everywhere. The present

study investigated only single word usage, thus providing the most basic stylometric analysis possible. More sophisticated methods would likely improve the accuracy of findings.

Word selection and content overlap pose additional problems. Some authors of the New Testament intentionally constructed their letters to speak directly to certain individuals, or persons residing in a particular community. The extent to which authors essentially deviated from their normal writing style to tailor letters to a particular audience could introduce some error into the measurement system. Further, the extent to which various authors speak about common topics can provide some distortion to the measures as well. For instance, although the gospels are closely correlated this does not suggest they were written by the same individual. Many passages in the gospels provide quotes from Jesus or other important historical persons. The extent to which these commonalities exist and introduce noise into the measurement system is not accounted for in the present study.

Additional problems arise when texts were written by more than one person. It is possible that some New Testament texts could share joint authorship. A researcher's ability to parse out this type of information would be quite limited without a great deal of additional contextual information. As such, the combination of styles used by two authors might produce an invalid measure of a presumed single author, thus distorting the findings.

Translation error is sure to be a problem as well. The present analysis relies entirely on Strong's Concordance, which relies in turn on the King James Version of the Bible. The KJV (and therefore Strong's) ignores the issue of textual variation in Biblical manuscripts. Since the KJV reflects a particular majority consensus, it is entirely possible that the authorship question is additionally confounded by the aggregate effect of an arbitrarily large number of copying errors and other small changes, as well as centuries of minor editorial decisions.

Future studies that investigate New Testament authorship may wish to do a number of things differently to minimize sources of error. One example might be to create a baseline measure of texts that were undoubtedly written by the Apostle Paul and compare those texts to the remaining books of the New Testament. The extent to which other texts correlate with known Pauline texts could provide a more accurate measure of authorship of text traditionally thought (or assumed) to be written by Paul. This approach could also more accurately reveal which texts were not written by Paul, and when combined with other historical information (e.g., presumed time stamp, physical location of discovery, etc.) could lead to potentially more accurate insights about New Testament authorship.

Future studies may also wish to study apocryphal texts in relation to other accepted texts. One particular apocryphal text, the Epistle to the Laodiceans, has been discerned by most Biblical scholars as a pseudepigraphical letter composed under the guise of the Apostle Paul that was presumed to be lost. Although this letter has been largely rejected by Biblical scholars, it would be interesting to see the extent to which its author who attempted to imitate the Apostle Paul was successful in a stylometric manner. Regardless of the topic or purpose of investigation, stylometric studies such as this one offer a unique, empirical perspective on evaluating literary authorship. The present study focused on Biblical authorship and it is the researcher's hope that in some small way the methodology presented here will assist Biblical scholars in their pursuit to better discern Biblical authorship. However, stylometric analyses should not be limited to studies of the Bible. There is also a great deal of potential for these techniques in other arenas, such as discerning famous literary works, documents of the deceased, and disputed legal documents, for example.

Implications for Evaluation Studies

The stylometric techniques demonstrated in this paper have a number of potential implications for evaluation. In fact, stylometrics would make a great complement to many forms of qualitative research. For example, stylometrics could be used to categorize interview transcripts into themes or topics that can help evaluators understand how people felt about a program. One could generate a dataset much like the one used in this study. The column variables would be the name of the interviewee, rows would be populated with substantive words used to describe the program, and each cell would be populated with the number of times the word was used. A Rasch-based PCA could be performed and the output map investigated. The results could potentially help evaluators to see the connections between different participants and their responses.

Another example might include a study in which one is attempting to discern similarities and differences between various programs. For example, imagine a state university system that consists of multiple institutions and each have the freedom to develop their own institutional policies

for a given framework. In such instances when multiple facets are being compared across a number of institutions, easy comparisons can be made with the help of a stylometric analysis. In particular, institutions that have the most similar/different policies can be identified and reported in graphical form to conveniently display such policy similarities and differences to readers. Such findings might illustrate which institutions possess the most stringent policies, and which possess the most lax. This information could be quite convenient to a reader, as s/he would not have to plow through mounds of text and attempt to make the connections for his or herself.

Of course, numerous other possibilities exist for stylometrics in evaluation studies as well. One very exciting avenue for this methodology would be cheating detection. Plagiarism software could be developed based on these techniques. One example of a useful analysis might involve the comparing of student essays to a literary work. The extent to which students borrowed words or lifted phrases could be discerned and flags could be generated to inform the instructor of the similarities. The instructor could then provide a qualitative review of the two texts and make a judgment about cheating. The potential for stylometric techniques in the larger evaluation arena is truly unknown at the present time due to the infancy of this methodology. However, researchers and practitioners are encouraged to explore this methodology in a wide array of contexts to better determine its utility.

Conclusion

The methodology presented in this study appears to have a great deal of promise in the evaluation arena. While the substantive findings presented herein are in no way espoused to be definitive, it is hoped that the findings will nevertheless be helpful for providing an additional perspective to the question of Biblical authorship. It should be noted that the study of stylometrics is somewhat in its infancy, thus there are many lessons yet to be learned. With that said, much work needs to be done to improve these techniques. Similarly, the implications for the methods are not yet fully realized either. It is the researcher's hope that this study will spark the creativity of others and serve as a useful framework for future stylometric studies.

Acknowledgements

The author would like to thank Mr. Greg Hillis for his computer programming assistance acquiring the data to make this study possible. Additionally, the author would like to thank two anonymous reviewers for comments that greatly improved the final product.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Arrindell, W. A., & van der Ende, J. (1985). An empirical test of the utility of the observations-to-variable ratio in factor and components analysis. *Applied Psychological Measurement*, *9*(2), 165-178.

Barr, G. K. (2003). Two styles in the New Testament epistles. *Literacy & Linguistic Computing*, *18*(3), 235-248.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Kenny, A. (1986). *A stylometric study of the New Testament*. Oxford, UK: Clarendon Press

Linacre, J. M. (2001). Who wrote Paul's epistles? *Rasch Measurement Transactions*, *15*(1), 800-801.

Linacre, J. M. (2010). WINSTEPS® (Version 3.70.1.1). Computer Software. Beaverton, OR: Winsteps.com.

Linacre, J. M. (2011). *Dimensionality: Contrasts & variances*. Retrieved Jan. 31, 2012 from http://www.winsteps.com/winman/index.htm?principalcomponents.htm

Whissell, C. (2006). Comparison of the books of the New Testament (English Translation) in terms of emotion and word use. *Psychological Reports*, *98*, 57-64.