A Look at "Defoe's Contributions to *Robert Drury's Journal*: A Stylometric Analysis," By Irving N. Rothman et al.—Are the Results Valid?

Joseph Rudman

Abstract: This paper is a reply to an article written by Irving N. Rothman, Rakesh Verma, Thomas M. Woodell, and Blake Whitaker—"Defoe's Contribution to *Robert Drury's* Journal: A Stylometric Analysis" (2017). That study claimed to support the consensus of traditional attribution studies that *Madagascar; or, Robert Drury's Journal* (1729) is a collaborative work to which Defoe contributed. This paper points out the many flaws of the Rothman group's attribution study—flaws not only in the non-traditional authorship attribution experimental plan but also in the eighteenth-century literary and editorial production aspects of their paper. Rothman et al.'s work was based on Stieg Hargevik's non-traditional authorship study of *Memoirs of an English Officer* which in turn was based on Alver Ellegård's non-traditional authorship work on the *Junius Letters*. This paper also explicates the errors carried over by the Rothman group from the Hargevik and Ellegård studies. The conclusion of this paper is that the Rothman group's results are not valid.

Keywords: Defoe, Authorship Attribution, Statistics, Stylistics, Accountability

> And even if some studies have proved faulty, the vigorous discussion of their shortcomings is a resource for those who follow. (Craig 287)

IN APRIL 2017, I was asked by *The Scriblerian* to review an article titled, "Defoe's Contribution to *Robert Drury's Journal*: A Stylometric Analysis" written by Irving N. Rothman, Rakesh Verma, Thomas M. Woodell, and Blake Whitaker and published in the Festschrift for Jim Springer Bork, *An Expanding Universe*. The article is in the area of my ongoing studies—non-traditional authorship attribution (non-traditional

meaning making use of statistics, and stylistics).[1] The review was published in 2018 ("Review"). I concluded the review by recommending that a proper "ripost" to the Rothman et al. study be undertaken.[2] The more time that went by without a ripost, the more I felt that I myself should take on this necessary but personally unpleasant task, since their article has the potential to misguide scholars and researchers not knowledgeable in the nuances or even the basics of non-traditional authorship studies. This task was made more complicated by the death in 2019 of the principal investigator, Irving N. Rothman. Rothman was a pre-eminent scholar and writer—especially knowledgeable and widely published in the area of eighteenth-century literature and all things Daniel Defoe. I had started to correspond with Rothman about their paper but had to shift the correspondence to Rakesh Verma who graciously, promptly, and completely answered my questions. I would ask the reader to bear in mind while reading the following article that Irving Rothman is not able to defend his experimental plan. [3]

I. Overview

> When I have had to criticize particular methods or scholars it has not been without respect for the courage with which they have addressed themselves to significant problems in a collective enterprise in which failure has often been as instructive as success. (Love 13)

> The task will be approached with such modesty as we can muster, for nothing is more instructive in surveying the errors of others than the salubrious suspicion that we ourselves are likewise fallible. (Ashley 8)

As the title of the Rothman et al. article indicates, they attempted to establish what contributions Defoe may have made to *Madagascar; or Robert Drury's Journal* (1729). The first third of the article presented a thoroughly researched and well written overview of the traditional authorship attribution studies that look at Defoe's involvement with the production of the *Journal*. The Rothman group reported that the consensus of these studies is that the *Journal* is a collaboration. They also reported that their stylistic study comes to the same conclusion—the *Journal* is a collaboration, with sections attributable to Defoe. Unfortunately, there is no consensus in the literature as to what type of collaboration it is, who the collaborators were, or what parts are collaborative.[4] This first third of their paper provides much needed background for what followed in their article—a non-traditional authorship attribution study of the *Journal* using Stieg Hargevik's experimental design and techniques based on his 1972 dissertation. Rothman's multi-disciplinary group at the University of Houston

consisted of an eighteenth-century literary scholar and Defoe specialist, Rothman; a computer scientist, Verma; a linguist, Woodell; and a graduate student in the English Department, Whitaker. The Rothman group's article and the Hargevik book are difficult reads, both because of their technical aspects and their many errors. Appendix A of this paper looks at what may be considered incidental errors—errors introduced or not detected in the production process by the authors, the reviewers, or the copy editors. Some of these errors have the potential to undermine the correct creation of the various corpora used in the Rothman group's experiment.

What follows is a discussion of several problems that bring into question the validity of the entire study. Problems of omission and commission are explicated, and this includes the carrying over of the Hargevik errors.

II. Synopsis of the Hargevik Techniques

> So even if my shortcomings are many and my method is unacceptable to some readers, I think it is high time this nest was stirred. (Hargevik I, 2)

Stieg Hargevik set out to determine if Daniel Defoe was the author of the 1728 tract *The Memoirs of an English Officer*. Hargevik's study was influenced by Alvar Ellegård's 1962 work, *A Statistical Method for Determining Authorship: The Junius Letters*. Without going into great detail, the following (distilled from Hargevik 21-32) makes up the Hargevik experimental plan:

1. Select and gather the Defoe Corpus: A one-million-word sample using first editions if possible. (Corpus one)
2. Select and gather the non-Defoe control corpus: A one-million-word sample of writings not by Defoe, using first editions if possible. (Corpus two)
3. Obtain *Memoirs of an English Officer* with a text length of about 75,252 words. (Corpus three)
4. Compile Defoe's favored words and phrases. (From corpus one)
5. Compile rare Defoe words and phrases. (From corpus one)
6. Search for and list all favored and rare items. (In the three corpora)
7. Analyze the data and determine if the *Memoirs* is by Defoe.

It is interesting (and telling) to note that aside from some earlier Rothman work (Rothman, "Stylometric Study" and "Defoe De-Attribution Scrutinized") and the current work by the Rothman group, the only other non-traditional authorship attribution study using the Hargevik criteria was done by Richard Newsome on the

continuation of *Roxana*.[5] Newsome found the method wanting:

> But in attempting to replicate them [Hargevik's results] I find that they do not provide an answer to the question of whether an unsigned work was written by Defoe or not. (5)

The results of (4) and (5) of the Hargevik experimental plan make up the bulk of the 709 basic words and phrases (and permutations) that he determined to be unique to Defoe's style. Hargevik carried out this phase of the study in three ways: (1) he compiled all words that begin with the letter *i* and the prefixes *dis-* and *ex-* (because he reasoned that these words are rare in Anglo-Saxon vocabulary); (2) he gathered marker words that other stylistic studies used (e.g., those of Ellegård and of Mosteller and Wallace); (3) however, Hargevik compiled the bulk of his stylistic markers by carefully reading the two million words in his corpora in the hope of finding items that Defoe liked and disliked to use (Hargevik 31-32). Hargevik told us that "[a]fter memorizing the preliminary testing list … I re-read all the 267 text portions forming the two million [word]-samples, recording as carefully as possible the occurrences of the different items in the testing list" (36). This "testing list" consisted of 709 items and their permutations each of which could be either (1) individual words, (2) whole phrases, (3) collocations, (4) orthographical oddities in spelling, or (5) frequently occurring phrases in foreign languages. Referring to item (6) in the experimental plan, Hargevik states:

> The result of this operation would, at best, be (1) a list of items more rarely used by Defoe than by most of the writers in the comparative sample, and (2) a list of items which could be regarded as characteristic of Defoe, because they occur more frequently in texts by him than in texts by the majority of contemporary writers. (24)

Another major Hargevik carryover problem I would like to point out is that using only his memory, he read through two million words noting the occurrences of the items from the preliminary testing list. Think about this—he read one million words of Defoe and one million words of non-Defoe and determined Defoe's favorite expressions! Hargevik knew he had a problem: "It would be unwise and presumptuous to pretend that all the occurrences of the items concerned were listed, for it is quite obvious they were not" (36). He was not even able to calculate a systematic error: "At various times I checked my own performance in order to estimate the rate of occurrences missed out. This attempt was balked, however, as my efficiency varied at different times of the day" (36).[6] Hargevik justified the continuation and publication of his study by this quote from Elegård: "To a large extent, however, the mistakes cancel each other out by affecting both plus and minus expressions. The consistency of the results is a guarantee that the occurrences missed through inadvertence have not seriously affected the classification" (qtd. in Hargevik 36). A computer was available to Hargevik to do this sorting and counting but he did not use it. He determined that he could get "similar results but at a smaller cost and with less labor" without a

computer (31). The Rothman group should have compiled their own list of plus and minus Defoe words not from the two-million-word Hargevik corpora but from valid corpora of their own construction. The Rothman group chose to use Hargevik's fatally flawed corpora and style markers rather than redo the initial compilation. This problem alone is enough to question the validity of the entire Rothman group undertaking.

III. The Primary Data —the Input Texts

> 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 28)

Keep this quotation in mind as you read the rest of this paper. It advocates deception by exclusion—something that Hargevik does much too often.

The first major problem of the Rothman group's study is the Hargevik selection and subsequent Rothman group adoption of the two input corpora: the one-million-word Defoe corpus and the one-million-word non-Defoe corpus. It is crucially important that all texts in non-traditional authorship attribution studies be of absolutely certain authorship. Hargevik used ten anonymous selections in his one-million-word non-Defoe control group. This is ~75,000 words (7.75%). He knew this but chose to ignore it.[7] And more damning, as Hargevik admitted, "it is of course also possible that one or two of these anonymous texts were written by Defoe himself" (27). Hargevik was aware that some of the Defoe texts were of questionable authorship: "Defoe's production is vast, and there is a great deal of uncertainty as to the authorship of several texts ascribed to him" (22). But again he did nothing about this crucial problem. Hargevik is guilty of cherry picking his samples and only from the low-hanging branches.[8] In his one-million-word Defoe sample (90 selections) my count is that ~467,000 words (46.7%) were de-attributed by Furbank and Owens—their count for this is 449,500 words (464). Obviously Hargevik did not know this. However, any study done since Furbank and Owens must delete these questionable texts from the Defoe sample before style markers are selected—even if the practitioner disagrees with the de-attribution. Rothman was aware of this and even lists some of the de-attributed Defoe works in his earlier *PBSA* publications but doesn't do anything about it in the Rothman group paper (Rothman, "Defoe De-Attribution Scrutinized" and "Response"). Furbank and Owens called this "an extraordinary fallacy" (464). As I have noted elsewhere in my scholarship, any study of Defoe's style will be degraded by each inclusion of a text not of certain Defoe authorship (Rudman,

"Unediting" 7). This problem alone is enough to question the validity of the entire Rothman group undertaking.

Another problem with the Hargevik input corpora is that they do not distinguish among works of different genres.[9] The consensus of non-traditional attribution practitioners is that genre trumps authorship—many of an author's style markers that are consistent within a genre are not consistent across genres. The practitioner must stick to the genre of the questioned text. Hargevik recognized this problem but did not control for it. He even was aware of the problems caused by not eliminating sub-genres: "Defoe's texts often contain dialogue, and it is well-known that the spoken language of any period differs from the written language" (22). Hargevik conflated nine genres in his control sample: essays; speeches and debates; sermons; histories; novels; journals and diaries; letters; dictionaries and lexicons; and play reviews. Think about this: Hargevik did not conflate just two genres but at least nine. He equated the style of such disparate genres as dictionary compilation with novel writing—play reviews with sermons. And it is important to know if Hargevik selected and analyzed only one side of a printed debate. Hargevik mixed at least four genres in his one-million-word Defoe sample: essays, histories, novels, and letters. The Rothman group was also aware of the problems caused by mixing genres but like Hargevik chose to ignore it.[10] The genre in the corpora should all be what the Rothman group called "seventeenth- and eighteenth-century terrestrial and sea journeys" (Rothman et al. 95). This problem alone is enough to question the validity of the entire Rothman group undertaking.

The next major problem with the input corpora is chronological overextension. A suggested chronological range for non-traditional authorship studies is ± five years from the date of the questioned work.[11] The tighter the chronological span, the better. This is to control for an author's change of style over time, the zeitgeist style change over time, and the change of a genre's style over time. The Hargevik corpora both span thirty years (1700-1730). His questioned tract, *Memoirs of an English Officer* was published in 1728. The Rothman group's questioned work, *Robert Drury's Journal*, was published in 1729. The chronological range of the two corpora should be from 1723 to 1731 (1731 being the year of Defoe's death). This would eliminate most of the two million words in the Defoe and control corpora. This problem alone is enough to question the validity of the entire Rothman group undertaking.

Two other problems with the two corpora are (1) non-random sampling—Hargevik chose his samples by availability and convenience (non-random sampling has the potential to cause problems in the statistics and to introduce experimental bias); and (2) not always using first editions. Hargevik tried to locate first editions but was not always successful: "It is, unfortunately, possible that where later editions were used the results of the investigation were affected" (29). The Rothman group did use

first editions for the several tracts they used in their analysis. But they did not tell us where they obtained all of their texts, if they were in electronic form, or if not in electronic form how they were entered. Nor do they report what types of errors and how many errors were introduced in the process.

IV. Unediting, De-Editing, Editing

When preparing a corpus for analysis, it is essential to attend to three elements of the process: unediting, de-editing, and editing. I have previously defined these terms this way:

> **Unediting**—The Process of removing everything that has been added to the author's manuscript over the ages by editors, printers, or other like "commentators;"
> **De-editing—**The removal of any and all "extraneous" text (e.g. quotations, foreign languages) that would interfere with a valid non-traditional attribution study;
> **Editing—**In this context, the encoding, regularizing, and lemmatizing of the text. (Rudman, "Unediting" 6)[12]

However, the Rothman et al. article does not tell us what they unedited, de-edited, or edited in any of their corpora. They do tell us some of what Hargevik excised:

> He eliminates titles of "texts and headings," "simple ranks and titles in direct apposition, e.g.: King William and Lord Galway," but others were counted when they bore special titles or titles that, if eliminated, would distort the narrative image, such as "the Earl of Peterborough, the King of France." He excludes abbreviations except *viz.* and counted pronouns in two words as one—"every thing" or "some body." He omitted "numerals in the names of regents (e.g., Henry the **Eighth**)," "passages in foreign languages," although he retained words in common usage, such as "en passant," and he omitted questions [sic, read quotations] because Defoe seldom used quotation marks. He also eliminated "Proper names of persons and places," although he counted "names of months, festivals, and similar phenomena." **"When in a quandary**," he writes, "**I omitted such passages altogether**." (108; emphasis mine)

And since the Rothman group followed the Hargevik criteria we can assume they followed suit. But in another study, Rothman advocated removing "tagwords" such as "mother" and "father" as speaker designations so as not to distort Defoe's average sentence length. He also advocated removing Latin quotes but not biblical quotes ("Stylometric Study"). We can guess but do not know if Rothman continued removing those items in the Rothman group study. However, a few problems remain: Did Hargevik also excise the word *the* along with *Eighth* in "Henry the Eighth"? The last sentence of the above quote (in bold) actually refers to Hargevik's comments about quotations, not "similar phenomena." Since Hargevik excluded quotations, why did the Rothman group include them? Did the Rothman group eliminate the 666 separate words and phrases of the native Madagascan languages from the texts? Did

the Rothman group eliminate the 27 word paragraph on page 241 of the *Journal* that is in a native language and/or did they eliminate the 39 word translation of the prayer that immediately follows (Drury 1729, 241-242)? Why did the Rothman group not quote Hargevik's comments about his treatment of hyphenation? On this point, Hargevik had written that

> Hyphenation presented a major problem. I followed Yule's system: "familiar and accepted instances were entered as single nouns … but compound words made up for the nonce … were divided." (Yule, *The Statistical Study* …, p. 125) Needless to say, consistency is feasible only with immense labour. (29)

I would also like to add something else Hargevik said that the Rothman group left out. When discussing foreign languages he writes that "in these cases, the choice between inclusion and exclusion was of necessity very subjective" (29). Hargevik's statement that he "excluded items which appeared to be over-represented in certain texts and thus caused disproportion between the two million-samples" (36) is obviously subjective. But worse, he did not tell us what they are—making it impossible to replicate his study.

As we have seen, Hargevik was not always successful in obtaining first editions. And he realizes the problems this can cause: "Defoe's language has been changed in a most arbitrary way by certain editors … e.g., the word 'further' occurs twice in the first edition and twenty seven times in Aitken's edition" (30). The Rothman group talked about using "The Stoke Newington Daniel Defoe Edition[s] published by AMS Press" (111) based on first editions but did not tell us what they did about changes made by the editors. We can assume that both Hargevik and the Rothman group excised catchwords and signatures—but they were silent on this.

Again, it is of vital importance that we know the exact makeup of the final input text—both studies are based on the number of words in the texts and the number of words in their study blocks. Both Hargevik and the Rothman group knew this but did not let us know exactly what they excised. But they do let us know how difficult and subjective these choices can be. This problem alone is enough to question the validity of the entire Rothman group undertaking.

V. The Rothman Group Experimental Plan

The Rothman group set out to determine if there were sections of the *Journal* that could be attributed to Defoe. Their "procedure conceptualized an analysis of four types of text in *Robert Drury's Journal* as the basis for an assessment of authorship" (107). These four types of text are listed below in Appendix A.

To make sure that I understood how closely the Rothman group followed the Hargevik experimental plan, I asked Verma to confirm the following statements:

1. The study used the same exact one-million-word sample of Defoe's works that Hargevik used.
2. The study used the exact one-million-word sample of the control group.
3. The study used the same Defoe 709 basic words and phrases Hargevik identified.
4. The study used the same Defoe rare words and phrases that Hargevik identified.

Verma confirmed all of the statements, adding that Rothman "wrote to, and even visited, several libraries in the USA and Britain to get hold of the exact editions that Hargevik used" (Verma 2019). This shows to what extraordinary lengths Rothman went to so that he would exactly duplicate the Hargevik criteria. This highlights the fact that Rothman was an exemplary traditional scholar.

The Rothman group then basically followed the Hargevik experimental plan but changed the text from *Memoirs of an English Officer* to that of *Robert Drury's Journal*. Rather than the four selections, the entire *Journal* should have been subject to analysis; the entire text must be subjected to analysis to avoid experimenter's bias. The Rothman group might have avoided this problem by using David Kaufer's Docuscope techniques that use over 40 million English language patterns that are classified into over 100 rhetorical functions that found collaboration in the Federalist papers (Collins et al.) or other techniques such as Eder's rolling stylometry that look for interpolations in texts by breaking the text into equal and overlapping blocks for analysis. By pre-selecting sections, an experimenter's bias was introduced. This problem alone is enough to question the validity of the entire Rothman group undertaking.

VI. Replication and Duplication

In stylometric analysis, as for any other experimental method, a study's results must be replicable if they are to be considered valid. As I have previously explained, replication "means to follow the experimental plan of the original study in every detail without the slightest deviation"; this is distinct from duplication, which "means to reproduce the results using a different experimental plan, such as different style markers, different statistical tests, different control groups" (Rudman, "Shakespeare's Canon" 311). Neither the Hargevik nor the Rothman et al. study can be replicated. We do not know the input data (the texts). If Rothman were still alive, I am sure that all questions about the study would have been answered. I asked Verma for a few pages of the "log" that they referenced in their paper. He sent a file that, "contains all

the matches using the Monoconc software for the Group 4 plus queries of Hargevik on the 11K word extract from Drury's Journal" (Verma 2017). And Verma is willing to answer other questions, which shows the group's willingness to be as transparent as possible. We do not know enough to duplicate the study. However, the problems of creating a valid Defoe sample and a valid control sample are (as we have seen) indomitable. There is a reason that non-traditional authorship attribution scholars in the main do not tackle the Daniel Defoe canon.

VII. Conclusion

The conclusion of the Rothman et al. article was that Defoe authored some sections of the *Journal* but not others. These sections are specified in figure 1. Defoe is identified as the author of two sections: selection two—the 8,889 word passages on religion; and selection three—the 2,965 word Drury's second voyage. The group identified two other sections as not by Defoe: selection one—the 11,254 word initial narrative—and selection three—the 4,917 word speeches or stories. They concluded that the *Journal* should remain in the Defoe canon, "with the understanding of the limitations of Defoe's authorship" (114). The conclusion of this paper is that the results of the Rothman et al. article are not valid and are not to be believed.[13] Many problems may result from non-vetted articles published in a well-respected venue: the results may be incorporated into an author's canon, and the techniques and methodologies (although fatally flawed) may be incorporated into other studies. The following quote from Hargevik is telling:

> Mistakes may breed mistakes if one text is accepted on too loose grounds as written by Defoe and then other texts are then assigned to Defoe on the basis of the first assignment. It appears to be necessary … to establish methods of defining authorship which are as unaffected by human prejudice and subjective thinking as possible. (4)

This is why I felt a pointed critique was in order. By following Hargevik's choice of corpora and his choice of marker words, the Rothman group's study was doomed from the outset. I write this essay in part to warn Defoe scholars to ignore the results of these studies and to warn non-traditional attribution practitioners to use more modern techniques, letting Hargevik's work take its place as a flawed historical step on the road to acceptable practices.

I would be remiss if I did not compliment the work of Woodell (a linguist) and Verma (a computer scientist). They did an admirable job of taking Hargevik's 709 words and permutations into "more than 7,000 terms to query" (Rothman et al. 104) and analyzing the staggering mountains of data. It is understandable that they would accept Rothman's lead on the corpora construction and other areas of the Hargevik criteria. Rothman was aware of the many pitfalls facing practitioners of non-

traditional authorship attribution studies that are listed above. He cited two articles that discussed the problems (Rudman, "State of Authorship Attribution Studies" and "Unediting") but chose to ignore the caveats.

Appendix A: Editing and Production Errors

Many if not most of the production problems that appear in the Rothman at al. article can perhaps be attributed to the state of the AMS press in its waning years —the volume containing the Rothman group's article was one of the last publications of the press before bankruptcy and liquidation. The AMS press sat on some of the submissions for this volume for almost ten years. The Rothman group article "was submitted to the Festschrift in 2007-08, I believe" (Verma 2019). There was little or no anonymous peer reviewing of the articles. There was little or no copyediting by the press near its end. According to Verma, "As far as I know, there were no interactions with a copy editor" (2019). The startling number of typographical and other minor errors in the essay tend to confirm that it received very little editorial attention. The authors and guest editors were left in the dark for a good portion of the publication process.[14] But it is important to keep in mind as you read this paper that the Rothman group had not seen their initial submission for over ten years and that it was published before they had a chance to correct or modify it. They had no chance to read any reviewers' comments and suggestions. Much of the turmoil in this esteemed press was caused by the declining health of its founder and operator, Gabriel Hornstein. His contributions to eighteenth-century studies cannot be overstated. Sadly, he passed away on February 17, 2017—a week before the publication of *An Expanding Universe*. This does not completely exonerate the editors or the authors but explains how the undetected errors could slip through. However, the other essays in the volume do not evidence the kind of errors found in the Rothman et al. article.

The first problems to be pointed out have to do with the presentation of the four sections of the *Journal* that the Rothman group selected to be tested to see if any or all of them were written by Defoe—problems with identification and pagination. The four times that these selections are printed in the paper are listed below in figs. 1 through 4.

Problem: Inconsistency in listing the content of the four selections. Note that the Rothman group transposes selection 3 and 4 in figs. 1 and 2. They then go back to the original order of fig. 1 in fig. 3. However, they again transpose selection 3 and 4 in fig. 1 and fig. 4 (observe that they also change from Arabic numerals to letters in fig. 4).

Problem: Incorrect inclusive page and line numbers for the four selections (in fig. 3). In the first selection the page and line numbers are given as 39:1-71:11. The actual numbers are 1:1-56:20. It did not take long to determine that the 39:1-71:11 numbers are from a different edition—the 1890 edition that was edited and expurgated by Pasfield Oliver even though the Rothman group stated that these numbers are from the 1729 edition (112). There are two sections that make up their second selection. They got the first of the two correct. The second of the two they gave as 230:8-88. The numbers should be 230:8-256:9. The '88' is a mystery—it is not the page number, the number of pages, nor the number of lines. The numbers of the third selection are correct. The fourth selection has three sections. The first two are correct. The third one is given as 105:17-105:37. The actual numbers are 105:17-115:25. There are no pages of the 1729 edition of the *Journal* with 37 lines.

Problem: There are other (perhaps inconsequential) irregularities and inconsistencies. In fig. 1 under Selection 1 and Selection 2, "a" should be "an." In fig.3 under 2 note that there are no quotation marks before "He" or after "them." Also in fig. 3 under 4 note that the phrase "enter'd the Country without Opposition" is extraneous and should not be there. In fig. 4 under C, "words" should be singular.

([Selection] 1) **a** [sic] 11,254-word introduction to Drury's experience extending from the beginning of the text;
([Selection] 2) **a** [sic] 8,889-word compilation that focuses on Drury's assessment of the natives' religious tenets and one example of fraudulent religious rites;
([Selection] 3) a 2,965-word selection from Drury's experience as a freed man at his return to Madagascar in a second voyage; and
([Selection] 4) a 4,917-word compilation of stories by others, which may appear to be in a different voice from that of the narrator.

Fig. 1 —From Rothman et al. 1. Note: I changed Rothman et al.'s word "corpus" in this figure to "selection" in order to avoid confusion with their other uses of the word corpus.

1. An analysis of the beginning narrative passage.
2. A compilation of several distinct passages on religious affairs….
3. Passages in which characters told their own stories to Drury, in a first-person syntax independent of the first-person narration of the *Journal* —Drury's voice —presumably understood to be the words of a reliable narrator.
4. Drury's return to Madagascar in a narrative appended to the original narrative….

Fig. 2 —From Rothman et al. 112

1. **pp. 39:1 —71:11** [sic] —"My design in….till I was swell'd with water."
2. pp. 181:24 —194:10 —**He** [sic] then desir'd me….as I did not affront the**m**; also, **pp. 230:8 —88** [sic]-- "Here is no one….to make him keep the secret."
3. pp. 444:26 —456:26 —"When I was a boy….may seem doubtful."
4. pp. 16:28 —24:5 —"I am an *English-man*….our Numbers are increas'd"; also, pp. 86:3 —90:13 —"That *Dean Woozington*, the king…**enter'd the Country without Opposition** [sic]; …their respective homes."; also, pp. 105:17 —**105:37** [sic] —"Now it happened….alive off the Island."

Fig. 3 —From Rothman et al. 113

A. 11,254-word, initial narrative….
B. 8,889-word, passages on religion….
C. 4,917-**words** [sic], speeches or stories….
D. 2,965-word, Drury's 2nd voyage….

Fig. 4 —From Rothman et al. 113

However, there are serious problems with these four selections. The Rothman group re-used the "Sam's Story's" seven pages from selection one. They were included in selection four (see fig. 1). This duplication is not obvious from reading the Rothman group article because of the errors in the page numbers. It only becomes obvious when you look at the correct page numbers and read the selections. We do not know what that does to the results for selection one but it does cast a cloud over the results. The Rothman group also tainted selection four by including some explanatory material before their third part of selection four, part three (see Selection 4 of fig. 1). Furthermore, this selection is by the narrator (which is not a "different voice" as advertised). This also taints the selection.

Of course, most of the problems highlighted in this appendix do not invalidate the results of their study—it is the rare scholar who has not seen an error creep into a published work (and I will not cast the first stone). However, the errors are frustrating as the reader tries to understand the authors' methods. And these many errors do give rise to the specter of other undetected errors in the reporting of the experimental plan, the analysis, and the results.

Carnegie Mellon University

Notes

[1] See Holmes for a good basic overview. For two encyclopedia entries that give an overview of

the topic, See Rudman, "Authorship Attribution" and "Stylometrics." Also, see Rudman, "State of Authorship Attribution Studies" and "State of Non-Traditional Authorship Attribution" for two more comprehensive articles.

[2] The term *ripost* is used in the non-traditional authorship attribution community to refer to the totality of a multi-faceted pointed critique of a non-traditional attribution experiment.

[3] The scope of this paper does not allow for an up-to-date general survey of the field of non-traditional authorship attribution studies. Nor does it allow for a complete presentation of a proper experimental plan for a valid way to do a non-traditional study of the Defoe canon. At the time of publication, this Zotero group offers a searchable bibliography of about 4,000 entries of non-traditional studies, as well as a list of suggested readings for those new to the field. For two truncated exempla of Defoe experimental plans see Rudman, "Non-Traditional Authorship" and "Unediting."

[4] See Rudman, "Shakespeare's Canon" for a more complete treatment of the collaboration concept.

[5] The Furbank and Owens ("Stylometry and the Defoe Canon") vs Rothman ("Defoe De-Attribution Scrutinized," "Response") give and take that took place in *The Papers of the Bibliographical Society of America* covered some of the same problems with the work of Ellegård and Hargevik that are explicated in this paper.

[6] See Beers for a detailed treatment of Hargevik's errors.

[7] I compiled the numbers and percentages about the texts from Hargevik, Appendix I (pp. I-VIII) and Appendix II (pp. IX-XVIII).

[8] See Rudman, "Cherry Picking" for a more complete treatment of cherry picking.

[9] Genre is an important variable in non-traditional authorship attribution studies. It must be controlled for. If genre cannot be controlled for, the practitioner must calculate a systematic error and fold it into the final result. The studies that come closest to questioning this are ones where the genres follow similar linguistic rules, such as tragicomedy and comedy. No one questions the need to separate sonnets from essays. Another point to keep in mind is that genre separation must also include separating sub-genres—e.g., poetry within a novel, a song within a drama.

[10] The Rothman group was aware of my 2005 paper that had the following comment on genre:

> It has been shown empirically that style-markers vary significantly over different genres (Karlgren and Cutting) (Stamatatos, Fakotakis and Kokkinakis). Burrows has shown that stylistic differences are greater among the various genres written by the same author than they are between different authors writing in the same genre. He has a telling graph that "shows a complex pattern in which genre transcends authorship" (Burrows 101-102). Binongo reinforces this: "When the essays and plays are brought together into one picture…the differences in genre predominate over other factors (Binongo 114).

[11] This suggested range was arrived at by looking at all of the studies that determined a stylochronological change—e.g., Boyd 7, Evans 128, Bramer and Miltos, Stamou, Hoover, Pennebaker and Stone.

[12] See Rudman, "Unediting" and "Shakespeare's Canon" for a more complete treatment.
[13] Note that this paper does not discuss Hargevik's "distinctiveness groups" or the way his statistics determine authorship. This essay's focus is on the validity of the Rothman team's results. If the input data (the texts) are invalid, the results of any tests would be invalid.
[14] This information was garnered from conversations with James E. May, who published a chapter in the volume, and Kevin L. Cope, one of the volume editors.

Works Cited

Ashley, Leonard R.N. *Authorship and Evidence*. Librairie Droz, 1968.

Beers, Yardly. *Introduction to the Theory of Errors.* Addison Wesley, 1958.

Binongo, J.N.G. "Stylometry and Implementation by Principal Component Analysis." 2000. University of Ulster, Ph.D. Dissertation.

Boyd, Ryan. "Mental Profile Mapping: A Psychological Single-Candidate Authorship Attribution Method." *PLoS ONE* vol.13, no. 7, 2018, pp. e0200588. doi:10.1371/journal.pone.0200588.

Bramer, Max, and Miltos Petridis. "Stylochronomerty: Timeline Prediction in Stylometric Analysis." *Research and Development in Intelligent Systems XXXII*, edited by Carmen Klaussner and Carl Vogel, Springer, 2015, pp. 91-106.

Burrows, John. "Not Unless You Ask Nicely: The Interpretive Nexus Between Analysis and Information." *Literary and Linguistic Computing*, vol. 7, 1992, pp. 91-110.

Carleton, George. *The Memoirs of an English Officer: who serv'd on the Dutch war in 1672. To the peace of Utrecht, in 1713.* London, 1728.

Collins, Jeff, et al. "Detecting Collaborations in Text: Comparing the Authors' Rhetorical Language Choices in The Federalist Papers." *Computers and the Humanities*, vol. 38, no. 1, 2004, pp. 15-36.

Craig, Hugh. "Stylistic Analysis and Authorship Studies." *A Companion to Digital Humanities,* edited by Susan Schreibman et al., Blackwell, 2004, pp. 273-288.

Drury, Robert. *Madagascar; Or, Robert Drury's Journal, During Fifteen Years' Captivity on That Island; And a Further Description of Madagascar by the Abbé Alexis Rochon,* edited by Pasfield Oliver. Macmillan Co., 1890.

---. *Madagascar: Or, Robert Drury's Journal, During Fifteen Years Captivity on That Island*. London, 1729.

Eder, Maciej. "Rolling Stylometry." *Digital Scholarship in the Humanities*, vol. 31, no. 3, 2016, pp. 457-469.

Ellegård, Alvar. *A Statistical Method for Determining Authorship: The Junius Letters*. Acta Universitatis Gothoburgensis, 1962.

Evans, Melanie A. "Style and Chronology: A Stylometric Investigation of Aphra Behn's Dramatic Style and the Dating of *The Young King*." *Language and Literature,* vol. 27, no. 2, 2018, pp. 103-132.

Furbank, Philip N. and W. R. Owens. *Defoe De-Attribution: A Critique of J. R. Moore's* Checklist. The Hambledon Press, 1994.

Hargevik, Stieg. *The Disputed Assignment of* Memoirs of an English Officer *to Daniel Defoe.* Vol. 1, Almqvist & Wiksell, 1974.

Holmes, David. "The Analysis of Literary Style—A Review." *The Journal of the Royal Statistical Society* (Series A [General]), vol. 148, no. 4, 1985, pp. 328-341.

Hoover, David. "Corpus Stylistics, Stylometry, and the Styles of Henry James." *Style*, vol. 41, no. 2, 2007, pp. 174-203.

Karlgren, Jussi and Douglas Cutting. "Recognizing Text Genres With Simple Metrics Using Discriminant Analysis." *In Proceedings of COLING 94, The 15th International Conference on Computational Linguistics,* August 5-9, 1994, pp. 1071-1075.

Love, Harold. *Attributing Authorship: An Introduction*. Cambridge University Press, 2002.

Mosteller, Frederick and David L. Wallace. *Applied Bayesian and Classical Inference: The Case of* The Federalist Papers. Springer Verlag, 1984.

Newsome, Richard. "An Investigation into the Authorship of the 1745 Continuation of Defoe's 'Roxana.'" 10 June 1987. Author's collection.

Owens, W. R. and P. N. Furbank. "Stylometry and the Defoe Canon: A Reply to Irving Rothman." *The Papers of the Bibliographical Society of America,* vol. 96, no. 3, 2002, pp. 463-465.

Pennebaker, James, and Lori D. Stone. "Words of Wisdom: Language Use Over the Life Span." *Journal of Personality and Social Psychology*, vol. 35, no. 2, 2003, pp. 291-301.

Rothman, Irving N. "A Response to P. N. Furbank and W. R. Owens." *The Papers of the Bibliographical Society of America,* vol. 96, no. 3, 2002, pp. 465-469.

---. "Defoe De-Attribution Scrutinized under Hargevik Criteria: Applying Stylometrics to the Canon." *The Papers of the Bibliographical Society of America,* vol. 94, no. 3, 2000, pp. 375-398.

---. "A Stylometric Study of the Variant Styles of Daniel Defoe." South Central Society for Eighteenth Century Studies Convention, 25 Feb. 1999, Shreveport, LA. Conference Presentation.

Rothman, Irving N., et al. "Defoe's Contribution to Robert Drury's Journal: A Stylometric Analysis." *An Expanding Universe: The Project of Eighteenth-Century Studies, Essays Commemorating the Career of Jim Springer Borck*, edited by Kevin L. Cope and Cedrick D. Reverend II, AMS, 2017, pp. 93-116.

Rudman, Joseph. Review of 'Defoe's Contribution to *Robert Drury's Journal*: A Stylometric Analysis, by Irving N. Rothman et al. *Scriblerian,* vol. 51, no. 1, 2018, pp. 10-11.
---. "Non-Traditional Authorship Attribution Studies of William Shakespeare's Canon: Some Caveats." *Journal of Early Modern Studies*, vol. 5, 2016, pp. 307-328.
---. "The State of Non-Traditional Authorship Attribution–2012: Some Problems and Solutions." *English Studies*, vol. 93, no. 3, 2012, pp. 259-274.
---. "Stylometrics." *Cambridge Encyclopedia of the Language Sciences*, edited by P. C. Hogan, Cambridge University Press, 2011, pp. 817-819.
---. "Authorship Attribution: Statistical and Computational Methods." *Encyclopedia of Language and Linguistics,* edited by K. Brown, Elsevier, 2006, pp. 611-617.
---. "Unediting, De-Editing, and Editing in Non-Traditional Authorship Attribution Studies: With an Emphasis on the Canon of Daniel Defoe." *The Papers of the Bibliographical Society of America*, vol. 99, no. 1, 2005, pp. 5-36.
---. "Cherry Picking in Non-Traditional Authorship Attribution Studies." *Chance*, vol. 16, no. 2, 2003, pp. 26-32.
---. "Non-Traditional Authorship Attribution Studies in Eighteenth Century Literature: Stylistics Statistics and the Computer." *Jahrbuch für Computerphilologie*, vol. 4. 2002, pp. 151-166.
---. "The State of Authorship Attribution Studies: Some Problems and Solutions." *Computers and the Humanities*, vol. 31, 1998, pp. 351-365.
Stamatatos, E., N. Fakotakis, and G. Kokkinakis. "Text Genre Detection Using Common Word Frequencies." *Proceedings of COLING 2000, The 18th International Conference on Computational Linguistics*, July 31 – August 4 2000, vol. 2, pp. 808-814.
Stamou, Constatina. "Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating." *Literary and Linguistic Computing*, vol. 23, no. 2, 2007, pp. 181-199.
Verma, Rakesh. E-mail to the author. 23 Sept. 2019.
---. E-mail to the author, 18 Aug. 2017.
Yule, G. Udny. *The Statistical Study of Literary Vocabulary.* Cambridge University Press, 1944.