The Philippine Review of Economics
Vol. LI No. 1, June 2014 pp. 162–178

When did we begin to spell "heteros*edasticity" correctly?

Alfredo R. Paloyo

Using digitized texts scanned by Google and subjected to optical character recognition, I show that *heteroskedasticity* overtook *heteroscedasticity* as the preferred spelling in 2001 and has continued to dominate, except for 2005, up to 2008. The latest trends indicate that writers are moving toward the *k* variant. However, for words such as *homoskedasticity*, *heteroskedastic*, and *homoskedastic*, the corresponding spellings using *c* are still overwhelmingly dominant, albeit slowly shifting.

JEL classification: A20, B19, B29
Keywords: culturomics, econometric orthography, Google Books, heteroskedasticity, philology

1. Introduction

In a brief article in *Econometrica*, J. Huston McCulloch [1985b] contended that "the most pressing issue in econometric orthography today is whether heteros*edasticity should be spelled with a k or with a c." At the time his note went to press, the majority of published manuscripts spelled it as *heteroscedasticity*. Arguing that since it was coined directly from Greek source words into English without French or Latin distillation, he declared confidently that "[h] eteroskedasticity is therefore the proper English spelling."

In an econometric context, heteroskedasticity is the phenomenon wherein the random disturbance term exhibits a nonconstant conditional variance. Consider the linear regression model in matrix notation: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Let the observation index i run from 1,…, N. Heteroskedasticity is present if the conditional variance of the error term, $\boldsymbol{\varepsilon}$, is $\text{Var}[\varepsilon_i|\mathbf{X}] = \sigma_i^2$—that is, if the spread or dispersion of ε is a function of specic values of $\mathbf{X}$. A stock example is the variance of food expenditure conditional on income, where higher income corresponds to a higher variance in food outlay. This has implications for the consistent estimation of the standard deviation of the estimated coefficients or the standard error, s.e.$(\widehat{\beta})$. The details are beyond the scope of this article, but a textbook treatment of the

topic is readily available for both novices [Gujarati and Porter 2008] and experts [Wooldridge 2010].

Using digitized texts scanned by Google and subjected to optical character recognition (OCR), I show that *heteroskedasticity* did in fact eventually overtake *heteroscedasticity* as the preferred spelling but only did so consistently 15 years after the publication of McCulloch [1985b]. However, for the noun *homoskedasticity* and the adjectives *heteroskedastic* and *homoskedastic*, the corresponding spellings using *c* still dominate, although the recent trend has been to move toward the *k* variants. For instance, *homoskedasticity* recently surprassed *homoscedasticity* in published books.

2. The etymology of heteroskedasticity

Heteroskedasticity has its roots in two Greek words: *éteros*, meaning "other" or "different"; and *skedánnymi*, meaning "to scatter" [McCulloch 1985b].[1] The word *skedánnymi* actually comes from Ancient Greek; its Modern Greek rendition is *skorpízo*. There is no contention regarding the first part of the word, but other sources attribute different root words for -*skedasticity*, such as *skedastós* ("capable of being scattered")[2] and *skédasis* ("scattering" or "dispersion").[3] In its entry on *homoscedasticity*, the *Merriam-Webster Dictionary* refers to *skedastikos* ("able to disperse"), which comes from *skedannynai* ("to disperse").[4] Clearly, however, these alternative root words are merely derivatives of the same idea. The word in question is therefore a neoclassical compound, for it is a combination of Ancient Greek words (what the *Oxford English Dictionary* calls "combining forms") and is imbued with a specific technical meaning.

McCulloch mentioned in a footnote that he found the earliest use of either *heteroskedasticity* or *heteroscedasticity* in a 1923 statistics text by Truman L. Kelley. He does not quote the original text at the request of the editor, but he presumably referred to the following instance on section 48 ("Properties of correlation surfaces") of the book *Statistical Method*: "If the standard deviation of the successive *x* arrays are equal, the distribution is homoscedastic in the *x* variable" [Kelley 1923:172]. Aldrich (see footnote 2) and David [1998][5] write,

[1] McCulloch spells it as *skedánnumi* but I render the Greek letter *υ* as *y* to emphasize its earlier correspondence with the Latin *y*. This is also the transliteration recommended by ISO 843:1997. In Greek, heteroskedasticity is *heteroskedastikótita*.

[2] This is from the contribution of John Aldrich to the Web site *Earliest Known Uses of Some of the Words of Mathematics* maintained by Miller [2011]: http://jeff560.tripod.com/h.html.

[3] This is from the Wikipedia entry on *heteroscedasticity*, which incidentally uses the *c* variant (replacing the first *c* with a *k* in the following will redirect the browser to the former): http: //en.wikipedia.org/wiki/Heteroscedasticity.

[4] See http://www.merriam-webster.com/dictionary/homoscedasticity. Schwartzman [1994] traces the roots of the Greek words even further to the Indo-European *sek*, *sked*, and *skel*, which ultimately gave rise to the English words such as *schizophrenic*, *shatter*, *scoliosis*, and *isosceles*.

[5] An update is available at http://www.stat.iastate.edu/preprint/articles/2011-10.pdf.

however, that the term was introduced earlier by Karl Pearson in 1905 in the article "On the general theory of skew correlation and non-linear regression".[6]

A search through Google Books (see section 3) returned results indicating that the first instance appeared in an earlier Manuscript by Pearson, "On the theory of contingency and its relation to association and normal correlation", which was published in 1904. This is an obvious technical mistake since searching through that particular paper does not reveal any variant of the words *heteroskedasticity* or *heteroscedasticity*. Thus, I confirm that the Pearson [1905] publication contains the first incarnation of the term. Specifically, Pearson introduces the term for the first time as follows:

> the variability of an array, *i.e.*, the standard deviation of an array … may or may not be the same for all arrays. If it is the same, or all arrays are *equally scattered* about their means, I shall speak of the system as a *homoscedastic* system, otherwise it is a *heteroscedastic* system. [Pearson 1905]

Since the origin of *heteroscedasticity* is clearly Greek, it is useful to briefly enumerate the three channels through which the Greek language has contributed to the modern English lexicon. First, a Greek word could be a direct donor, such as *kinetic* from the Greek *kinetikos* (from *kinein*, to move). Second, it may come to English through an intermediate language, such as Latin[7] or French.[8] Aldrich [2011] cites the example of *geometry*, which was Greek *geometria* (from *gē*, "earth", and *metréō*, "to measure"), then Latin *geometria*, and then French *géométrie*, after which it was adopted into English. This percolation through the sieves of various languages naturally has implications for both the word's pronunciation and its orthographic manifestation. Third, a lexical gap in English may be filled by coining a modern word using Greek roots: the neoclassical compound.[9] It is this third way that *heteroscedasticity* came to English.

Aldrich [2011] notes that, in the context of coining new technical words in mathematics and statistics, Pearson's "specialty was the Greek-based neologism" and gives the examples of *histogram*[10] and *heteroscedasticity*. Pearson introduced many other such terms in two lecture series in 1892 while he was professor of

[6] The *Merriam-Webster Dictionary* also lists 1905 as the year of first-known use of *homoscedasticity* but does not mention its source. Oddly, Van Loco [2007] writes—in a footnote explaining the etymology of heteroskedasticity—something extremely similar to what Aldrich contributes to *Earliest known uses of some of the words of mathematics* [Miller 2011] but instead of referencing Aldrich, he cites McCulloch [1985b], who did not mention Pearson at all and who referred to *skedánnymi* instead of *skedastós*.

[7] Latin was primarily introduced by the Christian missionaries beginning in the 5th to 6th centuries. It was also, at that time, the lingua franca of Europe.

[8] French was brought in via the Norman conquest of England when William, Duke of Normandy, defeated King Harold II of England at the Battle of Hastings in 1066.

[9] For the special case of neoclassical compounds in English morphology, see Bauer [2005].

[10] The term has Greek root words, specifically *istos* ("mast", as on a ship) and *gramma* ("something written"). See Ioannidis [2004].

geometry at the prestigious Gresham College: "The syllabi show an abundance of fancy terminology: stigmograms, euthygrams, epipedograms, histograms, chartograms, hormograms, topograms, stereograms, radiograms, and isodemotic lines", writes Stigler [1986]. A search through Miller's [2011] compendium of *Earliest known uses of some of the words of mathematics* reveals the substantial contribution of Pearson to modern statistical terminology.

Direct Greek contributions to modern scientific terms in English are relatively recent phenomena. Before the Renaissance (which began around the 14th century), Greek-sourced words came into English primarily after having already undergone a Latin or French transliteration. At the dawn of the Industrial Revolution around the 18th century, terms from earlier Greek works were being directly imported into English along with Latin terms with Greek roots to represent new knowledge. Modern coinages (e.g., *heteroscedasticity*), particularly classical compounds (which were popular in scienitific and technical applications), reached their peak in the 19th century [Aldrich 2011].

Since Pearson invented the term in 1905, McCulloch is therefore correct in noting that *heteroscedasticity* is a modern coinage and that "[t]he letter in question is … the transliteration of the Greek kappa (κ)." How, then, is κ transliterated into English? While certain Greek letters, such as γ, δ, υ, and χ, have varying transliterations, κ is nowadays invariably represented with the letter *k* in English. The Beta Code,[11] the BGN/PCGN romanization,[12] the ISO 843:1997 system,[13] the UN romanization system,[14] and the US Library of Congress transliteration chart (for both ancient/medieval and modern Greek) all recommend transliterating κ as *k* (or *K*, as the Beta Code recommends capital letters). However, these attempts at standardization are all recent initiatives and were codified after Pearson had coined the term.

Nonetheless, Greek words with κ directly imported into English typically would have had the κ replaced with a *k* even before formal attempts at standardization (a notable exception is the aforementioned *isosceles*). McCulloch provides the examples *skeptic* and *skeleton* but these words were not lifted directly from the Greek. *Skeptic*[15] entered via the French (*sceptique* from the Latin *scepticus*) while

[11] Developed by David W. Packard in the 1970s, the Beta Code is a system of representing ancient Greek using ASCII (American Standard Code for Information Interchange) characters.

[12] The US Board on Geographic Names (BGN) and the Permanent Committee on Geographic Names (PCGN) for British Official Use have their own transliteration convention for geographic names.

[13] This system applies to Greek script regardless of the period in which it was used.

[14] This is based on the ELOT 743 conversion system of the Greek Standardization Organization, which also forms the basis for ISO 843:1997.

[15] This is the preferred spelling in American English; the variant *sceptic* is predominantly found in British English. Indeed, since the word entered through French, it used to be spelled with a *c* on both sides of the Atlantic. However, largely successful US spelling reforms initiated by Noah Webster (of *Merriam–Webster* fame) emphasized phonetic faithfulness (hence, *color* [AE] for *colour* [BE], among many others). Oddly enough, despite the soft *c* [s] pronunciation in French, the persistent pronunciation in English is [k].

skeleton appeared in Late Latin (c.300–c.700) as *sceletus*. A more appropriate example would have been *kinesis* or *kinetic* from the Greek *kinesis*.

As one can already see, when Greek words entered Latin or French, the κ was transliterated as *c*. It is through this channel that the original κ in Greek may appear in English as *c* subject to palatalization in Late Latin.[16] McCulloch provides *scepter* (Old French [c.900–c.1400] *sceptre*, Latin *sceptrum*, Greek *skeptron*), *scene* (Middle French [c.1400–c.1600] *scène*, Latin *scaena*, Greek skené), and *cyclic* (French *cyclique*, Latin *cyclicus*, Greek *kyklikos*) as relevant instances.[17]

The lack of consensus over the spelling of heteroskedasticity has carried over to its pronunciation. Within the scientific community, the questionable *k* or *c* is typically enunciated as a voiceless velar plosive [k], regardless of how the word is actually spelled. McCulloch reiterates this in *Econometrica*. *Merriam-Webster*, however, indicates that it is the voiceless alveolar sibilant (grooved fricative, [s]), consistent with its suggested spelling of *homoscedasticity*. This is commonly known as the soft *c*, which is the usual phoneme when *c* appears before *e*, *i*, and *y* (*sceptic* in British English is a notable exception). Schwartzman [1994] is more permissive: "The word *heteroscedastic* may be pronounced as if the first *c* were a *k* or as if the first *c* were omitted." The question of pronunciation even found its way to *The math forum @ Drexel*, a leading online resource for teaching mathematics and statistics, when Jeff Miller of *Earliest known uses* [Miller 2011] raised the matter, citing the soft-*c* recommendation of the instructor's guide to the book *Statistics in action*.[18]

3. Word history and prevalence using Google Books

The data used in this paper are made publicly available by Google through its Google Books project and the multi-institutional team behind Culturomics. Originally known as Google Print that started in 2004, the service allows users to search the full text of—as of October 2010—"15 million books from more than 100 countries in over 400 languages" [Crawford 2010]. This represents about 12 percent of all published books. Searching through the contents of each book is made possible by subjecting the digitized copies to optical character recognition (OCR). Google Books holds perhaps the largest corpus of collected human

[16] In phonetic terminology, there are two different phonemes here: [k] and [c], where the former is a voiceless velar plosive and the latter is a voiceless palatal plosive.

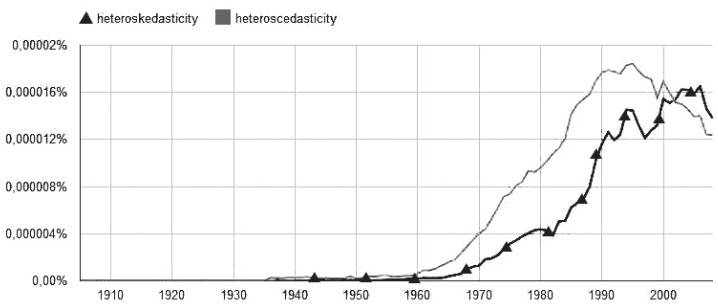
[17] See the *Online Etymology Dictionary* maintained by Douglas Harper for details: http://www.etymonline.com/.

[18] See http://mathforum.org/kb/message.jspa?messageID=5906833 and the discussion therein. Like Aldrich, as narrated by Miller, I have never heard the soft *c* pronunciation in English. McCulloch rightly points out that it is *hétéroscédasticité* in French and is consequently pronounced as [s] in that language. A French colleague informs me that when French economists speak in English, they are much more likely to use [s] in this case.

knowledge and it is still rapidly growing, as even more published materials are scanned.[19]

Starting with the collection of Google Books, Michel et al. [2011]—the "Culturomics team"—selected about five million books based on the quality of the resulting OCR output and the metadata. The assembled database contains over 500 billion words, of which 361 billion are English. For a particular word to appear in the dataset, it must have appeared in at least 40 books. The dataset begins in the 1500s, but I restricted the subsequent analysis to published works beginning in 1905, the year when Pearson coined the term *heteroscedasticity*. The series ends in 2008. I used the English corpus[20] for the graphs below.

Figure 1 shows the prevalence of the words *heteroskedasticity* and *heteroscedasticity* over time, with the horizontal axis representing years and the vertical axis representing the share of these two words in the body of published (English) words for a particular year. The lines have been smoothed using a three-period moving average (MA(3)). For much of the 20th century, the *heteroscedasticity* variant coined by Pearson outnumbered the *k* variant.



Source: Google Books Ngram Viewer, http://books.google.com/ngrams.

Note: The graph was generated on 29 November 2012. The vertical axis is the share (in percent) of the word of interest in the corpus of published works.

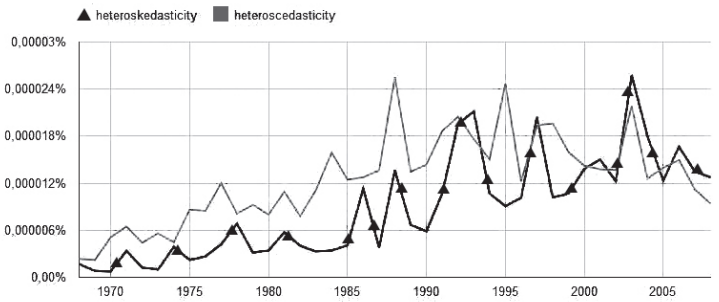
FIGURE 1. Heteroskedasticity and heteroscedasticity, 1905–2008, MA(3)

It was not until 1974 when the *k* variant overtook the *c* spelling, as seen in Figure 2, where I restrict the time dimension to 1968–2008 and remove the smoothing to show the raw counts. This *k*-dominance, however, lasted only

[19] Sections of the following materials were retrieved using Google Books: Statistical method [Kelley 1923], The history of statistics [Stigler 1986], The words of mathematics [Schwartzman 1994], "On the general theory of skew correlation and non-linear regression" [Pearson 1905], Problems in education [Western Reserve University 1927], "The borderline between derivation and compounding" [Bauer 2005], and History of Friedrich II of Prussia [Carlyle 1858].

[20] The other corpora are American English, British English, Chinese (simplified), English Fiction, English One Million, French, German, Hebrew, Spanish, and Russian.

one year. Subsequently, *heteroskedasticity* briefly appeared more times than *heteroscedasticity* for the years 1986, 1992, and 1993. Except for 2005, all years after 2001 show a consistent preference for *heteroskedasticity*.

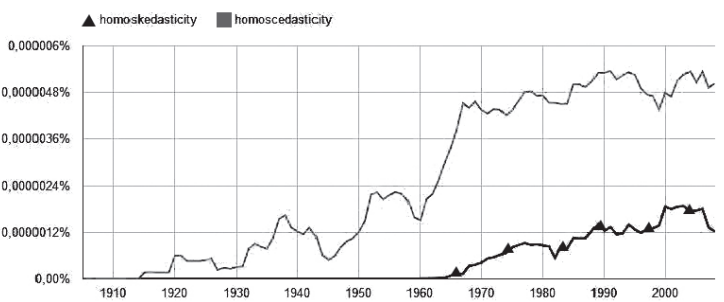


Source: Google Books Ngram Viewer, http://books.google.com/ngrams.

Note: See note on Figure 1.

FIGURE 2. Heteroskedasticity and heteroscedasticity, 1968–2008, MA(0)

Interestingly, for the phenomenon of homoskedasticity, the *c* spelling dominates for all but one year (see Figure 3). Although not evident from the figure because of the MA(3) smoothing, *homoskedasticity* only appeared more than *homoscedasticity* for the final year of the dataset (2008). It seems, therefore, that *heteroskedasticity* is leading the way and that *homoskedasticity* is catching up with changes in spelling preference. Perhaps this is because *heteroskedasticity* tends to be more often used than *homoskedasticity* anyway, so any changes in spelling will be reflected in the former ahead of the latter.

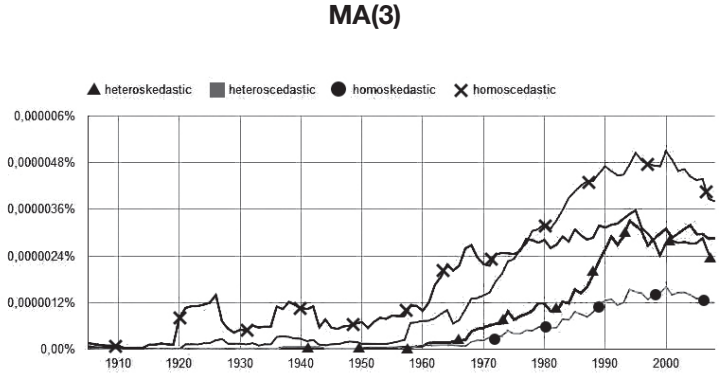


Source: Google Books Ngram Viewer, http://books.google.com/ngrams.

Note: See note on Figure 1.

FIGURE 3. Homoskedasticity and homoscedasticity, 1905–2008, MA(3)

In Figure 4, I plot the corresponding adjectives over time with MA(3) smoothing. Here, it is quite clear that the *c* variant dominates for the whole period. It is rather odd that, while the noun *heteroskedasticity* has overtaken its *c* variant, the adjective seems to have lagged so far. However, the overall trend seems to indicate that the two variants of the adjective will converge any time soon.

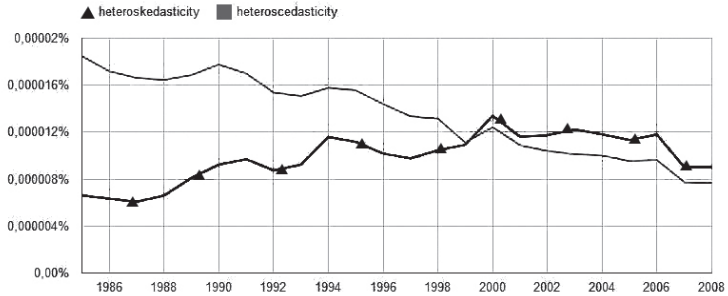


Source: Google Books Ngram Viewer, http://books.google.com/ngrams.

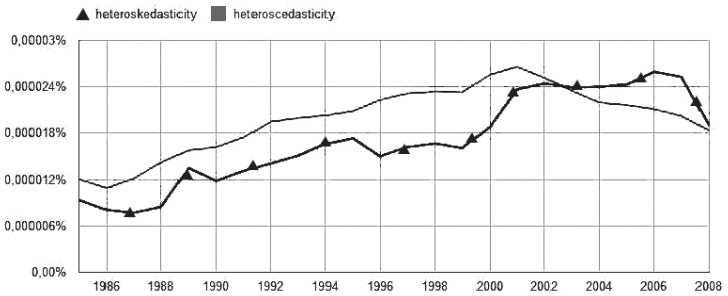
Note: See note on Figure 1.

FIGURE 4. Heteroskedastic, heteroscedastic, homoskedastic, and homoscedastic, 1905–2008,

Is there a difference in the prevalence of each variant of heteroskedasticity between American and British usage? After all, Pearson is British and the influential economists who used *k* are American (see section 4). To see this, I graph *heteroskedasticity* and *heteroscedasticity* separately for the American and British English corpora of Google Books for the period 1985–2008 in Figure 5. These corpora are restricted by the place of publication of the book in the dataset. Here, one can see that British publications are only recently moving toward *heteroskedasticity*, and recent years are showing a rapid decline in the usage of *heteroscedasticity*. In comparison, American orthography has been more accepting of the newer *k* variant and is shifting away from Pearson's original spelling.



(a) American English



(b) British English

Source: Google Books Ngram Viewer, http://books.google.com/ngrams.

Note: See note on Figure 1.

FIGURE 5. Heteroskedasticity and heteroscedasticity, 1985–2008, MA(3), American and British English corpora

The corpus of English texts made available by the Culturomics team is limited because it is only a sampling of the books that the Google Books project has subjected to OCR. Other materials were excluded, such as journals, periodicals, magazines, and the like. Moreover, the material found on the Internet is not included. Thus, I tabulate the number of search hits in Google and Google Scholar, which indexes only scholarly literature, of the terms of interest to complement the numbers retrieved from Google Books (see Table 1). Similar to the figures above, search data from Google indicate that it is only with the noun *heteroskedasticity* that the *k* variant has become more common. For the other instances, the *c* variant is still heavily dominant (in particular, for the noun *homoscedasticity*).

TABLE 1. Google and Google Scholar search hits

Word	Google	Google Scholar
Heteroskedasticity	866,000	72,900
Heteroscedasticity	460,000	65,300
Homoskedasticity	48,000	8,840
Homoscedasticity	212,000	30,800
Hetereoskedastic	152,000	22,100
Heteroscedastic	209,000	30,600
Homoskedastic	59,200	7,720
Homoscedastic	80,100	11,600

Note: Data were retrieved on 24 November 2011.

I also queried JSTOR (Journal Storage) for the relevant terms to examine their prevalence in the academic journals in JSTOR's archive. Column (1) of Table 2 is the result of queries in the JSTOR database of economics, mathematics, political science, sociology, and statistics journals. Column (2) is restricted only to economics and statistics journals. In this case, it can be seen that the academic literature is still largely faithful to the original spelling of Pearson. However, when I restrict the sample to the years 2000 and after, the dominance of *heteroskedasticity* once again becomes apparent (see column [3], which is the post-1999 restricted version of [1]).

TABLE 2. JSTOR search hits

Word	(1)	(2)	(3)
Heteroskedasticity	4,729	4,659	1,821
Heteroscedasticity	5,244	5,176	1,349
Homoskedasticity	498	492	160
Homoscedasticity	1,298	1,266	205
Hetereoskedastic	1,583	1,566	538
Heteroscedastic	2,764	2,739	774
Homoskedastic	583	576	203
Homoscedastic	1,438	1,426	322

Note: Data were retrieved on 24 November 2011.

Incidentally, a comparison between columns (1) and (2) of Table 2 illustrates that the concept of heteroskedasticity is almost exclusively the concern only of economists and statisticians. It has yet to substantially cross-pollinate the allied social sciences with a significant quantitative subfield, such as political science and sociology.

Overall, the term *heteroskedasticity* has been the lone shooting star among the *k* siblings. While *heteroskedasticity* has overtaken *heteroscedasticity* and has been

steadily outnumbering its *c* rival, the other *k*'s have consistently failed to surpass their *c* cousins. The trends from the figures above, however, seem to indicate that the profession is marching toward the *k* variant, albeit perhaps tentatively. One reason for this transition is presumably the phonetic faithfulness of the *k* spelling and, to a rather minor extent, the standardized transliteration of the Greek letter *κ* into k, which the *k* crusaders such as McCulloch like to point out. Of course, incredibly influential papers such as that of White [1980] massively contribute to this trend, as well as the "simultaneous growth in intellectual prevalence and authority of US journals and authors in economics".[21]

4. Let the econometricians speak!

Although one cannot conclude with absolute certainty, the spike of *heteroskedasticity* in 1986 (Figure 2) may be the result of McCulloch's 1985b publication. In private correspondence, McCulloch mentions that the original motivation for the *Econometrica* article is the insistence of an editor of the *Journal of Banking and Finance* to spell heteroskedasticity with a *c* in McCulloch [1985a], "Interest-risk sensitive deposit insurance premia: stable ACH estimates", where ACH means adaptive conditional heteroskedasticity. McCulloch wanted to show that *k* is "not only acceptable, but preferable." The journal then went along with *k* and he sent his note to *Econometrica*.

However, five years before the publication of McCulloch, Halbert White [1980] already published his influential paper, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity", in *Econometrica*. According to Kim, Morse, and Zingales [2006], this is the most cited paper in economics since 1970. By their count based on a selection of top general-interest and field journals, White [1980] has been cited 4,318 times, with the next-most-popular paper cited a little less than 300 times White's citation count. Google Scholar[22] reports 16,704 citations. Computing "White standard errors" (or "robust standard errors") is now standard in most applications.

In Kim, Morse, and Zingales's [2006] paper, "What has mattered to economics since 1970", *heteroskedasticity* appears five times in the titles of their list of influential papers. Those before 1990 include the aforementioned White [1980] paper and also Newey and West [1987] and Bollerslev [1986], cited 9,585 and 12,722 times, respectively, in Google Scholar. *Heteroscedasticity* appears in two papers—namely in Engle [1982] and Breusch and Pagan [1979] for a combined Google Scholar citation count of 15,367, which is 23,644 counts less than just the pre-1990 influential *heteroskedasticity* papers.

[21] I thank a reviewer for the last point.

[22] The citation numbers from Google Scholar reported here and the next paragraph were retrieved on 29 November 2012.

In an email, White suggests that he had perhaps three sources for his preference for the *k* variant: first, his PhD supervisor at Massachusetts Institute of Technology (MIT), Jerry Hausman; second, the *Econometric theory* textbook of Arthur S. Goldberger [1964]; third, the textbook of Jan Kmenta [1971], *Elements of econometrics*, which he "very much liked". He believes that he most probably picked it up from Goldberger [1964] since he was "strongly influenced" by that book. Among other factors, White's choice of using *k* and the subsequent popularity of his 1980 publication most certainly pulled the profession away from the *c* variants.

Hausman writes that he did, in fact, use *k* in his lectures at MIT. He does not recall how exactly he came to prefer that spelling but notes that he might have followed "the Greek approach". He could have also influenced other people at MIT, such as Whitney Newey and Kenneth D. West (now at the University of Wisconsin), both of whom were his students in econometrics and who jointly developed a technique to estimate heteroskedasticity- and autocorrelation-consistent (HAC) standard errors.[23] Although West notes that they were familiar with White [1980] and the White [1984] textbook, *Asymptotic theory for econometricians*, there was, to the best of his recollection, no discussion between him and Newey on how to spell heteroskedasticity: spelling it with a *k* just seemed "the natural one."

Of course, what was natural in the '80s may not have been so in the '70s. Trevor Breusch recalls that the *c* variant was dominant at the time he and Adrian Pagan wrote Breusch and Pagan [1979]. This is apparent in the references cited in their paper. The majority of the textbooks he owned also used the *c* spelling and thus their decision preferring *c* was not entirely a deliberate one. As an indication of the enduring influence of the McCulloch [1985b] publication, Breusch writes that he was convinced by the arguments laid out therein and has since switched to the *k* variant.

Evidently, while the authority of Karl Pearson over Greek-based mathematical and statistical neologisms is redoubtable, his coinages are not immune to orthographic mutations, especially when such mutations are driven by a well-argued position published in a reputable scholarly journal. Indeed, *Caesar non supra grammaticos*—Caesar is not above the grammarians.[24]

[23] West also acknowledges Daniel McFadden (now at the University of California at Berkeley) and Franklin M. Fisher as his former econometrics professors. Both McFadden and Fisher seem to have oscillated between the *c* [McFadden 1987; Fisher et al. 1966] and *k* [McFadden 1974; Fisher, McGowan and Evans 1980] variants, though. But, in fact, so did Hausman in 1987 when he and Paul A. Ruud spelled it with a *c* in the same issue of the *Journal of Econometrics* where McFadden [1987] appeared. So, perhaps that was the call of the editor, Richard Blundell, who predominantly writes with a *c*.

[24] In opening the Council of Constance (1414–1418), Sigismund, Holy Roman Emperor, had used the word *schisma* as feminine instead of neuter: "*Date operam, ut illa nefanda schisma eradicetur*", referring to Hussite Bohemia during the Western Schism (otherwise known as the Three-Popes Controversy). A cardinal reminded him, "*Domine, 'schisma' est generis neutrius*" [Your Majesty, schisma is neuter]. To which Sigismund replied, "*Ego sum Rex Romanus, et super grammaticam!*" [I am King of the Romans, and above grammar] [Carlyle 1858].

5. Conclusion

Writing for the *Financial Times* in 1998, economist John Kay says, "Every serious subject has its jargon. Economists need to know about heteroscedasticity. I take this example because it is virtually impossible to pronounce, and impossible to use the word in front of a class without everyone bursting out into laughter. Indeed, most spell-check programs reject it, and offer improbable or embarrassing alternatives. Yet heteroscedasticity is an important concept." With this I agree, which is why I think McCulloch's early attempt at settling an important issue of orthography in the profession is commendable and why I believe documenting its philological development remains relevant, especially as an example of how knowledge is diffused within the scientific community and eventually to popular literature.[25]

Based on the millions of books digitized by the Google Books project, the answer to the question posed in the title, "When did we begin to spell *heteros*edasticity* correctly?", is 1927, when Western Reserve University[26] published *Problems in education*, in which the authors noted—quite serendipitously in our context—that teaching spelling "is an enigma". The specific instance appeared in the following question: "Correlation, concentration, apperception, interest are not strange terms, but what about standard deviations, I.Q.'s, accomplishment quotients, tetrachoric *r*, multiple correlation or heteroskedasticity?" Much earlier, David [1998] had already noted that 1927 was the debut year of *heteroskedasticity* but in a different work, that of Frank M. Weida, who wrote in the *Annals of Mathematics*:

> If the limited mean error of y remains constant for all possible values of x, the connection of y with x is said to be *homoskedastic*; and if the limited mean error of y does not remain constant for all possible values of x, the connection of y with x is said to be *heteroskedastic*. [Weida 1927:303]

For the most part of recent history, however, the *k* variant of the word heteroskedasticity was never mentioned more than its *c* counterpart. It was only in 2001 when it has consistently dominated (with the exception of 2005). If the trend since the turn of the century persist, then we can expect *heteroskedasticity* to remain the dominant spelling in published works in succeeding years. Whether this will influence the spelling of the related words *homoskedasticity*, *heteroskedastic*, and *homoskedastic*, all of which are still outnumbered by their *c* variants, remains to be seen. There are, however, indications that it has already had an effect. Notably, *homoskedasticity* overtook *homoscedasticity* for the first time in 2008.

[25] By any reckoning, it is hard to overlook the institutional contribution of MIT in advancing standard statistical inference in the presence of nonspherical errors, considering the potent "one-two combo" of White [1980] and Newey and West [1987].

[26] The university is now known as Case Western Reserve University after it merged with the Case Institute of Technology in 1967.

This being an article about language, I end by acknowledging my sloppy use of language. Admittedly through my own fault, the reader might surmise that implicit in this discussion is the notion that *heteroskedasticity* is indeed the "correct" spelling. However, orthographic issues are hardly ever black and white (e.g., "a history" or "an history"?). As E. B. White is wont to remind us, "The language is perpetually in flux: it is a living stream, shifting, changing, receiving new strength from a thousand tributaries, losing old forms in the backwaters of time" [Strunk and White 1959].

Dictionaries list *heteroscedasticity* as an acceptable variant and it is certainly still used in modern textbooks both at the undergraduate (e.g., Gujarati and Porter [2008]) and graduate (e.g., Greene [2011]) levels. The textbook of Greene, in particular, remains the standard textbook for a first-year graduate econometrics sequence. It will undoubtedly have an enduring influence on how future economists will spell heteroskedasticity. Moreover, that *Wikipedia* prefers *c* is significant in an age when information is increasingly retrieved from the Internet, especially for the nonpracticioner.[27]

The aim of this manuscript is to be descriptive rather than prescriptive.[28] Despite appearances, it is not my objective to elevate one orthographic manifestation over the other. More modestly, I only describe trends in spelling variations of an important concept in statistics and econometrics. The present *communis opinio doctorum* is that both spellings are correct. No etymological undertaking can overrule that. The real answer to the question in the title, therefore, is that we have spelled (spelt?) it correctly all the time.

University of Wollongong
School of Accounting, Economics, and Finance
Centre for Human and Social Capital Research

Rheinisch-Westfälisches Institut für Wirtschaftsforschung

(I thank Vaia Karapanou for comments with respect to the Greek words in this article, and Dominik Cremer-Schulte and Florent Fremigacci for help with the French ones. Thanks are also due to Ronald Bachmann, Christoph M. Schmidt, Colin Vance, and an anonymous reviewer for constructive suggestions, as well as to Trevor Breusch, Jerry Hausman,

[27] A journalist will more likely head over to *Wikipedia* to learn about heteroskedasticity rather than pick up White [1980]. Indeed, since *Merriam-Webster* and *Wikipedia* both seem to prefer *heteroscedasticity* while the academe is moving toward *heteroskedasticity*, it is conceivable that the orthography in academic journals will be different from popular literature in the future.

[28] McCulloch, Guy Judge (University of Portsmouth), and David E.A. Giles (University of Victoria) are the latter: using *k* is the "proper English spelling", according to McCulloch, "you spell it with a k and not a c!" is the call of Judge [2007], and "Yes, this word should indeed be spelled with a 'k', and not another 'c'," says Giles [2011]. Both Judge and Giles cite McCulloch [1985b]. McCulloch says, however, that as an editor or reviewer, he would never insist on an author spelling these words with a *k*.

Kenneth D. West, and Halbert White for responding to queries about their influences concerning their orthographic preferences. I am extremely grateful to J. Huston McCulloch and Lilian Coronel, who both provided an extended commentary on an earlier draft, which substantially improved the paper. A short editorial based on this article appears in the *Journal of the Royal Statistical Society: Series A* (*Statistics in Society*) **176**(2): 291–293.)

References

Aldrich, J. [2011] "Mathematical words: origins and sources", http://www.economics.soton.ac.uk/staff/aldrich/Mathematical%20Words.htm.

Bauer, L. [2005] "The borderline between derivation and compounding" in: W. U. Dressler, ed., *Morphology and its demarcations: selected papers from the 11th Morphology Meeting, Vienna, February 2004*. John Benjamins Publishing Company.

Bollerslev, T. [1986] "Generalized autoregressive conditional heteroskedasticity", *Journal of Econometrics* **31**(3): 307–327.

Breusch, T. and A. Pagan [1979] "A simple test for heteroscedasticity and random coefficient variation", *Econometrica* **47**(5): 1287–1294.

Carlyle, T. [1858] *History of Friedrich II of Prussia, called Frederick the Great. Book II: Of Brandenburg and the Hohenzollerns (928–1417)*. Project Gutenberg 2008. http://www.gutenberg.org/dirs/2/1/0/2102/2102.txt.

Crawford, J. [2010] "On the future of books", Inside Google Books, http://booksearch.blogspot.com/2010/10/on-future-of-books.html.

David, H. A. [1998] "First (?) occurrence of common terms in probability and statistics: a second list, with corrections", *The American Statistician* **52**(1): 36–40.

Engle, R. F. [1982] "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation", *Econometrica* **50**(4): 987–1007.

Fisher, F. M., V. E. Ferrall Jr., D. Belsley, and B. M. Mitchell [1966] "Community antenna television systems and local television station audience", *The Quarterly Journal of Economics* **80**(2): 227–251.

Fisher, F. M., J. J. McGowan, and D. S. Evans [1980] "The audience–revenue relationship for local television stations", *The Bell Journal of Economics* **11**(2): 694–708.

Giles, D. E. A. [2011] "The second-longest word in the econometrics dictionary", *Econometrics beat: David Giles' blog*, http://davegiles.blogspot.com/2011/03/second-longest-word-in-econometrics.html.

Goldberger, A. S. [1964] *Econometric theory*. Wiley.

Greene, W. H. [2011] *Econometric analysis*. Seventh edition. Pearson Education Limited.

Gujarati, D. N. and D. C. Porter [2008] *Basic econometrics*. Fifth edition. McGraw-Hill.

Hausman, J. A. and P. A. Ruud [1987] "Specifying and testing econometric models for rank-ordered data", *Journal of Econometrics* **34**(1–2): 83–104.

Ioannidis, Y. [2004] "The history of histograms (abridged)" in: M. A. Nascimento, M. T. Özsu, D. Kossman, R. J. Miller, J. A. Blakeley, and B. Schiefer, eds., *Proceedings of the 30th International Conference on Very Large Datasets*.

Judge, G. [2007] "How do you spell 'heteroskedasticity'?" *Guy's econometrics blog*, http://econometricsstuff.blogspot.com/2007/03/how-do-you-spell-hetero skedasticity.html.

Kay, J. [1998] "In defence of endogenous growth theory", http://www.johnkay. com/ 1998/03/02/in-defence-of-endogenous-growth-theory.

Kelley, T. L. [1923] *Statistical method*. The Macmillan Company.

Kim, E. H., A. Morse, and L. Zingales [2006] "What has mattered to economics since 1970", *Journal of Economic Perspectives* **20**(4): 189–202.

Kmenta, J. [1971] *Elements of econometrics*. Macmillan.

McCulloch, J. H. [1985a] "Interest-risk sensitive deposit insurance premia: stable ACH estimates", *Journal of Banking and Finance* **9**(1): 137–156.

McCulloch, J. H. [1985b] "Miscellanea: on heteros*edasticity", *Econometrica* **53**(2): 483.

McFadden, D. [1974] Conditional logit analysis of qualitative choice behavior" in: P. Zarembka, ed., *Frontiers in econometrics*. Academic Press.

McFadden, D. [1987] "Regression-based specification tests for the multinomial logit model", *Journal of Econometrics* **34**(1–2): 63–82.

Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Novak, and E. L. Aiden [2011] "Quantitative analysis of culture using millions of digitized books", *Science* **331**(6014): 176–182.

Miller, J. [2011] "Earliest known uses of some of the words of mathematics", http: //jeff560.tripod.com/mathword.html.

Newey, W. K. and K. D. West [1987] "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix." *Econometrica* **55**(3): 703–708.

Pearson, K. [1904] "Mathematical contributions to the theory of evolution. XIII. On the theory of contingency and its relation to association and normal correlation" in: *Draper's company research memoirs: biometric series (vol. I)*. Dulau and Co.

Pearson, K. [1905] "Mathematical contributions to the theory of evolution. XIV. On the general theory of skew correlation and non-linear regression" in: *Draper's company research memoirs: biometric series (vol. II)*. Dulau and Co.

Schwartzman, S. [1994] *The words of mathematics: an etymological dictionary of mathematical terms used in English*. Mathematical Association of America.

Stigler, S. M. [1986] *The history of statistics: the measurement of uncertainty before 1900*. Harvard University Press.

Strunk, W. Jr. and E. B. White [1959] *The elements of style*. Macmillan.

Van Loco, J. [2007] “Method validation for food analysis: concepts and use of statistical techniques” in S. Caroli, ed., *The determinants of chemical elements in food: applications for atomic and mass spectrometry*. John Wiley & Sons.

Weida, F. M. [1927] “On various conceptions of correlation”, *Annals of Mathematics (Second Series)* **29**(1/4): 276–312.

Western Reserve University [1927] *Problems in education*. Western Reserve University Press.

White, H. [1980] “A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity”, *Econometrica* **48**(4): 817–838.

White, H. [1984] *Asymptotic theory for econometricians*. Fifth edition. Academic Press.

Wooldridge, J. M. [2010] *Econometric analysis of cross section and panel data*. Second edition. MIT Press.