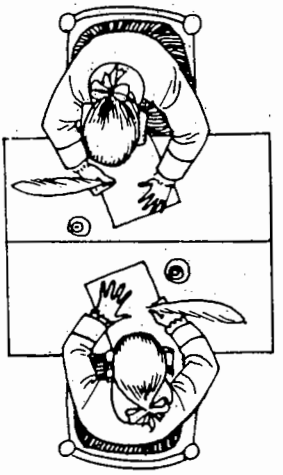

Deciding Authorship

Frederick Mosteller *Harvard University*
David L. Wallace *University of Chicago*

Art, music, literature, and the social, biological, and physical sciences share a common need to classify things: What artist painted the picture? Who composed the piece? Who wrote the document? If paroled, will the prisoner repeat the crime? What disease does the patient have? What trace chemical is damaging the process? In the field of statistics, we call these questions *classification* or *discrimination* problems.

Questions of authorship are frequent and sometimes important. Most people have heard of the Shakespeare-Bacon-Marlowe controversy over who wrote the great plays usually attributed to Shakespeare. A less well known but carefully studied question deals with the authorship of a number of Christian religious writings called the Paulines, some being books in the New Testament: Which ones were written by Paul and which by others? In many authorship questions the solution is easy once we set about counting something systematically. But we treat here an especially difficult problem from American history, the

controversy over the authorship of the 12 *Federalist* papers claimed by both Alexander Hamilton and James Madison, and we show how a statistical analysis can contribute to the resolution of historical questions.

The Federalist papers were published anonymously in 1787–1788 by Alexander Hamilton, John Jay, and James Madison to persuade the citizens of the state of New York to ratify the Constitution. Seventy-seven papers appeared as letters in New York newspapers over the pseudonym Publius. Together with eight more essays, they were published in book form in 1788 and have been republished repeatedly both in the United States and abroad. *The Federalist* remains today an important work in political philosophy. It is also the leading source of information for studying the intent of the framers of the Constitution, as, for example, in decisions on congressional reapportionment, since Madison had taken copious notes at the Constitutional Convention.

It was generally known who had written *The Federalist,* but no public assignment of specific papers to authors occurred until 1807, three years after Hamilton's death as a result of his duel with Aaron Burr. Madison made his listing of authors only in 1818 after he had retired from the presidency. A variety of lists with conflicting claims have been disputed for a century and a half. There is general agreement on the authorship of 70 papers—5 by Jay, 14 by Madison, and 51 by Hamilton. Of the remaining 15, 12 are in dispute between Hamilton and Madison, and 3 are joint works to a disputed extent. No doubt the primary reason the dispute exists is that Madison and Hamilton did not hurry to enter their claims. Within a few years after writing the essays, they had become bitter political enemies and each occasionally took positions opposing some of his own *Federalist* writings.

The political content of the essays has never provided convincing evidence for authorship. Since Hamilton and Madison were writing a brief in favor of ratification, they were like lawyers working for a client; they did not need to believe or endorse every argument they put forward favoring the new Constitution. While this does not mean that they would go out of their way to misrepresent their personal positions, it does mean that we cannot argue, "Hamilton wouldn't have said that because he believed otherwise." And, as we have often seen, personal political positions change. Thus the political content of a disputed essay cannot give strong evidence in favor of Hamilton's or of Madison's having written it.

The acceptance of the various claims by historians has tended to change with political climate. Hamilton's claims were favored during the last half of the nineteenth century, Madison's since then. While the thorough historical studies of the historian Douglass Adair over several decades support the Madison claims, the total historical evidence is today not much different from that which historians like the elder Henry Cabot Lodge interpreted as favoring Hamilton. New evidence was needed to obtain definite attributions, and internal statistical stylistic evidence provides one possibility; developing that evidence and the methodology for interpreting it is the heart of our work.

The writings of Hamilton and Madison are difficult to tell apart because both authors were masters of the popular *Spectator* style of writing—complicated and oratorical. To illustrate the difficulty, in 1941 Frederick Williams and Frederick Mosteller counted sentence lengths for the undisputed papers and got averages of 34.5 and 34.6 words for Hamilton and Madison, respectively. For sentence length, a measure used successfully to distinguish other authors, Hamilton and Madison are practically twins.

MARKER WORDS

Although sentence length does measure complexity (and an average of 35 words shows that the material is very complex), sentence length is not sensitive enough to distinguish reliably between authors writing in similar styles. The variables used in several studies of disputed authorship are the rates of occurrence of specific individual words. Our study was stimulated by Adair's discovery—or rediscovery as it turned out—that Madison and Hamilton differ consistently in their choice between the alternative words *while* and *whilst.* In the 14 *Federalist* essays acknowledged to be written by Madison, *while* never occurs whereas *whilst* occurs in 8 of them. *While* occurs in 15 of 48 Hamilton essays, but never a *whilst.* We have here an instance of what are called *markers*—items whose presence provides a strong indication of authorship for one of the men. Thus the presence of *whilst* in 5 of the disputed papers points toward Madison's authorship of those 5.

Markers contribute a lot to discrimination when they can be found, but they also present difficulties. First, *while* or *whilst* occurs in less than half of the papers. They are absent from the other half and hence give no evidence either way. We might hope to surmount this by finding enough different marker words or constructions so that one or more will always be present. A second and more serious difficulty is that from the evidence in 14 essays by Madison, we cannot be sure that he would never use *while.* Other writings of Madison were examined and, indeed, he did lapse on two occasions. The presence of *while* then is a good but not sure indication of Hamilton's authorship; the presence of *whilst* is a better, but still imperfect, indicator of Madison's authorship, for Hamilton too might lapse.

A central task of statistics is making inferences in the presence of uncertainty. Giving up the notion of perfect markers leads us to a statistical problem. We must find evidence, assess its strength, and combine it into a composite conclusion. Although the theoretical and practical problems may be difficult, the opportunity exists to assemble far more compelling evidence than even a few nearly perfect markers could provide.

RATES OF WORD USE

Instead of thinking of a word as a marker whose presence or absence settles the authorship of an essay, we can take the *rate* or *relative frequency* of the use of each word as a measure pointing toward one or the other author. Of

course, most words won't help because they were used at about the same rate by both authors. But since we have thousands of words available, some may help. Words form a huge pool of possible discriminators. From a systematic exploration of this pool of words, we found no more pairs like *while-whilst,* but we did find single words used by one author regularly but rarely by the other.

Table 1 shows the behavior of three words: *commonly, innovation,* and *war.* The table summarizes data from 48 political essays known to be written by Hamilton and 50 known to be by Madison. (Some political essays from outside *The Federalist,* but known to be by Hamilton or Madison, have been included in this study to give a broader base for the inference. Not all of Hamilton's later *Federalist* papers have been included. We gathered more papers from outside *The Federalist* for Madison.)

Neither Hamilton nor Madison used *commonly* much, but Hamilton's use is much more frequent than Madison's. The table shows that in 31 of 48 Hamilton papers, the word *commonly* never occurs, but that in the other 17 it occurs one or more times. Madison used it only once in the 50 papers in our study. The papers vary in length from 900 to 3,500 words, with 2,000 about average. Even one occurrence in 900 words is a heavier usage than two occurrences in 3,500 words, so instead of working with the number of occurrences in a paper, we use the rate of occurrence, with 1,000 words as a convenient base. Thus, for example, the paper with the highest rate (1.33 per 1,000 words) for *commonly* is a paper of 1,500 words with two occurrences. *Innovation* behaves similarly, but it is a marker for Madison. For each of these two words, the highest rates are a little over 1 per 1,000.

Table 1 Frequency distributions of rate per 1,000 words in 48 Hamilton and 50 Madison papers for *commonly, innovation,* and *war*

Commonly			*Innovation*			*War*		
Rate per 1,000 Words	*H*	*M*	*Rate per 1,000 Words*	*H*	*M*	*Rate per 1,000 Words*	*H*	*M*
0 (exactly)*	31	49	0 (exactly)*	47	34	0 (exactly)*	23	15
0^+–0.2	Cannot occur†		0^+–0.2	Cannot occur†		0^+–2	16	13
0.2–0.4	3	1	0.2–0.4		6	2–4	4	5
0.4–0.6	6		0.4–0.6	1	6	4–6	2	4
0.6–0.8	3		0.6–0.8		1	6–8	1	3
0.8–1.0	2		0.8–1.0		2	8–10	1	3
1.0–1.2	2		1.0–1.2		1	10–12		3
1.2–1.4	1					12–14		2
						14–16	1	2
Totals	48	50	Totals	48	50	Totals	48	50

*Each interval, except 0 (exactly), excludes its upper endpoint. Thus a 2,000-word paper in which *commonly* appears twice gives rise to a rate of 1.0 per 1,000 exactly, and the paper appears in the count for the 1.0–1.2 interval.

†With the given lengths of the papers used, it accidentally happens that a rate in this interval cannot occur. For example, if a paper has 2,000 words, a rate of 1 per 1,000 means 2 words, and a single occurrence means a rate of 0.5 per 1,000. Hence a 2,000-word paper cannot lead to a rate per thousand greater than 0 and less than 0.5.

Source: Mosteller and Wallace (1984).

The word *war* has spectacularly different behavior. Although absent from half of Hamilton's papers, when present it is used frequently—in one paper at a rate of 14 per 1,000 words. *The Federalist* papers deal with specific topics in the Constitution and huge variations in the rates of such words as *war, law, executive, liberty,* and *trade* can be expected according to the context of the paper. Even though Madison uses *war* considerably more often than Hamilton in the undisputed papers, we explain this more by the division of tasks than by Madison's predilection for using *war.* Data from use of a word like *war* would give the same troublesome sort of evidence that historians have disagreed about over the last hundred years. Indeed, the dispute has continued because evidence from subject and content has been hopelessly inconclusive.

USE OF NONCONTEXTUAL WORDS

For the statistical arguments to be valid, information from meaningful, contextual words must be largely discarded. Such a study of authorship will not then contribute directly to any understanding of the greatness of the papers, but the evidence of authorship can be both strengthened and made independent of evidence provided by historical analysis.

Avoidance of judgments about meaningfulness or importance is common in classification and identification procedures. When art critics try to authenticate a picture, in addition to the historical record, they consider little things: how fingernails and ears are painted and what kind of paint and canvas were used. Relatively little of the final judgment is based upon the painting's artistic excellence. In the same way, police often identify people by their fingerprints, dental records, and scars, without reference to their personality, occupation, or position in society. For literary identification, we need not necessarily be clever about the appraisal of literary style, although it helps in some problems. To identify an object, we need not appreciate its full value or meaning.

What noncontextual words are good candidates for discriminating between authors? Most attractive are the filler words of the language: prepositions, conjunctions, articles. Many other more meaningful words also seem relatively free from context: abverbs such as *commonly, consequently, particularly,* or even abstract nouns like *vigor* or *innovation.* We want words whose use is unrelated to the topic and may be regarded as reflecting minor or perhaps unconscious preferences of the author.

Consider what can be done with filler words. Some of these are the most used words in the language: *the, and, of, to,* and so on. No one writes without them, but we may find that their rates of use differ from author to author. Table 2 shows the distribution of rates for three prepositions—*by, from,* and *to.* First, note the variation from paper to paper. Madison uses *by* typically about 12 times per 1,000 words, but sometimes he has rates as high as 18 or as low as 6. Even on inspection though, the variation does not obscure Madison's systematic tendency to use *by* more often than Hamilton does. Thus low rates for *by* suggest Hamilton's authorship, and high rates Madison's. Rates for *to* run in the opposite direction. Very high rates for *from* point to Madison but low

Table 2 Frequency distribution of rate per 1,000 words in 48 Hamilton and 50 Madison papers for *by, from,* and *to*

By			*From*			*To*		
Rate per 1,000 Words	*H*	*M*	*Rate per 1,000 Words*	*H*	*M*	*Rate per 1,000 Words*	*H*	*M*
1–3*	2		1–3*	3	3	20–25*		3
3–5	7		3–5	15	19	25–30	2	5
5–7	12	5	5–7	21	17	30–35	6	19
7–9	18	7	7–9	9	6	35–40	14	12
9–11	4	8	9–11		1	40–45	15	9
11–13	5	16	11–13		3	45–50	8	2
13–15		6	13–15		1	50–55	2	
15–17		5				55–60	1	
17–19		3						
Totals	48	50	Totals	48	50	Totals	48	50

*Each interval excludes its upper endpoint. Thus a paper with a rate of exactly 3 per 1,000 words would appear in the count for the 3–5 interval.

Source: Mosteller and Wallace (1984).

rates give practically no information. The more widely the distributions are separated, the stronger the discriminating power of the word. Here, *by* discriminates better than *to,* which in turn is better than *from.*

PROBABILITY MODELS

To apply any of the theory of statistical inference to evidence from word rates, we must construct an acceptable probability model to represent the variability in word rate from paper to paper. Setting up a complete model for the occurrence of even a single word would be a hopeless task, for the fine structure within a sentence is determined in large measure by nonrandom elements of grammar, meaning, and style. But if our interest is restricted to the rates of use of one or more words in blocks of text of at least 100 or 200 words, we expect that detailed structure of phrases and sentences ought not to be very important. The simplest model can be described in the language of balls in an urn, so common in classical probability. To represent Madison's usage of the word *by,* we suppose there is a typical Madison rate, which would be somewhere near 12 per 1,000, and we imagine an urn filled with many thousands of red and black balls, with the red occurring in the proportion 12 per 1,000. Our probability model for the occurrence of *by* is the same as the probability model for successive draws from the urn, with a red ball corresponding to *by* and a black ball corresponding to all other words. To extend the model to the simultaneous study of two or more words, we would need balls of three or more colors. No grammatical structure or meaning is a part of this model, and it is not intended to represent behavior within sentences. What is desired is that it explain the variation in rates—in counts of occurrences in long blocks of words, corresponding to the essays.

We tested the model by comparing its predictions with actual counts of word frequencies in the papers. We found that while this urn scheme reproduced variability well for many words, for other words additional variability was required. The random variation of the urn scheme represented most of the variation in counts from one essay to another, but in some essays the authors changed their basic rates a bit. We had to complicate the theoretical model to allow for this, and the model we used is called the ***negative binomial distribution.***

The test showed also that pronouns like *his* and *her* are exceedingly unreliable authorship indicators, worse even than words like *war.*

INFERENCE AND RESULTS

Each possible route from construction of models to quantitative assessment of, say, Madison's authorship of some disputed paper, required solutions of serious theoretical statistical problems, and new mathematics had to be developed. A chief motivation for us was to use the *Federalist* problem as a case study for comparing several different statistical approaches, with special attention to one, called the Bayesian method, that expresses its final results in terms of probabilities, or odds, of authorship.

By whatever methods are used, the results are the same: overwhelming evidence for Madison's authorship of the disputed papers. For only one paper is the evidence more modest, and even there the most thorough study leads to odds of 80 to 1 in favor of Madison.

Figures 1 and 2 illustrate how the 12 disputed papers fit the distributions of Hamilton's and Madison's rates for two of the words finally chosen as discriminators. In Figure 1 the top two histograms portray the data for *by* that was given earlier in Table 2. Madison's rate runs higher on the average. Compare the bottom histogram for the disputed papers first with the top histogram for Hamilton papers, then with the second one for Madison papers. The rates in the disputed papers are, taken as a whole, very Madisonian, although 3 of the 12 papers by themselves are slightly on the Hamilton side of the typical rates. Figure 2 shows the corresponding facts for *to.* Here again the disputed papers are consistent with Madison's distribution, but further away from the Hamilton behavior than are the known Madison papers.

Table 3 shows the 30 words used in the final inference, along with the estimated mean rates per thousand in Hamilton's and Madison's writings. The groups are based upon the degree of contextuality anticipated by Mosteller and Wallace (1984) prior to the analysis.

The combined evidence from nine common filler words shown as group B was huge—much more important than the combined evidence from 20 low-frequency marker words like *while-whilst.* These 20 are shown as groups C, D, and E.

There remains one word that showed up early as a powerful discriminator, sufficient almost by itself. When should one write *upon* instead of *on*? Even authoritative books on English usage don't provide good rules. Hamilton and

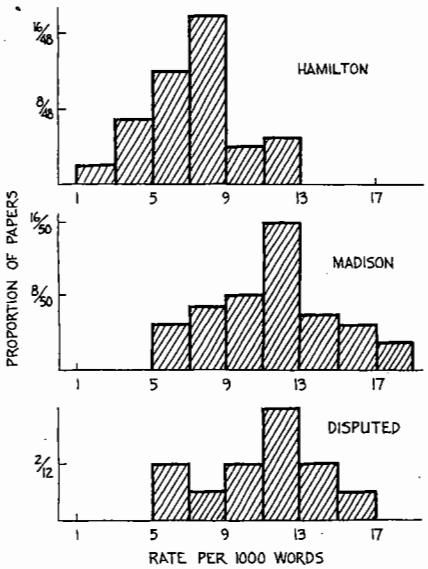

Figure 1 *Distribution of rates of occurrence of* by *in 48 Hamilton papers, 50 Madison papers, 12 disputed papers.*

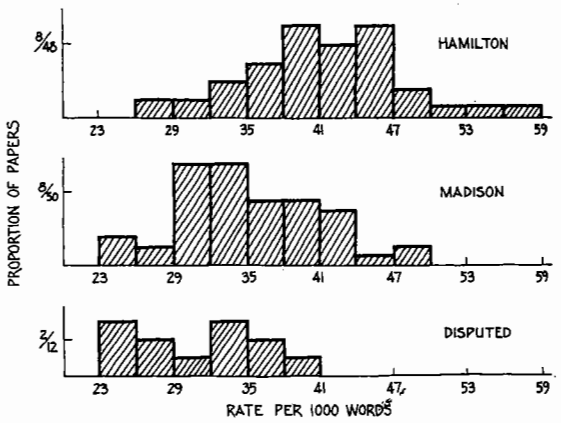

Figure 2 *Distribution of rates of occurrence of* to *in 48 Hamilton papers, 50 Madison papers, 12 disputed papers.*

Madison differ tremendously. Hamilton writes *on* and *upon* almost equally, about 3 times per 1,000 words. Madison, on the other hand, rarely uses *upon.* Table 4 shows the distributions for *upon.* In 48 papers Hamilton never failed to use *upon;* indeed, he never used it less than twice. Madison used it in only 9 of 50 papers, and then only with low rates. The disputed papers are clearly Madisonian with *upon* occurring in only 1 paper. That paper, fortunately, is strongly classified by the other words. It is not the paper with modest overall odds.

Table 3 Words used in final discrimination and adjusted rates of use in text by Madison and Hamilton

Word	Rate Per 1,000 Words		Word	Rate Per 1,000 Words	
	Hamilton	*Madison*		*Hamilton*	*Madison*
Group A			*Group D*		
Upon	3.24	0.23	Commonly	0.17	0.05
Group B			Consequently	0.10	0.42
Also	0.32	0.67	Considerable(ly)	0.37	0.17
An	5.95	4.58	According	0.17	0.54
By	7.32	11.43	Apt	0.27	0.08
Of	64.51	57.89	*Group E*		
On	3.38	7.75	Direction	0.17	0.08
There	3.20	1.33	Innovation(s)	0.06	0.15
This	7.77	6.00	Language	0.08	0.18
To	40.79	35.21	Vigor(ous)	0.18	0.08
Group C			Kind	0.69	0.17
Although	0.06	0.17	Matter(s)	0.36	0.09
Both	0.52	1.04	Particularly	0.15	0.37
Enough	0.25	0.10	Probability	0.27	0.09
While	0.21	0.07	Work(s)	0.13	0.27
Whilst	0.08	0.42			
Always	0.58	0.20			
Though	0.91	0.51			

Source: Mosteller and Wallace (1984).

Table 4 Frequency distribution of rate per 1,000 words in 48 Hamilton, 50 Madison, and 12 disputed papers for *upon*

Rate per 1,000 Words	*Hamilton*	*Madison*	*Disputed*
0 (exactly)*		41	11
0^{+}–0.4		2	
0.4–0.8		4	
0.8–1.2	2	1	1
1.2–1.6	3	2	
1.6–2.0	6		
2.0–3.0	11		
3.0–4.0	11		
4.0–5.0	10		
5.0–6.0	3		
6.0–7.0	1		
7.0–8.0	1		
Totals	48	50	12

*Each interval, except 0 (exactly), excludes its upper endpoint. Thus a paper with a rate of exactly 3 per 1,000 words would appear in the count for the 3.0–4.0 interval.

Source: Mosteller and Wallace (1984).

Of course, combining and assessing the total evidence is a large statistical and computational task. High-speed computers were employed for many hours in making the calculations, both mathematical calculations for the theory and empirical ones for the data.

You may have wondered about John Jay. Might he not have had a hand in the disputed papers? Table 5 shows the rates per thousand for nine words of highest frequency in the English language measured in the writings of Hamilton, Madison, Jay, and, for a change of pace, in James Joyce's *Ulysses*. The table supports the repeated assertion that Madison and Hamilton are similar. Joyce is much different, but so is John Jay. The words *of* and *to* with rate comparisons 65/58 and 41/35 were among the final discriminators between Hamilton and Madison. See how much more easily Jay could be discriminated from either Hamilton or Madison by using *the, of, and, a,* and *that*. The disputed papers are not at all consistent with Jay's rates, and there is no reason to question his omission from the dispute.

SUMMARY OF RESULTS

Our data independently supplement the evidence of the historians. Madison is extremely likely, in the sense of degree of belief, to have written the disputed *Federalist* papers, with the possible exception of paper 55, and there our evidence yields odds of 80 to 1 for Madison—strong, but not overwhelming. Paper 56, next weakest, is a very strong 800 to 1 for Madison. The data are overwhelming for all the rest, including the two papers historians feel weakest about, papers 62 and 63.

For a more extensive discussion of this problem, including historical details, discussion of actual techniques, and a variety of alternative analyses, as well as for a brief review of authorship studies since 1969, see Mosteller and Wallace (1984).

Table 5 Word rates for high-frequency words (rates per 1,000 words)

	*Hamilton (94,000)**	*Madison (114,000)**	*Jay (5,000)**	*Joyce (Ulysses) (260,000)**
The	91	94	67	57
Of	65	58	44	30
To	41	35	36	18
And	25	28	45	28
In	24	23	21	19
A	23	20	14	25
Be	20	16	19	3
That	15	14	20	12
It	14	13	17	9

*The number of words of text counted to determine rates.

Sources: Hanley (1937); Mosteller and Wallace (1984).

PROBLEMS

1. Why can't the authorship of the disputed papers be determined by literary style or political philosophy?
2. a. What is a discriminator?
   b. Distinguish at least two categories of discriminators.
   c. Why is *by* a good discriminator? (Refer to Table 2.)
3. What is a "noncontextual word"?
4. Why do the authors use word frequency per thousand words instead of just the number of occurrences?
5. Refer to Table 1. In how many of the Hamilton papers studied does the word *commonly* appear at least once?
6. Refer to Table 2. In what percentage of the Madison papers studied does *from* occur 3–7 times per 1,000 words? (Note: The interval 3–7 uses the authors' convention on intervals.)
7. Consider the "balls in an urn" model. How many colors of balls would we need to extend the model to the simultaneous study of five words? Of n words?
8. Consider Figure 1. True or false: More than ⅓ of the Hamilton papers studied use *by* 3–7 times per 1,000 words.
9. Study Figure 2. Does the graph for the disputed papers look more like the graph for the Hamilton or the Madison papers?
10. Consider Table 3. Looking at group B, which word would you say was the best Hamilton/Madison discriminator? What was your word-selection criterion? Answer the same questions for group D.
11. Table 1 shows the relative frequency of *war*. Why doesn't *war* appear in Table 3?

REFERENCES

Miles L. Hanley. 1937. *Word Index to James Joyce's "Ulysses."* Madison, Wis.: University of Wisconsin.

F. Mosteller and D. L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist.* Reading, Mass.: Addison-Wesley.

F. Mosteller and D. L. Wallace. 1984. *Applied Bayesian and Classical Inference: The Case of "The Federalist Papers"* (2nd ed. of *Inference and Disputed Authorship: The Federalist).* New York: Springer-Verlag.