A Statistical Approach to Judicial Authorship: A Case Study of Judge Easterbrook

Kelly Bodwin, Jeffrey S. Rosenthal, and Albert H. Yoon[1]

(August 2012; revised December 2012)

Abstract

Legal scholars have long been interested in the extent to which judges rely on their law clerks in writing judicial opinions. Recent scholarship provides compelling evidence that Supreme Court justices increasingly rely on their clerks. Distinguishing between judge-written and clerk-written opinions is difficult, since this information is typically kept within chambers. The opinions of Judge Frank Easterbrook of the U.S. Court of Appeals provides a unique opportunity: he writes his own opinions, aside from allowing each of his clerks to write a first draft of one opinion (typically towards the end of her clerkship). This article examines linguistic and stylistic elements of Judge Easterbrook's opinions to establish analytically the degree of similarity in his writing style compared to his clerks and fellow 7th Circuit judges. We find that Judge Easterbrook's opinions are clearly identifiable from those of his fellow jurists. By contrast, his clerks' writings are statistically distinct but less identifiable. Our study supports the view that even with close judicial oversight, clerks' writing cannot be made exactly equivalent to their judges. It also provides evidence for the effectiveness of a clerkship experience such as Judge Easterbrook's, in which clerks are given a lengthy opportunity to learn both his stylistic and substantive approach to the law.

[1]University of North Carolina – Chapel Hill, Department of Statistics; University of Toronto, Department of Statistics; University of Toronto Faculty of Law, respectively. Authors are listed in alphabetical order. The authors would like to thank Ben Alarie, Jonathan Cardi, Ian Caines, Helen Levy, Tom Miles, Eric Posner, Richard Posner, and Simon Stern for their helpful comments. We would especially like to thank Frank Easterbrook for his generous cooperation and help with this project. This research was partially supported by NSERC and by the Russell Sage Foundation.

I Introduction

It is commonly known that in the modern era, judges rely on their law clerks to keep up with the demands of opinion writing (Peppers 2006; Ward & Weiden 2006; Lazarus 1998). While jurists have defended this practice (Liptak 2008), it is often criticized under the assumption that the writing and legal thought of a clerk, typically a recent law school graduate, is inferior to that of the judge (Posner 2002). Herein we make no comment on quality of legal analysis in judicial opinions, but focus on the question whether judges and their law clerks possess stylistically distinct writing styles.

Recent scholarship provides persuasive evidence that over time, Supreme Court justices have increasingly relied on their clerks in writing judicial opinions. Early work focused on citation practices (Choi & Gulati 2005), a measurable but questionable metric. More recent efforts examined writing style using common function words, and found that justices' variability increased, both within and across years (Rosenthal & Yoon 2011a, 2011b).

Although a secular, upward trend in writing variability has been clearly established, it is still difficult, if not impossible, to determine the extent to which a specific opinion is judge-written versus clerk-written because the relative contributions of each is typically a closely-guarded secret within chambers. Archival research on opinion-drafts reveals differences imputed to individual clerks on the Supreme Court (Wahlbeck et al 2002), but such opportunities remain rare and still leave unanswered the relative contributions of judge and clerk.

The opinions of Judge Frank Easterbrook, Chief Judge of the 7th Circuit Court of Appeals, provides a unique opportunity to disentangle the relative contributions of judges and their clerks. In contrast to most federal judges (Choi and Gulati 2005), Judge Easterbrook has a long-standing practice of writing all of his opinions, with the exception that he allows each of his clerks to draft a single opinion, typically towards the end of their clerkship year. This non-reliance on clerks produces a writing style that has a consistently low variability, based on common function words, which stands

in stark contrast to any contemporary justice on the Supreme Court (Rosenthal & Yoon 2011b). This institutional practice makes it possible to separately examine each type of opinion. We reasonably expect that clerks have an incentive to write their opinions[2] in a manner as similar to Easterbrook's own as realistically possible. Thus, these opinions provide an extreme but clear benchmark for answering our question: Can a law clerk successfully emulate the writing style of his or her judge?

This paper is organized as follows. In Section II we briefly describe our data, consisting of Judge Easterbrook's opinions (including those drafted by his clerks) together with certain "Imposter" opinions, described below. In Section III we introduce and justify a method of identifying incongruous opinions from among a corpus of judicial opinions. Our work builds on the earlier contributions by statisticians on the issue of authorship (Airoldi et al 2006; Mosteller and Wallace 1964; Ellegard 1962). We develop our method by testing various canonical metrics of authorship identification on known outlier opinions – "Imposter" pieces taken from Easterbrook's judicial contemporaries on the 7th Circuit of the U.S. Court of Appeals, and the corpus consisting of Judge Easterbrook's opinions. As we will show, our methods are remarkably successful in unmasking these Imposter opinions. In Section IV we describe our outlier score, based on the metrics used in Section III.

We report our results in Section V, in which we examine the full set of Easterbrook opinions. For each year we produce a ranked list of texts, specifying which opinions we presume are most likely to be clerk-authored. We also describe our approach, to avoid over-fitting of the data, of using opinions from judicial terms beginning with even-numbered years (e.g., 1990–1991) as training data, and opinions from sessions beginning with odd-numbered years (e.g., 1987–1988) as testing data. For each stage, after coming up with our yearly rankings, Judge Easterbrook generously revealed which opinions his clerks drafted. Comparing this list to the true

[2]We are here referring specifically to opinions credited to Judge Easterbrook, for which he permits his clerks to write the initial draft. Our understanding is that Judge Easterbrook reviews and edits his clerks' drafts, although we do not observe the degree to which he does so for each such opinion.

clerk opinions, we find that our methods identify clerk-authored opinions: not with complete reliability, but significantly better than random guesswork. Section VI discusses our results, including potential implications for the judiciary and statistical approaches to identifying authorship. Section VII concludes.

II Data

We consider the corpus of all court opinions attributed[3] to Easterbrook, from 1984 (his first year on the 7th Circuit Court of Appeals) through 2010.[4]

As described above, Judge Easterbrook writes all of his own opinions, with the exception that he allows his clerks – typically two per term – to each draft a single opinion, typically toward the end of their clerkship year. For obvious reasons, Judge Easterbrook does not publicly reveal the clerk-written opinions. Thus, when embarking on our analysis, we did not know which of the opinions Judge Easterbrook had written and which ones his clerks had drafted.

We divided this corpus into two parts: training and testing. The training data were opinions from terms where the fall was an even-numbered year (e.g., 1986–87). The testing data were opinions from terms where the fall was an odd-numbered year (e.g., 1987–88). Judge Easterbrook generously offered to reveal which opinions his clerks had drafted (with the clear understanding that this information would remain confidential) upon completion of each stage of the analysis. Thus, we first attempted to identify the clerk-written opinions from the even-numbered years (training data). After we had done so, Judge Easterbrook then revealed the true clerk-written opinions for these years. We then used these results from the even-numbered years to try to improve upon our results for the odd-numbered (testing data).

As a test of our methods, we also included for each year in the Easterbrook corpus one "Imposter" opinion, written by one of Easterbrook's judicial contempo-

[3]We exclude *per curiam* opinions in which Judge Easterbrook was part of the panel.

[4]All textual data analyzed in this study comes from Westlaw (www.westlaw.com).

raries on the 7th Circuit in the same year.[5] Our resulting corpus consists of "Easterbrook+Imposters." This inclusion allows us to consider the extent to which methods that successfully identify outside opinions as outliers are also able to identify clerks' writings.

Finally, after completing our initial analysis, by contacting Judge Easterbrook and (with his permission) his former clerks, we were able to amass a list of the actual clerk opinions in the corpus. Although in principle there are two clerk-authored opinions per session (one for each of Easterbrook's clerks), the number of clerk-drafted opinions we identified do not consistently number two per year, for several reasons: (i) in his early judicial tenure (1990 and before), his clerks drafted *per curiam* opinions – we excluded these opinions from our analysis to be consistent with our approach of excluding *per curiam* opinions generally.[6]; (ii) some opinions written by clerks were not published until the following term; and (iii) a few former Easterbrook clerks could not be reached or were unable to recall which opinions they drafted. Nonetheless, we were able to identify 32 (of the 38) expected clerk-written opinions after 1990. We make use of this information in our second-round analysis, described in Section V.3 herein.

III Statistical Measures

In this section we describe our strategy to construct a scoring system under which we distinguish what we believe are clerk-authored opinions. To achieve this, we

[5]The available Imposter opinions are from Judges Bauer (138), Coffey (121), Cudahy (102), Eschbach (40), Evans (73), Fairchild (16), Flaum (180), Kanne (144), Manion (86), Pell (7), Posner (251), Ripple (141), Rovner (92), Swygert (1), Sykes (28), Tinder (10), Williams (45), and Wood (159), and we selected at random from this collection.

[6]We exclude *per curiam* opinions because, by design, the court does not reveal the identity of the authoring judge on the panel, precluding us from knowing which among these opinions Judge Easterbrook actually wrote.

begin by considering several diverse linguistic elements, and developing[7] statistical measures based on each of them. Our statistical strategy works in two stages. First, in this section, we apply these various measures to the Easterbrook+Imposters corpus (for which we already know which opinions come from Imposters), in an effort to determine how useful they are in identifying outlier opinions. Then, in Section IV, we combine these various measures to create a single scoring system, which we then apply to attempt to identify the clerk-authored opinions in the Easterbrook corpus.

III.1 Rare Words

One well-established approach to authorship identification involves word choice (Mosteller and Wallace 1964). Different authors will tend to favor or avoid particular words. Thus, if certain words appear only rarely in the overall Easterbrook corpus, but appear frequently in one particular opinion, this may indicate that the opinion was not written by Easterbrook but rather by one of his clerks (or, in the Easterbrook+Imposters case, by one of the Imposters).

One concern with this approach is topicality. That is, a word may be rare simply because it relates to the narrow topic of the particular case at hand. For example, one of Judge Easterbrook's opinions, Stockman v. LaCroix, 790 F.2d 584 (1986), involved deals with the disputed purchase of a racehorse; as such, it contains many rare words, like "stallion" or "foal," but this in no way indicates a clerk-authored opinion.

To avoid this problem, we define "Rare Words" as those which appear in more than five separate opinions, but appear 100 or fewer times in total.[8] This implies that most of the Rare Words will indeed be stylistic rather than topical.

[7]For all of the software that we have developed and used to download and analyze the opinion text, see: http://probability.ca/easterbrook/README

[8]While these cutoffs are somewhat arbitrary, we found that the results were fairly stable if the cutoffs were modified.

To make statistical use of these Rare Words, we let

$$p \;=\; \frac{\text{total Rare Words in corpus}}{\text{total words in corpus}} \;\doteq\; 0.0485$$

be the total fraction of Rare Words in the corpus. Then, suppose a given "typical" opinion has n_k words total. We would then expect that on average, the number of Rare Words it contains, say X, would be approximately equal to pn_k. In symbols, $\mathbf{E}(X) = pn_k$. Using the Poisson approximation, we can then model X as $X \sim \text{Poisson}(pn_k)$. If the opinion in fact contains a total of y Rare Words, then we can use the Poisson distribution to compute the tail probability $P(X \geq y)$. If this tail probability is far from zero, this means that y is not unexpectedly large, so the opinion does indeed appear to be typical in this sense. However, if this tail probability is close to zero, say[9] less than 0.05, then this suggests the opinion is far from typical and may indeed be clerk-authored.

Examining all of the opinions in the Easterbrook+Imposters corpus, we found that the Poisson tail probability was below the 0.05 cutoff in 8 of the 26 Imposter opinions (30.8%), but in only 63 of the 1630 actual Easterbrook opinions (3.8%). This suggests that the Rare Words approach identifies the non-Easterbrook opinions with some success, and is thus a somewhat promising method.

III.2 Function Words

Another approach is to consider "Function Words", i.e. words such as "the" or "and" which are very common and are not related to any particular topic. Much success has been found through the use of this method in authorship identification, most notably in the case of The Federalist Papers (Mosteller & Wallace, 1964). It has also been applied successfully to issues of clerk authorship on the Supreme Court (Rosenthal & Yoon 2011a, 2011b).

We consider the same 63 function words used in Rosenthal & Yoon, namely: *a, all, also, an, and, any, are, as, at, be, been, but, by, can, do, down, even, for, from,*

[9]This cutoff is also somewhat arbitrary, though quite standard; we again found that modifying it did not overly affect our results.

had, has, have, her, his, if, in, into, is, it, its, may, more, must, no, not, now, of, on, one, only, or, our, so, some, such, than, that, the, their, then, there, things, this, to, up, was, were, what, when, which, who, with, would.

For each function word i and each opinion j in the corpus, let $x_{i,j}$ be the proportion of words in opinion j which are equal to function word i. Let μ_i be the mean of this proportion over all opinions j, and let σ_i^2 be the corresponding variance. Then using the standard normal approximation (Rosenthal & Yoon 2011a), the log-likelihood for the observed proportions $x_{i,j}$ in opinion j is given by the formula

$$\text{loglike}(j) \;=\; \sum_{i=1}^{63} \left(-\frac{1}{2}\log(2\pi\sigma_i) - \frac{(x_{i,j}-\mu_i)^2}{2\sigma_i^2} \right) .$$

The smaller the value of loglike(j), the more surprising are the observed fractions $x_{i,j}$, and thus the more likely it is that opinion j is an outlier, i.e. was written by a clerk or an Imposter. This allows us to *rank* all of the opinions in a corpus, in terms of how likely they are to be an outlier.

We used this approach on the Easterbrook+Imposters corpus, to rank all the opinions within each year, in terms of how likely they are to be an outlier. The resulting ranks were as follows (where 1 is the most likely to be an outlier):
Table 1 shows that in most years, Imposters opinions are ranked in the upper third – and in fact, four opinions are ranked in the top three. This suggests that the Function Word approach provides another promising method.

To determine the statistical *significance* of this success, we simulated *random* rankings (where each ranking is equally likely to be any number from 1 to the number of opinions in that year) as a comparator method, 10,000 different times. We found that 1.58% of the time the random ranking performed better (i.e., had higher rankings in the majority of the years), and 3.67% of the time it was a tie, and the remaining 94.75% of the time the Function Word approach performed better. This indicates that the Function Words are indeed working well on the Easterbrook+Imposters corpus, far better than random chance.

Session	Imposter Rank	Total Opinions	Session	Imposter Rank	Total Opinions
1984-85	1	20	1997-98	23	76
1985-86	5	69	1998-99	42	66
1986-87	27	61	1999-2000	13	66
1987-88	44	70	2000-01	1	61
1988-89	31	50	2001-02	21	62
1989-90	35	70	2002-03	10	60
1990-91	42	63	2003-04	2	73
1991-92	37	68	2004-05	18	78
1992-93	10	60	2005-06	17	68
1993-94	19	78	2006-07	62	69
1994-95	26	67	2007-08	13	63
1995-96	26	61	2008-09	2	59
1996-97	16	62	2009-10	42	57

Table 1: Ranking of Imposter opinions by Function Word log-likelihood, showing generally good (low-number) rankings.

III.3 Punctuation

We next examine punctuation. Instead of the above normal approximation / log-likelihood model (which implicitly assumes of independence between the Function words), we assume a fixed number of places where punctuation can go in each opinion, corresponding to a Chi-Squared distribution. Specifically, we define $O_{j,i}$ to be the number of times that punctuation mark i appears in the j-th opinion. Then we assume that on average the punctuation marks are equally divided among the opinions, so that the expected count of punctuation mark i in opinion j is given by:

$$E_{j,i} \;=\; \left(\frac{1}{c}\sum_{k=1}^{c} O_{j,k}\right)\left(\frac{1}{r}\sum_{k=1}^{r} O_{k,i}\right)\,,$$

where c is the number of different punctuation marks, and r is the total number of opinions in the corpus.

To find outliers, we use the contribution of each opinion to the full Chi-Squared statistic. This is given by the equation

$$\text{Contribution}(j) \;=\; \sum_{i=1}^{c} \frac{(O_{j,i} - Ej,i)^2}{E_{j,i}}\,.$$

The higher the value of Contribution(j), the more the observed punctuation counts differ from their expected values, and thus the more likely that opinion j is an outlier.

Figure 1 compares boxplots of the contributions from Easterbrook opinions (left) and from Imposter opinions (right). We see from Figure 1 that the Imposter opinions tend to have larger Chi-Squared contributions. Indeed, a one-sided t-test confirms that Imposter opinions are larger than Easterbrook opinions with confidence > 99.9 percent. This illustrates that this measure of punctuation provides another promising method of identifying outliers from a corpus.

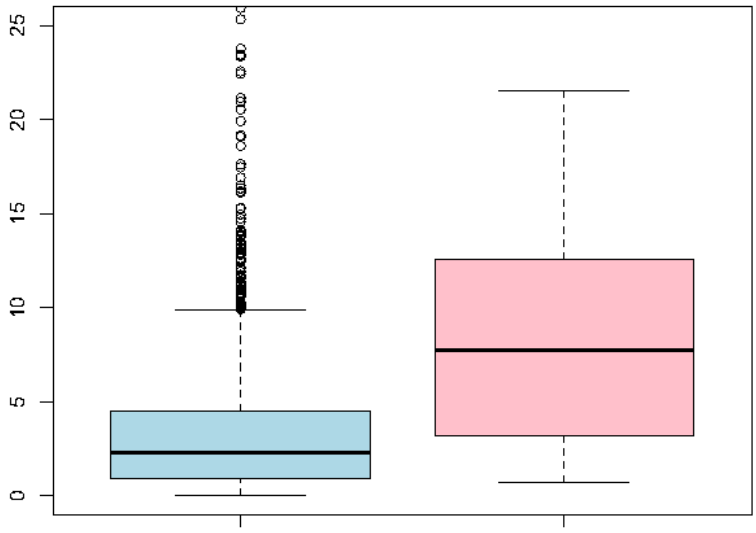

Figure 1: Boxplots of the Punctuation Chi-Squared contribution values of the Easterbrook (left) and Imposter (right) opinions, showing that the Imposter opinions generally have larger values.

We can also measure this in terms of rankings, similar to before. The results by ranking are presented in Table 2. That table indicates that in nearly every term, Imposters ranked in the top half, and for half the terms (13 out of 26), Imposters ranked in the top 10%. This provides further evidence that punctuation counts are able to identify the Imposter opinions with good success.

Session	Imposter Rank	Total Opinions	Session	Imposter Rank	Total Opinions
1984-85	5	20	1997-98	6	76
1985-86	4	69	1998-99	18	66
1986-87	4	61	1999-2000	25	66
1987-88	15	70	2000-01	2	61
1988-89	39	50	2001-02	30	62
1989-90	6	70	2002-03	4	60
1990-91	15	63	2003-04	6	73
1991-92	44	68	2004-05	1	78
1992-93	4	60	2005-06	23	68
1993-94	41	78	2006-07	1	69
1994-95	34	67	2007-08	1	63
1995-96	37	61	2008-09	23	59
1996-97	4	62	2009-10	1	57

Table 2: Ranking of Imposter opinions by their Punctuation Chi-Squared contribution, showing generally good (low-number) rankings.

III.4 Principal Component Analysis of Common Words

Our fourth, and final, metric uses Principal Component Analysis (PCA), a commonly used method in authorship attribution (see e.g. Burrows & Craig 2001). It identifies the components (i.e., the axes or directions or linear combinations) that account for the most variation in a set of values. The advantage of PCA is that we may explain most of the data using only a few components, so that a complicated high-dimensional dataset may be fairly well described by only a few values.

We begin with a list of the 100 most commonly-appearing words in the writings of Easterbrook and his contemporaries. Each of these words is treated as a spatial dimension, so that the relative frequencies of each word in a given opinion define a specific 100-dimensional vector for that opinion. We use a dataset consisting of the Easterbrook opinions together with an equal number of Imposter opinions. We then apply PCA to the resulting collection of 100-dimensional vectors of all the opinions in our dataset, and examine the corresponding values of the principal components.

Figure 2 is a graph of the value of the *first* principal component (Y-axis) from this PCA analysis, for each opinion in the dataset (X-axis). The dark plus signs represent

Imposter opinions, and the light circles represent Easterbrook opinions; the opinions are all arranged chronologically from left to right. This first principal component accounts for 22% of the total variance, but it does *not* appear to distinguish between Easterbrook and Imposter opinions. Instead, it seems to be capturing a temporal shift, in that both the Easterbrook and Imposter opinions have values in this component which gradually decrease with time. So, this first principal component does not appear to be a useful method of identifying Imposter opinions.

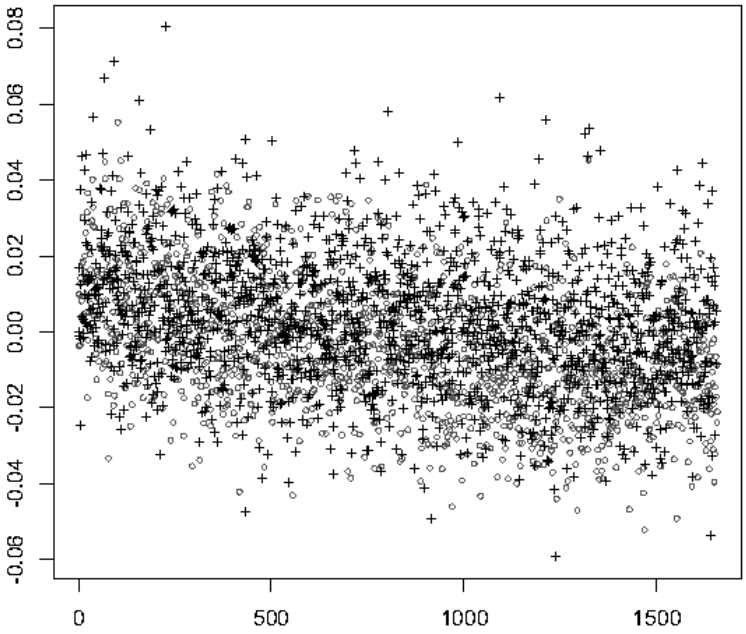

Figure 2: The value of the first principal component (Y-axis) of the PCA analysis for Easterbrook ('o') and Imposter ('+') opinions, arranged chronologically from left to right (X-axis), showing very little ability ability to identify the Imposters.

Figure 3 is a graph of the value of the *second* principal component (Y-axis) for each opinion in the dataset (X-axis), with the opinions again arranged chronologically from left to right. This component accounts for 9% of the total variance. It shows no temporal pattern, but it does reveal a clear separation between the two sets of points.

Specifically, for this component, the Easterbrook opinions tend to have larger values than the Imposter opinions. Indeed, a t-test indicates that these component values are greater for Easterbrook than for Imposter opinions at confidence level > 99.99 percent. This illustrates that the second PCA component is yet another promising method for identifying Imposters.[10]

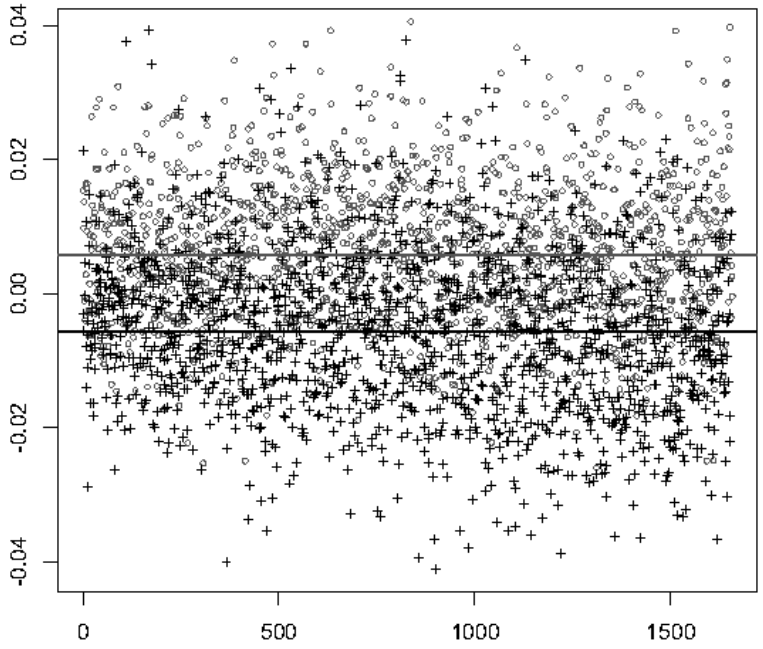

Figure 3: The value of the second principal component (Y-axis) of the PCA analysis for Easterbrook ('o') and Imposter ('+') opinions, arranged chronologically from left to right (X-axis), showing generally good ability to identify the Imposters. The horizontal lines show the means of the two different subsets.

[10]We also examined additional PCA components, but found that they did not aid in prediction, so we do not include them in our analysis.

III.5 Other Possible Approaches

We briefly mention several other methods of authorship identification, which we chose *not* to use in our analysis.

First, we excluded n-Grams, a method that compares frequency of sets of n letters (usually $n = 3$) in place of full words (see e.g. Suen 1979). In exploring this metric, we found it to be redundant with Function Words, ranking the Imposters almost exactly the same in each year. Since Function Words performed slightly better and are more intuitive, we omitted n-Grams from our analysis.

Second, we omitted an examination of *Instable Words*, i.e. words that have many near-synonyms so that their use involves lots of choice. While this approach has been successfully employed in certain other contexts (e.g. Koppel et. al. 2006), we were unable to identify clearly instable words which were relevant to legal writing, so we did not pursue this.

Finally, we considered a Bayes Factor analysis which compares the frequency of use of *all* words in one corpus to another, by calculating the probability of an opinion's words being drawn from all the words in the Easterbrook corpus versus all the words in the Imposter corpus. This method could very successfully identify the Imposter opinions, but only if it was *trained* using the Imposter opinions, i.e. it needed to "know" in advance which opinions were Imposters. For this reason, we did not believe that it could be successfully employed in identifying the clerk opinions, so we did not pursue it further.

IV Creating a Combined Outlier Score

Based on our analysis from the previous section, we wish to use the following four statistical measures to help identify outlier opinions:

- Rare Words, via the Poisson tail probabilities;
- Function Words, via the normal approximation log-likelihoods;

- Punctuation, via the Chi-Squared contribution terms;
- Principal Component Analysis, specifically the second principal component based on the 100 most common words.

We now discuss the challenge and our approach to combining these measures together to produce a single method of identifying outliers.

IV.1 Redundancy of the Different Measures

One issue when combining several different methods together is that they may be measuring essentially the same thing, and are thus just re-identifying the same outliers without adding any new power to the analysis. If so, this redundancy may make it inefficient or undesirable to combine these methods.

To investigate this, we compare our sets of results on the Easterbrook+Imposters corpus. The Rare Words approach correct identifies the Imposters as outliers (below the 0.05 probability cutoff) in the following sessions: 1985, 1986, 1987, 1991, 1992, 1995, 1998, 2007. The Function Word ranks for these sessions are, respectively: 5 of 20, 27 of 61, 44 of 70, 37 of 68, 10 of 60, 26 of 61, 23 of 76, 62 of 69. The Punctuation ranks for these same sessions are, respectively: 4, 4, 15, 44, 4, 37, 18, and 1.

Examining these rankings, we see that there is significant heterogeneity in ranking across methods, i.e. the different methods are *not* identifying precisely the same outliers. This result provides support for combining the different measures to produce a single overall score function.

IV.2 Combining Methods using Regression

Based on the above, we wish to combine our four different measures into one overall Outlier Score. We do this by taking a linear combination. To determine the coefficients of this linear combination, we run a linear regression again using the Easterbrook+Imposters corpus. The regression's output variable is simply the binary variable indicating whether or not an Imposter (judge) authored the opinion.

The input variables are the above four statistical measures, adjusted so that the Rare Words variable is taken to be 1 if the tail probability is less than 0.05 or otherwise 0, and the other three variables are given in terms of their percentile ranking over the entire corpus (to put them on a common scale). The regression will thus find the linear combination of our four measures which best identifies Imposters from the Easterbrook+Imposters corpus.

By repeating this regression approach 20 times (each with a fresh random selection of Imposter opinions) and average the results, we obtain the following as our best predictor:

$$\text{Predictor} = 0.00501 + 0.03539 \times \text{Punc} - 0.04592 \times \text{PCA} + 0.02992 \times \text{Func} + 0.00889 \times \text{Rare}\,,$$

where Punc is the percentile ranking of the Chi-Squared contribution from the Punctuation counts, PCA is the percentile ranking of the second component of the PCA analysis, Func is the percentile ranking of the log-likelihood from the Function Word counts, and Rare equals 1 if the Poisson tail probability from the Rare Word counts is less than 0.05 otherwise it equals 0. Here "Predictor" is designed to be close to 1 if the opinion is an outlier, and close to 0 if the opinion is truly written by Easterbrook himself.

V Results

We now apply our combined outlier score, in two contexts. First, apply it to the Easterbrook+Imposters corpus to identify Imposter opinions. Second, we apply it to the Easterbrook opinions only (excluding all Imposters) to identify clerk-drafted opinions.

V.1 Performance on Imposters

To test our combined outlier score on Imposter opinions, we use a fresh randomly-selected set of Imposter opinions (so we are testing the score on opinions that it

was not trained on). The resulting ranks of the Imposter opinions are presented in Table 3.

This table indicates that our method overall identifies Imposter opinions well. In eight of the 26 years, the Imposter opinion was ranked in first place. The opinions are ranked approximately 13th on average, i.e. in the 26th percentile. Across all years, our results are indeed statistically significant, i.e. they are significantly better than could be expected by random guessing.

Session	Imposter Rank	Total Opinions	Session	Imposter Rank	Total Opinions
1984-85	12	20	1997-98	5	76
1985-86	40	69	1998-99	6	66
1986-87	51	61	1999-2000	13	66
1987-88	6	70	2000-01	18	61
1988-89	1	50	2001-02	1	62
1989-90	14	70	2002-03	12	60
1990-91	1	63	2003-04	1	73
1991-92	2	68	2004-05	24	78
1992-93	20	60	2005-06	17	68
1993-94	22	78	2006-07	1	69
1994-95	1	67	2007-08	1	63
1995-96	49	61	2008-09	1	59
1996-97	19	62	2009-10	9	57

Table 3: Rankings of Imposters by our final outlier score, showing generally good ability to identify the Imposters.

We also summarise our results in terms of histogram of percentile rankings, in Figure 4. Consistent with Table 3, Figure 4 indicates that the percentile rankings are skewed heavily to the left, illustrating the ability of our combined method to identify Imposter opinions reasonably accurately.

V.2 Performance on Clerks – Training Data

We next rank all of the opinions attributed to Easterbrook according to our scoring method above, without implanting any Imposters. The resulting ranks of the true clerk opinions are then the object of interest: the lower the ordinal rank, the

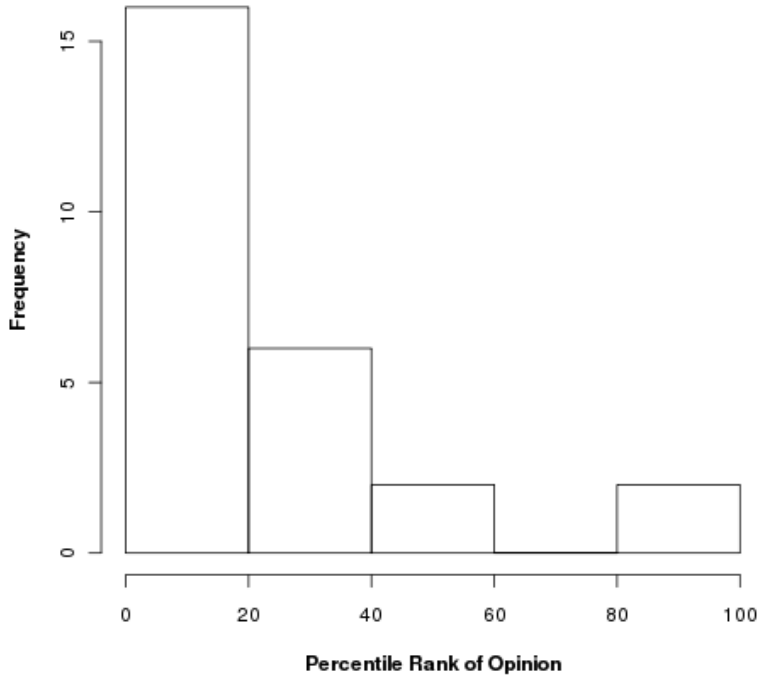

Figure 4: Histogram of rank percentiles for Imposter opinions, showing generally good ability to identify the Imposters.

better our method at identifying clerk-authored opinions. To preserve anonymity, we do not report the exact rankings of the clerk opinions, but provide only a histogram of the percentile ranks of the true clerk opinions, in Figure 5.

The histogram skews slightly left, towards lower rank numbers (i.e., percentiles closer to 0), illustrating that our method still performs better than random guessing. However, the results are not as dramatic as when identifying Imposters. Indeed, the average rank of the clerk opinions is approximately 29, which corresponds to approximately the 45th percentile of all Easterbrook opinions. These results are weakly statistically significant: a permutation test reveals that our rankings are indeed better than a random guess, significant at the 90% confidence level. That is, our results are modest; we can identify the clerk-authored opinions to some extent, but not with the same success as Imposter-authored opinions. Indeed, comparing Figures 4

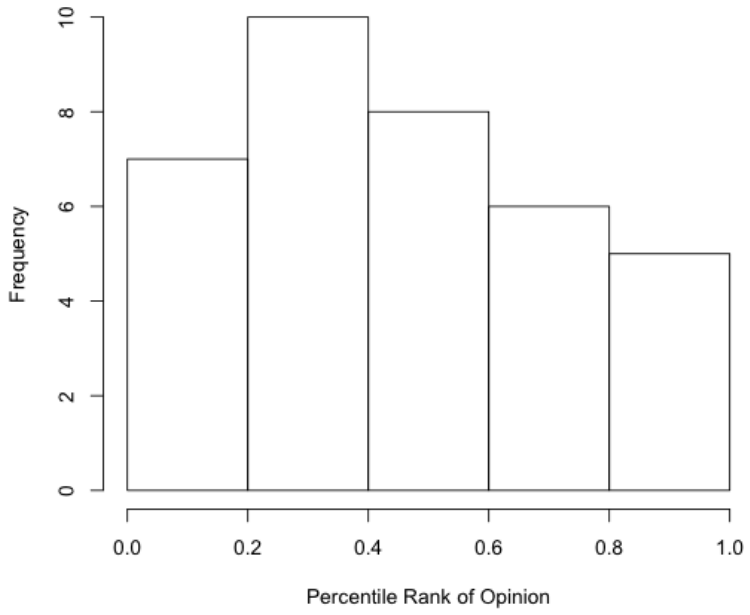

Figure 5: Histogram of rank percentiles for clerk-authored opinions, showing some (limited) ability to identify the Imposters.

and 5 shows that our method is much more successful in identifying Imposter opinions, than in identifying clerk-authored opinions in a corpus of Easterbrook-credited opinions.

V.3 Performance on Clerks – Testing Data

The relatively weak performance of our method on clerk-authored opinions raises the possibility that our scoring system may be ill-equipped to handle clerk-authored opinions specifically *because* it was originally designed to identify Imposter opinions.

To test this hypothesis, we performed a second round of statistical analysis, after *learning* the identities of the clerk-authored opinions for sessions beginning in even-numbered years (such as 1988–89 and 1994–95). Specifically, we used the same four statistical measurements, but we recalculated the regression coefficients (weightings) by performing a fresh linear regression not against Imposters but rather against the

clerk-authored opinions from these even-numbered sessions. (In this case, our data was limited, so we could not run multiple trials, but rather had to rely on a single linear regression.) Table 4 presents the resulting coefficients for each measure for this second analysis, and also compares them to those from the first analysis from the previous section.

	First Analysis	Second Analysis	Pct Change
Intercept	0.00501	-0.00747	-249%
Punctuation	0.03539	0.02707	-23%
PCA Value	-0.04592	0.00957	+121%
Function Words	0.02992	0.00123	-96%
Rare Words	0.00889	0.01093	+23%

Table 4: Change in regression coefficient values, from our first-round analysis (trained on Imposter opinions) to our second-round analysis (trained on clerk-authored opinions from even-numbered years), showing significant changes in the coefficients, although this does not lead to significantly better ability to identify clerk-authored opinions.

Table 4 indicates that the resulting coefficients do change somewhat in the second regression, i.e. when trained on true clerk-authored opinions (from the even-numbered sessions) instead of Imposter opinions. This raises the question of whether the new outlier scoring formula using these new coefficients will improve our ability to identify the clerk-authored opinions, when attempting to identify the (heretofore unknown) true clerk-authored opinions from the *odd*-numbered sessions.

However, we find that this is not the case: these new coefficients do not significantly improve our ability to identify clerk-authored opinions. In our original analysis, the odd-year clerk-authored opinions ranked at a mean percentile of 45. In this second-round analysis, after training on the even-year clerk-authored opinions, the odd-year clerk-authored opinions then rank at a mean percentile of 40. Although this does represent an improvement, a permutation test indicates that it is not statistically significant.

Illustrated graphically, Figure 6 shows the odd-year clerk-authored rankings from the original analysis (trained on Imposters), compared to that of the second-round analysis (trained using the even-year clerk-authored opinions). This plot indicates that there is no systematic improvement in the rankings of the clerk opinions between the first and second analysis. That is, training on half the clerk-authored opinions (as opposed to the Imposter opinions) gives us a different but roughly equivalent approach. This suggests that training was not the problem, i.e. that the clerk-written opinions are inherently less detectable than the Imposter ones.

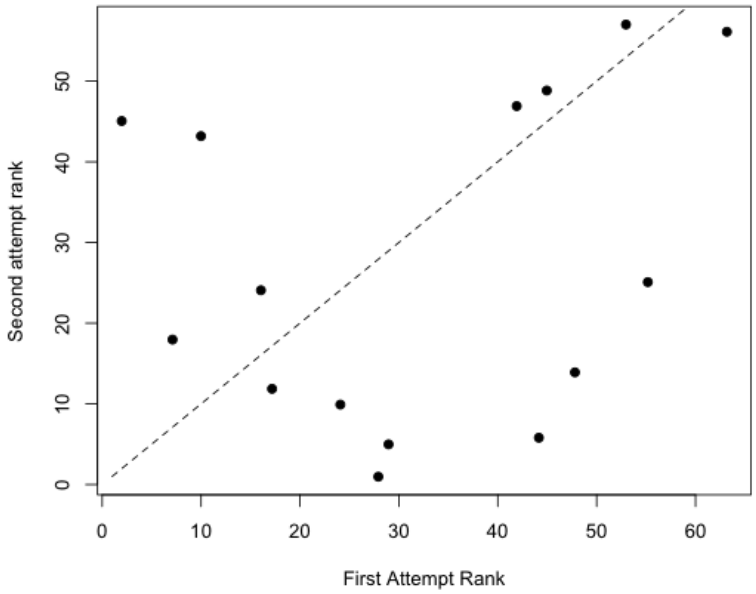

Figure 6: Comparison of rankings of clerk-authored opinions from our first-round analysis and from our second-round analysis, showing little improvement in our ability to identify the clerk-authored opinions. (Values are randomized to within 1 of the true ranking, to preserve confidentiality.)

VI Discussion

This article examines the extent to which, given a corpus consisting of opinions written almost exclusively by Judge Easterbrook, we can identify which opinions were *not* written by him. We consider a variety of statistical methodologies, and find that they can identify other judges' (i.e., Imposters') opinions quite successfully, but opinions written by Easterbrook's own clerks with only limited (but still statistically significant) success. This illustrates that the clerk-authored opinions are written in almost, but not quite, an identical writing style as Easterbrook's own opinions.

As stated from the outset, Judge Easterbrook represents a unique case. Anecdotal evidence strongly suggests that in most judicial chambers, it is common for clerks to draft multiple opinions (Choi and Gulati 2005). Easterbrook, by contrast, is known for only allowing clerks to write a single opinion after months of experience, and for both heavily supervising and heavily editing these writings. This helps to explain why Easterbrook's clerks' writing styles so closely mimic his own. This is in contrast to most other judges' clerks, for which statistical evidence suggests that their writing styles are sufficiently distinct from their judges' own writing styles as to lead to a detectable increase in overall writing style variability (e.g. Rosenthal & Yoon 2011b).

Despite the above, and despite what we assume to be Easterbrook's careful training of his clerks and editing of their draft opinions, our methods were still able to score the clerk-drafted opinions higher (on average) than the rest, and thus to statistically distinguish the former from the latter. It is worth noting that the manner by which Easterbrook operates his judicial chambers biases downward the differences between judge-written and clerk-drafted opinions that we likely would otherwise observe. First, the lengthy period that his clerks wait before drafting an opinion means that clerks have considerable time to learn Easterbrook's approach – stylistic, and quite possibly substantive – to judicial opinion-writing. Second, Easterbrook's editing of clerk-drafted opinions means that he may reduce – whether intentionally or not – the stylistic differences between the two.

Nevertheless, our results suggest that a real, quantifiable difference exists between

the writing of Easterbrook and of his law clerks. It is reasonable to assume that for a judge less dedicated to the quality of his or her clerks' work, these stylistic differences would be more pronounced. Of course, for other judges we do not have direct access to the information about which clerks participated in the writing of which decisions, so we are unable to test this hypothesis directly.

By contrast, we *were* able to identify the Imposter opinions with high success, even though the Imposter opinions were drawn from texts authored by (or attributed to) Easterbrook's fellow judges on the 7th Circuit Court. We assume that these judges are comparable in quality to Easterbrook and, given the random assignment of cases in the 7th Circuit, are writing on the same distribution of cases as Easterbrook.

If we had been unable to achieve success in ranking the Imposters, we might attribute our relatively modest success with the clerks to our choice of statistical measures; such a result would suggest either a homogeneity in writing style across judicial chambers, or more likely, flaws in our methodology. Or, if we could significantly improve our results by training on half the clerk opinions, we might attribute our modest success with the clerks to having trained our methods on the Imposter opinions. However, neither of these results manifested, suggesting that our moderate success with identifying the clerk-authored opinions is the result specifically of the clerks successfully managing (most likely strengthened by the aid of Easterbrook's close supervision and editing) to imitate Easterbrook's writing style. That is, even if clerk writing is not statistically identical to Easterbrook, it is closer to his writing than anything else is – and in particular, much closer to his writing than his fellow judges can approach.

It bears repeating that our analysis focuses on the writing style, not substance, of judicial opinions. We leave the important question of substance to another day. We recognize, however, that judges' reliance on law clerks – substantively even more than stylistically – is a practice not to be taken lightly. No amount of care and attention can produce writing that is qualitatively identical to the judge to whom it is being attributed. However, proper dedication, such as that shown by Easterbrook, can produce work by clerks that, stylistically across several measures, closely

approximates those by Easterbrook himself. In particular, clerk writings under such conditions can achieve not just judge-style quality of writing, but writing that is *closer to their particular judge*.

One can debate the relative merits of the effort needed to train such clerks, but our results suggest that it is not a futile exercise, and that a clerk can indeed learn to successfully replicate the writing style of the judge. And if form follows substance – a supposition that we hope to explore in future work – then our findings may perhaps give some reason to believe that if Easterbrook's clerks take pains to write like Easterbrook, then they may have learned to *think* like him as well.

VII Conclusion

This article has examined the degree to which judicial opinions drafted by law clerks can approximate those of their judge. We focused on the unique case of Judge Easterbrook, who writes all of his own opinions aside from allowing each of his clerks to write a first draft a single opinion. We found that while clerk-drafted opinions are statistically distinguishable from those of the judge, they are much more difficult to detect than are opinions written by any of his fellow jurists on the U.S. Court of Appeals for the 7th Circuit.

While our analysis focused on judicial writing, its applications can extend to any situation in which one is trying to determine authorship from a series of writings. Much of the focus surrounding author identification examines whether a writing can be attributable to one identifiable author or another. In this article we instead ask, in a corpus of writing, do certain writings appear stylistically inconsistent from the others? Our results suggest that individuals by nature possess distinct writing styles, but with the right motivation and supervision and editing, they can closely adopt another's writing style.

References

[1] E.M. Airoldi, A.G. Anderson, S.E. Fienberg, and K.K. Skinner (2006), Who wrote Ronald Reagan's radio addresses? Bayesian Analysis **1(2)**, 289–320.

[2] John Burrows (2002). 'Delta': a Measure of Stylistic Difference and a Guide to LIkely Authorship. *Literary and Linguistic Computing* **17 (3)**, 267–287, Oxford University Press, Oxford, UK.

[3] John Burrows and Hugh Craig (2001). Lucy Hutchinson and the Authorship of Two Seventeenth-Century Poems: A Computational Approach. *The Seventeenth Century* **16 (2)**, 259–282.

[4] Gordon Campbell, Thomas N. Corns, John K. Hale, Fiona J. Tweedie (2009). Milton and the Manuscript of De Doctrina Christiana. *Renaissance Quarterly* **62 (4)**, 1388–1389.

[5] Stephen J. Choi and G. Mitu Gulati (2005). Which Judges Write Their Opinions (And Should We Care)? *Florida State University Law Review* **32** 1077.

[6] Denise Dillon, David Cottrell, Joseph Reser (2008). Group differences in word use and meaning: a text analysis of the abstract word, 'Values'. 9^{es} *Journées internationales d'Analyse statistique des Données Textuelles*, JADT 2008, 389–396.

[7] A. Ellegard, A statistical method for determining authorship : the Junius letters, 1769–1772. Gothenburg studies in English, Vol. **13**. Göteborg, 1962.

[8] Moshe Koppel, Navot Akiva, and Ido Dagan (2006). Feature Instability as a Criterion for Selecting Potential Style Markers. *Journal of the American Society for Information Science and Technology* **57 (11)**, 1519–1525.

[9] Edward Lazarus. *Closed Chambers: The Rise, Fall, and Future of the United States Supreme Court*. New York: Penguin Books. 1998.

[10] Adam Liptak *A Second Justice Opts Out of a Long-Time Custom: The 'Cert' Pool. New York Times*, September 26, 2008 at A21.

[11] David Madigan, Alexander Genkin, David D. Lewis, Shlomo Argamon, Dmiriy Fradkin, Li Ye (2005). Author Identification on the Large Scale. Proceedings of CSNA-05.

[12] Fredrick Mosteller and David L. Wallace. *Applied Bayesian and Classical Inference: The Case of The Federalist Papers.* Springer-Verlag, New York. 1964.

[13] Todd C. Peppers. *Courtiers of the Marble Palace: The Rise and Influence of the Supreme Court Law Clerk*. Stanford: Stanford University Press. 2006.

[14] Richard A. Posner (2002). Diary. *Slate,* Jan. 15th, 2002. (available at http://www.slate.com/id/2060621/entry/2060742/)

[15] Jeffrey S. Rosenthal and Albert H. Yoon (2011a). Detecting Multiple Authorship of United States Supreme Court Legal Decisions Using Function Words. *Annals of Applied Statistics* **5 (1)**, 283–308.

[16] Jeffrey S. Rosenthal and Albert H. Yoon (2011b). Judicial Ghostwriting: Authorship on the Supreme Court. *Cornell Law Review* **96 (6)**, 1307–1343.

[17] Efstathios Stamatatos (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology* **60 (3)**, 538–556.

[18] Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis (1999). Automatic Authorship Attribution. *Proceedings of EACL* University of Patras, Patras, Greece.

[19] Ching Y. Suen (1979). N-Gram Statistics for Natural Language Understanding and Text Processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. PAMI-**1 (2)**, 164–172.

[20] Paul J. Wahlbeck, James F. Spriggs II and Lee Sigelman (2002). Ghostwriters on the Court? A Stylistic Analysis of U.S. Supreme Court Opinion Drafts *American Politics Research* **30** 166–172.

[21] Artemus Ward and David L. Weiden. *Sorcerers' Apprentices: 100 Years of Law Clerks at the United States Supreme Court* New York and London: New York University Press. 2006.