Cherry Picking in Nontraditional Authorship Attribution Studies

Joseph Rudman

How does "cherry picking" undermine the credibility of stylometry research, and why do improvements demand closer collaboration between literature experts and statisticians?

Cherry Picking in Nontraditional Authorship Attribution Studies

Joseph Rudman

Nontraditional authorship attribution studies are those studies that make use of the computer, statistics, and stylistics to identify the author(s) of an anonymous text. The literature leaves no doubt that there are major problems in this field—problems that must be addressed if there is to be general acceptance of the results not only by other practitioners of stylometry, but also by the general public. Cherry picking is one of these major problems and is the topic of this paper.

One definition of cherry picking is deliberately picking out the data or scientific studies that support your view, while ignoring the data or studies that oppose your view. (Norton, p.1) The first time I heard the term *cherry picking* in the context of nontraditional authorship attribution studies was at the 1995 Classification Society of North America (CSNA) meeting in Denver. A member of that audience stated that Ward Elliott "cherry picked" the style markers (i.e. quantifiable elements of style such as word length distributions or function word ratios) and statistical tests for his Shakespeare attribution study. I had been aware of the concept of selecting only favorable style markers to guarantee a pre-conceived statistical result, but I never really thought of it as cherry picking.
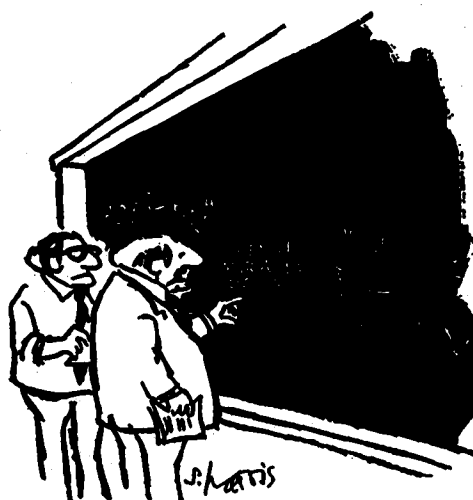
"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

After the 1995 CSNA meeting, Elliott and Valenza (1996 p. 203) used the term cherry picking in print. Foster (1996b) picked up on the term and their long running and at times personal feud was joined. There are limited references to the term cherry picking in the literature. Some of the references equate cherry picking to data mining, data dredging, or data snooping. However, cherry picking is an umbrella term that has multiple definitions even when confined to attribution studies. This paper presents an overview of six main types of cherry picking in nontraditional authorship attibution studies:

1. Selecting the 'text' — the primary data.
2. Selecting quantifiable style markers.
3. Selecting statistical tests.
4. Selecting controls.
5. Stopping the analysis.
6. Ex post facto analysis.

At first glance it might seem that statisticians would be more likely to cherry pick in 1, 2, and 4, and that non-statisticians would be more apt to cherry pick in 3, 5, and 6. But there are plenty of examples of cherry picking to go around for both groups in all six categories.

"OF COURSE YOU CAN'T REPLICATE MY EXPERIMENTS. THAT'S THE BEAUTY OF THEM."

Cherry Picking Violates the Scientific Method

Most investigators of similar stylo-statistical problems do not divulge how their samples were built up, or how sample size was estimated. It is, certainly, very sensible to leave out such compromising matter, for any attempt to lay down principles in these cases is liable to attract criticism. (Hargevik, 1974 Part I, p. 22)

Hargevik's (1974) study, *The Disputed Assignment of "Memoirs of an English Officer" to Daniel Defoe* is well respected in Defoe attribution circles. Hargevik is a literature scholar, yet his study violates principles required for a valid scientific experiment. Indeed many humanities scholars are not trained in scientific experimentation and method. I would argue that a majority of the nontraditional practitioners are not aware of the scientific demands of a valid attribution study — of the fact that all information must be provided so that the study can be replicated. On the other hand, many statisticians use authorship attribution as only a convenient area in which to test statistical techniques without appreciation for the complexities necessary to model the data meaningfully. For example, Gleser's statement that, "[Statistics] allows me to enter almost any field, and

without the need to get really deep information about the subject matter of that field..." (Leon Gleser, p. 21 in Clark). Literature scholars and statisticians must work together more closely so that each benefits from the strengths of the other, and to ensure that the principles of the scientific method are not violated.

Selecting the Text

There are various ways to cherry pick texts, most of which have to do with convenience:
- Using a 'bad' edition because it is available in machine-readable form
- Using collaborative texts
- Mixing genres or time periods because these are the only available texts
- Failing to look at the downloaded texts before starting the analysis

How many linguistic savvy practitioners (let alone hurried, harried, or unaware statisticians) have gone to the trouble to verify ambigous text? For example, homographs such as "to" can be either a preposition or an infinitive marker. The homograph "for" can be either a preposition or a conjunction. How many practitioners worry about clarifying the ambiguity between a single quotation mark and an apostrophe?

The Khmelev and Tweedie attribution study, "Using Markov Chains for

Identification of Writers" (2001), exemplifies one of the most rampant forms of cherry picking in attribution studies — taking textual data from any available source with little thought about the "authenticity" of the text or how editing may have changed it. The aim of their study was to use a large number of texts and avoid human intervention as much as possible. Yet their aim does not exculpate them from cherry picking the texts. As data for one third of their study, Khmelev and Tweedie downloaded 387 texts from the Project Gutenberg Archives (Project Gutenberg), representing forty-five authors writing in English. They do not take into consideration which editions were put into the Gutenberg Archives. An example of the problems that this poses can be seen in one of their selected authors, Daniel Defoe: Hargevik (1974, p. 30) mentions that the word "further" occurs twice in a first edition of the Defoe text but twenty seven times in a later edition. Khmelev and Tweedy make no genre distinction: poetry, drama, novels, and essays are all treated the same. They make no time distinction: they include John Milton (a 17th century author) and Upton Sinclair (a 20th century author). Notice also that these two authors are from different countries.

Other examples of 'selecting the text' are the Peng and Hengartner study (2002), Elliott and Valenza's "Shakespeare Clinic," the studies of the *Historia Augusta*, John Burrow's work on Sarah and Henry Fielding, and many other studies on authorship attribution in the works of Shakespeare. Peng and Hengartner analyzed critical apparatus (e.g. editorial comments and footnotes) as part of the author's text. Elliott and Valenza took editions from any source. "The Clinic took pride in having begged, borrowed, bought, typed, or scanned almost five million words of machine-readable Elizabethan text." (Elliott and Valenza 1991, p. 4) Foster (1996b) says that Elliott and Valenza stacked the deck by their choice of texts and that inconvenient data are simply omitted, generic and chronological variables ignored. Among the studies of the *Historia Augusta*, at least twelve by various authors simply use the text from the CD prepared by the Packard Humanities Institute — and only a few of the practitioners perform any type of editing. The

Historia Augusta text on the Packard CD was reconstructed from various manuscripts that are about 500 years removed from the original manuscript. There is no stemma chain — there is no chain of custody. There is no certainty or even preponderant probability as to how the *Historia Augusta* was written or as to how many authors were involved (Rudman 1998b).

John Burrows is one of the leading practitioners of stylistics and nontraditional authorship studies and one of the most scientifically rigorous. Burrows uses texts by Henry Fielding's sister, Sarah Fielding, both to analyze for Sarah's authorship of questioned texts and as controls in other studies. There is a major problem with all of Sarah's attributed writings. Burrows himself did a study that showed Henry Fielding wrote the beginning of the history of *Anna Boleyn* and that Sarah continued it for him. Burrows also showed that Henry either revised her ending, or added an ending in which he sought to imitate Anna's style (Burrows and Hassall). Henry is also known to have edited at least some of Sarah's work — e.g. "In correcting Sarah's *David Simple*..., Fielding changed every relative *that* to his preferred *which* for things and *who* or *whom* for persons. (Baker p.549)." Carolyn Woodward (2001) and other eighteenth century scholars argue that Sarah's writings are all collaborations, that there are no uncontaminated Sarah texts. Unless each text can be separated into Sarah and non-Sarah (and Burrows is one of the few who make an attempt), using her texts is cherry picking.

There are hundreds of studies on Shakespearean authorship, all with the same flaw — a cherry picked text. There is no Shakespeare dramatic text valid for nontraditional authorship attribution studies. Just what do we have when we look at the text of a Shakespeare drama? How many actual words and phrases were copied from his source material such as Holinshed's *Chronicles* or North's *Plutarch*? How many entire passages were paraphrased? What about the dramatic passages simply translated from the German drama? How many years elapsed from the date a Shakespeare play was first constructed and acted until the text we have was first published? Sixteen years is the closest we can come. And in those years, how

many directors, actors, copy scribes, pirate publishers, textual scholars, and editors made additions or changes, intentional or inadvertant? Drama is by its very nature a collaborative genre. And then we are doing authorship studies on these plays using hapax legomena (words that occur once) and rare word tests (McDonald).

Selecting Style Markers

Other researchers measure almost everything that is convenient, then look for variables that discriminate among texts of known authorship. But this can lead to cherry-picking. Between any two bodies of text, some of the enormously many measurable variables will be consistently similar solely by chance, even if both texts were written by different people; conversely, some features will be consistently different, even though the works have common authorship. (Banks p. 2)

The ... researcher can't just willynilly search through equidistant letter sequences until something interesting or unusual turns up. Statisticians call this cherry picking or data snooping, and it vitiates any statistical analysis. (Dembski p. 3)

The grouping now called BoB 2 was thrown out [by Elliott and Valenza] because it didn't produce enough of the expected rejections among the Claimant plays. (Foster 1996b, p. 251)

There is no standard methodology in the literature for selecting style markers. Practitioners have theorized that "unconscious" usage is a determinant — that "function" words are better than content-bearing words. But, however the style markers to be used are selected, it must be shown that the suspected "author" uses the chosen style markers in a consistent way in all of his known writings (of the same genre and same constricted time frame). The discriminating style markers are almost invariably different for each author.

Kenneth Neumann (1990) tested 617 style markers in his study of the Pauline Epistles. But that only scratches the surface — there are thousands of quantifiable style markers (Rudman 2000).

The cherry picking comes in when, for example, the practitioner identifies ten style markers that the "suspect author" uses consistently and that none of the controls uses in the same consistent manner. If only six of the ten markers show that the "suspect author" wrote the anonymous text and the practitioner ignores the other four markers that show the text was not by the "suspect author," this is cherry picking. It is also cherry picking to use only those markers that "worked" in other studies and ignore the markers that were not discriminatory in other studies. Again, the combination of discriminatory markers will be different in almost every case.

Selecting Statistical Tests (and/or Parameters)

You can take any data set, and if you cut it apart in a number of different ways, you'll find something eventually. But you can't just go in and pick out the one finding that's significant and say 'Aha, here it is,' when the other findings are not significant.

This tactic...exemplifies what's known in scientific research as the sharpshooter's fallacy, analogous to the way a gunslinger might empty his six-shooter into the side of a barn and then draw the bull's-eye around the bullet holes. (Richard P. Sloan [Quoted by Duenwald p. 2])

Some have criticized this process of trial and error for "multiple testing" or "cherry-picking:" bombarding our texts with so many tests that some would be bound to show some kind of distinction or identity between baseline text *a* and sample set *b*, simply by the law of averages, regardless of whether the distinction is true or false. (Elliott and Valenza 1996, p.203)

David Hoover has shown in his cluster analysis studies of texts by Emerson and Morson that the fifty most frequent words and the 100 most frequent words cluster these texts correctly. However, the thirty most frequent words and the 400 most frequent words fail to cluster these texts correctly (Hoover, p. 435).

What is the humanities practitioner to do but wonder about the validity of the previous published studies that have used cluster analysis — how many authors have cherry picked the parameters? Will future authorship studies that use cluster analysis have to test everything from the twenty most frequent words to the 500 most frequent words?

Practitioners cherry pick statistical tests not only from published attribution studies but also from other disciplines. Examples follow.

- The Efron-Thisted tests (expanded from Fisher) are from butterfly collecting.
- Simpson's index is based on the distribution of different species coexisting in a given ecosystem.
- The modal analysis used by Elliott's group is derived from signal processing.
- Morton's QSUM is based on industrial process and quality control monitoring.

The Efron-Thisted tests are based on the assumption that things (words) are well mixed in time. The assumption is that you will not capture all the members of one species early on and all of the members of another species later (Valenza). We must be leery of assumptions. We must be able to prove any assumptions behind statistical tests. Statistics should not be the tail that wags the dog of attribution studies (Rudman 1998a).

A variation under this heading is not showing the standard errors associated with statistical estimates. Many humanities practitioners have done this. The two plots in Figure 1 give different impressions. Plot (a) adds credence to Defoe being the author of the anonymous text. The error bars on (b) shows that uncertainty dominates.

Selecting Controls

In the review of Radday's work by Morton, it appears that Morton does not realize that erroneous methodological procedures have been used. Differences in usage of literary elements found in a text are not necessarily an indication of differences in authorship, unless inter-authorship parameters have been compared to intra-authorship parameters to establish critical levels of variation. (La Mar Adams p. 86)

Different types of controls are required in nontraditional authorship attribution studies:

- A random subset of the author's touchstone texts should be withheld as a 'penultimate analysis' control.
- A substantial random sample of all of the non-anonymous, non-questioned, non-collaborative authors of the same period in the same genre, from the same geographical region writing in the same native language.
- The other writers who have been mentioned as possible authors of the questioned text.

While my knowledge of the attribution literature is greatest in the type of study asking, "Did author A write the anonymous text?" I know of no published study that has correctly and completely employed these controls. Many researchers who are aware of the 'ideal' controls, time and financial constraints, unavailability of texts, or difficulty using the texts are stated reasons for falling short of the ideal.

Examples of cherry picking under this heading are rife. The most publicized is Donald Foster's attribution of *A Funerall Elegie* to Shakespeare. Foster now agrees that his data base of control texts was inadequate (Niederkorn). This kind of flaw in such a highly publicized study sets back the 'cause' of nontraditional authorship attribution and casts doubt on all other like studies. Foster himself writes:

But it must be viewed as a major setback when one highly visible study can be cited by skeptics as proof that the whole quantitative enterprise is fruitless, or a playground for fringe theories having
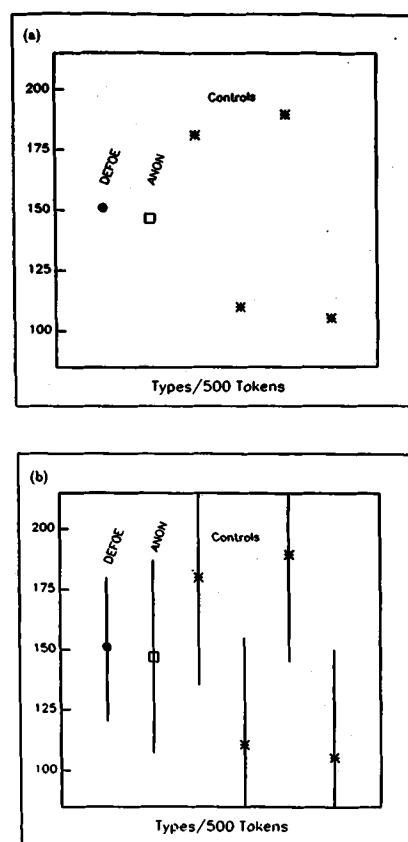

Figure 1. Error bars reflecting uncertainty of statistical estimates.

no historical or computational validity. (Foster 1996b, p. 255)

Foster is a brilliant and innovative practitioner, and most of his work is to be admired and emulated. Elliott and Valenza (2002, p. 459), in discussing Foster's work, explain why the use of the maxim, *one erroneous item means everything is erroneous*, should not apply to him or to other nontraditional practitioners. This applies to Elliott and Valenza's work as well. It is up to the "gatekeepers" to point out the mistakes and to explain what must be done to avoid them in future studies.

Other cherry picking examples under 'selecting controls' are the twelve attribution studies of the *Historia Augusta* mentioned earlier. All provide classic examples of invalid controls:

- The control texts were not written in the same time period — many are centuries apart.
- Most of the control texts are of a different genre — one, the *Codex*

Theodosianus, is a handbook of law codes.

- Some control texts were published anonymously.
- Some control texts were written by non-native speakers.

In *The Disputed Assignment of* "Memoirs of an English Officer" *to Daniel Defoe*, Steig Hargevik used a 1,000,000 word, non-random sample of selected "other" authors, cherry picked because of availability. Some of his control texts were published anonymously.

Joel Best (2001, p. 53), in his *Damned Lies and Statistics*, says, "...large samples aren't necessarily good samples ... the representativeness of a sample is actually far more important than sample size." The need for a random sample cannot be overpowered by increasing the size of a non-random sample. This may be a truism to statisticians, but is not obvious to other authorship practitioners. (See Johnstone (1989) for a treatment of the pros and cons of the necessity of random sampling.)
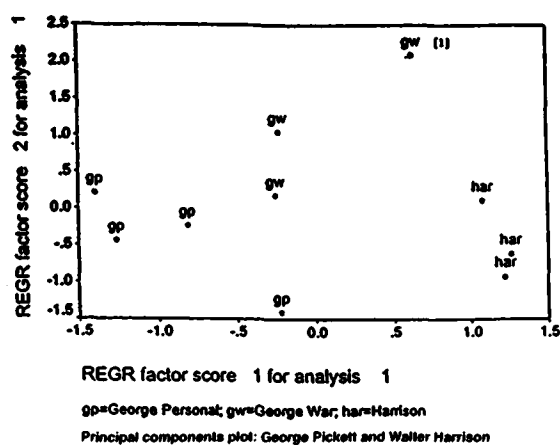
Of course, there is the other side of the large sample coin — the insufficient sample. Norton (2002, p. 1) quotes the famous example, "After treatment with the drug, one third of the mice were cured, one third died, and the third mouse escaped."

Stopping the Analysis

The esteemed Israeli mathematician Eliyahu Rips checked the 716 letter passage by computer for appearances of AHRN at equidistant intervals, and found 25. Mathematical analysis predicted 8.3, and the odds against 25 occurrences was calculated at 1 in about 2 million. (The math is undisputed, but the critics say that the statistical effect vanishes when a somewhat longer passage is tested.) (Johnson p. 4)

Foster (1996a), in his *Primary Colors* attribution study, was presented with thirty-five likely authors of the anonymous novel *Primary Colors*. Foster did not test all thirty-five candidates. He stopped after fifteen because he felt that he had the answer — Joe Kline. The fact that he was proven correct doesn't excuse his stopping the experiment — potentially, there was a "better" fit or some fits that could have clouded the issue.

The most famous example of prematurely stopping the experiment is from DNA anaylsis. In Great Britain, an innocent man was matched to a burglary by a DNA comparison test that used six loci. Only after he provided an irrefutable alibi was the DNA test rerun using ten loci. The second test showed that there was no match (Willing 2002). This might not seem to fit in with authorship attribution. However, Rudman's (2002) "DNA and Nontraditional Authorship Attribution: An Inclusive Model" shows a direct and meaningful analogy.

REGR factor score 1 for analysis 1

gp=George Personal; gw=George War; har=Harrison

Principal components plot: George Pickett and Walter Harrison

Dendrogram: George Pickett and Walter Harrison.

Figure 2. Figure from Holmes et al (2001). Reprinted with permission of the authors.

Ex Post Facto Analysis

[Rips] admitted that no protocol had been used in the selection process. In addition, the formula selected for measuring distance within the pairs was one of many possibilities — and favorable to the Rips group's results. Furthermore, if a rabbi on the list could not be located in Genesis, he was simply dropped. "They used what worked and ignored what didn't. ..." (Cohen p. 61)

Most ex post facto analysis, by its very nature, goes undetected. Practitioners usually don't report that they dropped texts, went back and altered texts, or changed controls to show a desired result.

Peng and Hengartner (2002) downloaded the texts for their study, "Quantitative Analysis of Literary Styles," from the Project Gutenberg website. It is obvious from their paper that they did not look at the texts before analysis. However, after they found some unexpected outliers on their principal component analysis plots, they went back and took a look at the sections of the texts in ques-
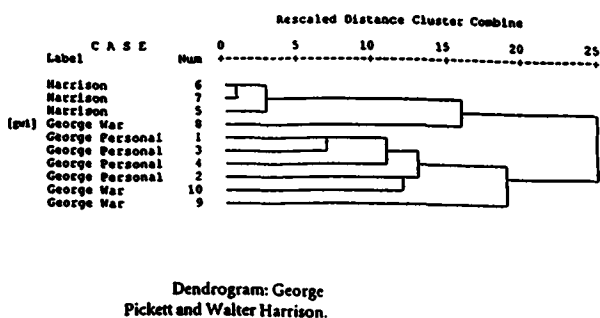
We had better publish this quick before our numbers change again

Note that 'gw1' is an outlier. They looked carefully at this outlier text sample and concluded that it was tainted — either by ghost writing or by collaboration. They then eliminated it from the rest of the study. Is this cherry picking or is it a valid experimental technique? A tainted sample must be expurgated at any stage of analysis. As long as Holmes et al. looked at all of the samples with the same criteria, this is not cherry picking. However, the question would never have arisen (at least in my mind) if the data selection had eliminated the text before statistical analysis began.

Suggestions for Improved Authorship Attribution Studies

Articulating the problem of cherry picking goes a long way towards its elimination. As a philosopher once said, "You have to know that you don't know." Practitioners should reach for the ideal of a perfect experiment — no mistakes, no cherry picking. The only tolerable level of cherry picking is that point at which it doesn't invalidate the experimental results. Practitioners must know, or at least be aware of the totality of nontraditional authorship attribution studies: how to set up a complete and valid experiment, how to avoid all fatal errors. Practitioners must agree on what questions to ask, such as, 'what is a word?' and 'what is a unique word?' These are not simple questions.

Nontraditional authorship attribution is interdisciplinary by nature, but the various disciplines cannot remain compartmentalized. The literature scholar (or scholar of history, religion, or law) must know the right questions to ask the statistician and the linguist — and vice versa in all the combinations — e.g. what assumptions about the text does a particular statistical test make, or another more specific example, is it "...assumed that the vocabulary is fixed and finite, and that the author writes ... by successively drawing words from this collection, independently of the previous configuration.(McNeil p. 92)" Only with thoughtful and educated collaboration can cherry picking in authorship studies be eliminated, or at least minimized to a point of demonstrable insignificance. **C**

tion. And they found that "...the downloaded versions...were critical editions containing extensive footnotes and commentary... (Peng and Hengartner p.179)." They deleted the editorial additions and the outliers went away. There is no indication that they also looked at the rest of the downloaded texts to see if anything else had to be dealt with by getting the correct edition, deleting editorial interventions, or editing. Had they done so, this example would be placed into the next section, 'Seeming Cherry Picking'.

Seeming Cherry Picking

The following are a few examples of "seeming" cherry picking — methods that might seem to belong to one side or the other (statistician versus the rest of the attribution world) as cherry picking. The first example is the Ward Elliott style marker selection that was mentioned earlier. It turned out that Elliott eliminated style markers that did not differentiate authors — a completely valid decision. Many style markers are zeitgeist driven — they are used equally (within errors) by almost all authors under consideration. Elliott also elimi-

nated those style markers that the suspect author did not use consistently. Again, a valid decision.

Another example of seeming cherry picking is a practitioner who switches to a log scale in order to more clearly show grouping tendencies. To the uninitiated, it seems that practitioners keep switching scales until one is found that clearly shows what you want it to show. Williams, in a sentence length study of Shaw, Wells, and Chesterton, discusses a switch of histograms from number of words per sentence to a log scale histogram. What caused him to switch was that the frequency distributions of the actual number of words per sentence struck him, "...by their general resemblance to certain skew distributions ... in some Entomological problems..." (Williams p. 356). He does go on to give other good statistical reasons for the switch. Posner (1963), in "Use and Abuse of Stylistic Statistics" discusses this in a little more detail.

The final example is what would seem to be ex post facto tweaking. In their authorship study of the George Pickett letters, Holmes (arguably the leading practitoner in the field) et al. (2001) show a plot and dendrogram reproduced here as Figure 2.

References

Baker, S. (1989), "Did Fielding Write 'A Vision?'" *Eighteenth Century Studies*, 22, 548–551.

Banks, D. (1996), "Questioning Authority," *Classification Society of North America Newsletter*, 44, 1.

Best, J. (2001), *Damned Lies and Statistics*, Berkeley: University of California Press.

Burrows, J. F., and Hassall, A. J. (1988), "*Anna Boleyn* and the Authenticity of Fielding's Feminine Narratives," *Eighteenth Century Studies*, 21, 427–453.

Clark, T. (1996), "A Passion for Statistics," *Ventures* (University of Pittsburgh), 6,2, 19–21.

Cohen, H. (1998), "God Only Knows," *Lingua Franca*, 8, (July/August).

Dembski, W. A. (1998), "The Bible by Numbers." [Review of *Cracking the Bible Code*. Jeffrey Satinover. Morow (1998)] *First Things*, 85, 61–64.

Duenwald, M. (2002), "Religion and Health: New Research Revives an Old Debate," *The New York Times*, Tuesday, May 7, 2002, D5.

Elliott, W. E. Y., and Valenza, Robert. J. (2002), "So Many Hardballs, So Few Over the Plate: Conclusions from our 'debate' with Donald Foster," *Computers and the Humanities*, 36, 455–460.

Foster, D. W. (1996), "Primary Culprit: An Analysis of a Novel of Politics," *New York*, 26 February 1996, 50–57.

——(1996), "And Then There was None: Winnowing the Shakespeare Claimants," *Computers and the Humanities*, 30, 191–245.

—— (1991), "Who Was Shakespeare?" (Update of Summary for *Chance*) Preprint supplied by the authors.

——(1996), "Response to Elliot [sic] and Valenza, 'And Then There Were None,'" *Computers and the Humanities*, 30, 247–255.

Hargevik, S. (1974), *The Disputed Assignment of* "Memoirs of an English Officer" *to Daniel Defoe*. (Part 1) (Stockholm Studies in English XXX), Stockholm: Almqvist and Wiksell.

Holmes, D. I., et al. (2001), "A Widow and Her Soldier: Stylometry and the American Civil War," *Literary and Linguistic Computing*, 16, 4, 403–420.

Hoover, D. L. (2001), "Statistical Stylistics and Authorship Attribution: An Empirical Investigation," *Literary and Linguistic Computing*, 16, 4, 421-444.

Johnson, P. E. (1998), "What Would Newton Do?" *First Things*, 87, 25–31.

Johnstone, D. J. (1989), "On the Necessity for Random Sampling," *British Journal for the Philosophy of Science*, 40, 4, 443–457.

Khmelev, D. V., and Tweedie, F. J. (2001), "Using Markov Chains for Identification of Writers," *Literary and Linguistic Computing*, 16, 3, 299–307.

La Mar Adams L. (1974), "Literary Statistics." (Under 'Correspondence') *ALLC Bulletin*, 2, 2, 85–87.

McDonald, R. (1996), *The Bedford Companion to Shakespeare: An Introduction with Documents*, Boston: Bedford Books of St. Martin's Press.

McNeil, D. R. (1973), "Estimating an Author's Vocabulary," *Journal of the American Statistical Association*, 68, 341, 92–96.

Neumann, K. J. (1990), *The Authenticity of the Pauline Epistles in the Light of Stylostatistical Analysis*, Atlanta, Georgia: Scholars Press (Society of Biblical Literature Dissertation Series Number 120).

Niederkorn, W. S. (2002), *The New York Times*, 20 June, B1, B5.

Norton, J. (2002), "Fallacies of Evidence and Causality," *members.aol.com/jimn469897/evidence.htm*, 04-20-02.

Peng, R. D. and Hengartner, N. W. (2002), "Quantitative Analysis of Literary Styles," *The American Statistician*, 56, 3, 175–185.

Posner, R. (1963), "The Use and Abuse of Stylistic Statistics," *Archivum Linguisticum*, 15, 111–139.

Project Gutenberg Web Site. *promo.net/pg/*

Rudman, J. (2002), "'DNA' and Nontraditional Authorship Attribution: An Inclusive Model." Paper delivered at the ALLC/ACH 2002 conference in Tübingen, Germany, July 2002.

——(2000), "The Style-Marker Mapping Project: A Rational and Progress Report." Paper delivered at the ALLC/ACH 2000 conference in Glasgow, Scotland, July 2000.

——(1998), "The State of Authorship Attribution Studies: Some Problems and Solutions," *Computers and the Humanities*, 31, 351–365.

——(1998), "Nontraditional Authorship Attribution Studies in the *Historia Augusta*: Some Caveats," *Literary and Linguistic Computing*, 13, 3, 151–157.

Valenza, R. J. (1991), "Are the Thisted-Efron Authorship Tests Valid?" *Computers and the Humanities*, 25, 1, 27–46.

Williams, C. B. (1939-1940), "A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary Style," *Biometrika*, 31, 356–361.

Willing, R. (2002), "Mismatch Calls DNA Tests Into Question," *USA Today*, 8 February 2000, 3A.

Woodward, C. (2001), "Sarah Fielding and the Inscription of Authorship in *The Cry: A New Dramatic Fable* (1754)," Paper delivered at the 32 Annual Meeting of the American Society for Eighteenth Century Studies, New Orleans, April 2001.