3-22-2013

The Readability of Principles of Macroeconomics Textbooks

Sarah Tinkler
Portland State University

James Woods
Portland State University

THE READABILITY OF PRINCIPLES OF MACROECONOMICS TEXTBOOKS

Sarah Tinkler and James Wood*

Abstract: The authors evaluated principles of macroeconomics textbooks for readability using Coh-Metrix, a computational linguistics tool. Additionally, they conducted an experiment on Amazon's Mechanical Turk Web site in which participants ranked the readability of text samples. There was a wide range of scores on readability indexes both between textbooks and within textbooks. Results from the Mechanical Turk experiment revealed that the Flesch Reading Ease Index does not predict which samples readers will prefer, but readers do prefer samples that are thematically similar, as identified by Latent Semantic Analysis. There were differences in the responses of native and non-native-but-proficient English speakers to the text samples, suggesting that the intended audience is an important determinant of readability.

Keywords: Coh-Metrix, economics textbooks, Mechanical Turk, readability

JEL codes: A2, A20, A21

Prepublication version of "The Readability of Principles of Macroeconomics Textbooks", *The Journal of Economic Education*, Vol. 44, No. 2. Published version available from publisher website at http://dx.doi.org/10.1080/00220485.2013.770345

*Sarah Tinkler (e-mail: tinkler@pdx.edu) is a professor of economics and James Woods (e-mail: woodsj@pdx.edu) is an assistant professor of economics, both at Portland State University. Tinkler is the corresponding author.

Perhaps the most entrenched tradition of all in economics teaching is the undergraduate textbook. In surveys of economics instructors over a 15-year period, Becker, Watts, and Schaur found that respondents almost always use a textbook when teaching principles of economics (Becker and Watts [1998, 2008], Watts and Schaur 2011). The desire to choose a good textbook is often a key component of the planning that goes into teaching an economics course. But choosing a textbook can be difficult for a number of reasons, not the least of which is the diversity of students encountered in the classroom. The typical principles course attracts both prospective majors and students seeking to fulfill one requirement or another in their course of study. While the instructor's sympathies may lie with future majors whose enthusiasm and aptitude for the subject support the selection of challenging course materials, the needs of less motivated students also must be considered. Economics instructors in search of a textbook also need to keep in mind that critical reading scores on the SAT test for both male and female college-bound seniors are currently at their lowest levels since 1972 (College Board 2011, 2).

McConnell observed that the main difference between an economics textbook from the 1940s and a contemporary textbook is the large increase in the number of diagrams which now total "150–200" per book (McConnell 1998, 32). It may be tempting to assume that diagrams add clarity to textbooks for students because diagrams tend to have this effect for economists. However, Larkin and Simon found that diagrams are useful "only to those who know the appropriate computational process for taking advantage of them" (Larkin and Simon 1987, 99). Principles students may lack the skills necessary to benefit from the proliferation of diagrams in their textbooks and may instead rely on the written text to make sense of the material.

Textbooks are clearly important to economics education but the subject has garnered little attention from economists since the 1980s. In light of substantial advances in readability research

coupled with technological advances in processing texts, a re-examination of textbook selection is warranted. We begin our study by introducing the reader to some of the most important concepts in readability research. We then evaluate the readability of 10 of the most popular macroeconomics principles texts using quantitative measures of readability, as well as an experiment conducted online using Amazon's Mechanical Turk (Amazon Mechanical Turk [Web site]). Finally, we provide some practical suggestions for economics instructors who may be struggling with selecting appropriate reading materials for their students.

DEFINING READABILITY

Readability research seeks "ways to enhance and improve factors which connect reader and writer in a text" (Horning 1991, 35). Of the numerous readability measures available, the Flesch-Kincaid Grade Level Score is the best known, partly because it is included in word processing programs such as Microsoft Office Word.[1] Many organizations, including the Department of Defense, require that all documents have a Flesch-Kincaid Grade Level Score below a certain cutoff (10th grade in the case of the Department of Defense). The Flesch-Kincaid Grade Level Score is closely related to the Flesch Reading Ease Score (usually expressed as an index between 0 and 100) because both predict readability using a formula that weights the number of syllables per word and the number of words per sentence (Flesch 1951; Kincaid et al. 1975).[2,3] A study of principles textbooks in the early 1980s using readability formulas available at the time found that *all* the textbooks included in the study were "readable" by undergraduates but, because different readability formulas ranked the books differently, it was impossible to say which book was the *most* readable (McConnell 1982, 1983). In this study, we substituted the Flesch Reading Ease index for the Flesch-Kincaid Grade Level Score because it has the convenient property that

higher scores on the index mean greater readability, as is the case with the other readability measures used here.

Linguists and educators have always had reservations about the usefulness of Flesch-Kincaid type indexes outside of the population for which they were originally developed, namely young children. For children, unlike older readers, word and sentence length may indeed be important determinants of readability (McConnell 1983). Flesch-Kincaid Scores emphasize simplicity of language and sentence structure but the cohesiveness of a text is more important to readability (Halliday and Hasan 1976, Horning 1991, Irwin 1980). A cohesive text contains linguistic ties that help the reader infer meaning. For example, the word "because" is a cohesive element because it signals a relationship between two concepts. Repetition of words across sentences is another way to add cohesion to a text. Cohesive elements have been found to be more important to "low-knowledge" as opposed to "high-knowledge" readers (McNamara, Louwerse, and McCarthy 2010). Principles students are more likely to be "low-knowledge" readers, which suggests that introductory textbooks should strive for cohesiveness rather than superficial readability in the form of word and sentence length. The Coh-Metrix tool seeks to assign numerical values to the cohesive elements found in texts (University of Memphis [Web site]).

In the end, writers should strive for *coherence*, defined as "the quality of the mental representation constructed by a reader" (McNamara et al. 2006). Underlying all quantitative readability measures, including Flesch-Kincaid, is the assumption that there is a relationship between the measurable properties of a text and its coherence. An emphasis on coherence also means that readability is a subjective concept. While a writer may strive to make a text readable, it is ultimately up to a reader to decide whether or not he or she has succeeded. McConnell,

finding the quantitative measures of readability available at the time unsatisfactory, suggested that readability be assessed by "a direct field test with a group of students for whom the material was intended" (McConnell 1983, 70). Our experiment with crowd-sourcing readability testing on Mechanical Turk is in the spirit of McConnell's suggestion.

ASSESSING READABILITY USING COH-METRIX

We used the latest editions of 10 introductory macroeconomics texts. Sales figures are proprietary information, so popularity was determined using the Amazon Best Sellers tool for the category "Macroeconomics" (Amazon.com). This category includes books other than principles of macroeconomics textbooks, as well as multiple versions and editions of some of the textbooks. We included the first 10 *authors* of principles of macroeconomics textbooks in our study and used their latest edition, even if an earlier edition of the same author's book ranked higher. The 10 textbooks, in alphabetical order, are Arnold (2011); Baumol and Blinder (2012); Case, Fair and Oster (2012); Colander (2010); Cowen and Tabarrok (2012); Hubbard and O'Brien (2013); Krugman and Wells (2013); Mankiw (2012); McConnell, Brue and Flynn (2012); and Miller (2012). Hereafter, the books will be referred to by the name of the first author only.

If readability were to be assessed using sentence-level features of the text only, as is the case with the Flesch-Kincaid Grade Level Score, then the preferred methodology would be to generate several random samples from each text. However, we require conceptually complete samples, i.e., parts of a text that can be understood without the text that immediately precedes or immediately follows. We sampled three topics in each textbook: "GDP" begins with the definition of Gross Domestic Product; "Money" begins at the start of the section that includes the definition of money; and "Unemployment" begins at the start of the section that includes

definitions of the types of unemployment. We selected text-heavy sections of the books with relatively few equations and diagrams because Coh-Metrix is a text processor. These topics are also core subjects in principles of macroeconomics courses and all 10 books include them. The samples vary in length from 400 to 600 words due to the requirement that the samples be conceptually complete. As a caveat, the need to sample text-heavy sections means that we are not assessing the power of the equations and diagrams incorporated in the textbooks to aid in the understanding of the more analytical material. Before processing the samples, we made minor edits to remove decimal points that are read by Coh-Metrix as demarcating the end of a sentence. We also removed hard returns in lists where it was obviously not the author's intent to begin a new paragraph. The 30 text samples are included in the Mechanical Turk Survey that is posted online.[4]

Coh-Metrix Indexes

According to its developers, Coh-Metrix is a "one-stop, computational linguistic department store" (McNamara, Louwerse, and McCarthy 2010, 292). Coh-Metrix evaluates text samples on a multitude of indexes related to readability. In a study of high- and low-cohesion versions of texts from the field of discourse psychology, the developers identified five indexes in Coh-Metrix 2.0 that they found to be most closely associated with readability. We used the identical or nearly equivalent indexes from the September 1, 2012, release of Coh-Metrix 3.0 (D. McNamara, pers. comm.)

Concrete (WRDCNCc in Coh-Metrix 3.0) is the mean concreteness for content words. Concrete words are words for which it is easy to form a mental picture; *house* is a good example of a concrete word. In contrast, abstract words, like *freedom*, are harder to conceptualize. A text with more concrete words than abstract words is presumed to be easier to read. Coh-Metrix

makes use of human-generated concreteness ratings from the *MRC Psycholinguistics Database* (Coltheart 1981). Higher scores on **Concrete** indicate that the text is more concrete and should be more readable.

WordFreq (WRDFRQmc in Coh-Metrix 3.0) is the mean log frequency for the least-frequent content word in a sentence. Frequency refers to how common a word is in the English language. For example, the word "beer" is common compared to the word "absinthe." (Beer ranks 2,927th, and absinthe ranks 85,991st in the Word Count ranking of English words by frequency of use (Word Count.org). An example in which a person buys a beer is likely to be understood more quickly by a student than an example in which a person buys a glass of absinthe. Calculating the score for the least frequent word in a sentence has a theoretical basis in the idea that it is the most unusual word in a sentence that determines its difficulty (McNamara, Louwerse, and McCarthy 2010). The *log* of the score for the least frequent content word is used because the ranking of a word is inversely proportional to its frequency, i.e., it follows a Zipfian distribution. Coh-Metrix derives frequency scores from *Celex*, a linguistic database containing over 160,000 words from 284 written texts (Baayen, Piepenbrock, and Gulikers 1995). Higher scores on **WordFreq** mean that the text should be easier to read. **Concrete** and **WordFreq** both attempt to quantify the complexity of the language used in a text sample. The next three indexes measure cohesive ties. Larger values for the three cohesive indexes described below indicate higher levels of cohesiveness in a sample of text.

The presence of connectives is important for cohesion because they help readers to "join the dots." The index **CausalRatio** (SMCAUSr in Coh-Metrix 3.0) is the ratio of causal particles to causal verbs plus 1. (The addition of 1 to the denominator is necessary because there may be no causal verbs.) Examples of causal particles are "*because*," "*since*," and "*thus*." A causal verb

is used to define a relationship between a person or thing and something that happens. (Example: Hyperinflation *causes* social unrest.) Coh-Metrix identifies causal verbs using WordNet, a lexical database that categorizes words based on semantic characteristics (Fellbaum 1998, Miller 1995). A high score for **CausalRatio** means the writer is making more connections for the reader and this should enhance cohesion.

Writers are often encouraged to use synonyms in order to make their writing more vivid but consistency of terminology (i.e., the avoidance of synonyms) is helpful when a text is challenging. Referential cohesion occurs when an argument (a noun, pronoun, or noun phrase) is repeated throughout a text. The Coh-Metrix index **ArgOverlap** (CRFAO1 in Coh-Metrix 3.0) is the mean score for the repetition of an argument between adjacent sentences. For any pair of sentences, **ArgOverlap** has a value of 1 if repetition of an argument occurs and a value of 0 if no repetition occurs. Repeated use of the word "argument" in the second and third sentences of this paragraph is an example of argument overlap. A higher score for **ArgOverlap** is associated with greater readability.

The Coh-Metrix index **LSA** (LSASS1 in Coh-Metrix 3.0) computes mean Latent Semantic Analysis cosines for adjacent sentences and provides a measures of how conceptually similar each sentence is to adjacent sentences (Deerwester et al. 1990, Landauer and Dumais 1997, Landauer, Foltz, and Laham 1998). The technique of Latent Semantic Analysis uses information about the proximity of words to infer meaning about the relationships among words. While the technique itself is complex, the intuition is simple; LSA permits identification of groups of words that are conceptually similar even if the words themselves are superficially different. For example, in a text about a brewery, **LSA** will detect that "beer," "brewery" and "hops" are thematically similar. Higher values of the index **LSA** imply higher levels of cohesion.

Coh-Metrix Results

Readability metrics are based on models of readability in much the same way that a utility function is a model of preferences. We wouldn't expect different readability metrics to rank the text samples in the same way because they measure different dimensions of readability. Table 1 shows results for the readability indexes used in this study. A radar plot of the data contained in this table can be viewed by following the link in the note to table 1. In addition to the Coh-Metrix indexes described above, we include the Flesch-Kincaid Reading Ease Index (**FleschEase**) as a point of comparison. (Recall that *higher* numbers mean greater readability on the Flesch-Kincaid Reading Ease Index, in contrast to the Flesch-Kincaid Grade Level Score where lower numbers mean greater readability.) All index values are normalized to between 0 and 1.

[Insert table 1 about here]

It is notable from table 1 that the text samples differ greatly in their scores on the different readability measures. For example, Krugman GDP has relatively low scores on **FleschEase**, **Concrete** and **WordFreq**, but is highly readable according to **LSA**, **ArgOverlap** and **CausalRatio**, which are cohesiveness measures. We attempt to determine which measures are the "best" measures of readability in the next section. Another observation is that the three samples from the same textbook are not consistent in terms of readability. As an example, Baumol Money performs well on some measures but Baumol Unemployment scores poorly on all measures of readability. Readability is not consistent across a textbook, perhaps because most textbooks have multiple authors, but it is also probably the case that some topics are inherently more difficult to explain than others. As evidence in support of the idea that some topics are harder to explain, the indexes for the Unemployment samples are nearly all lower than the

indexes for Money and GDP across all textbooks. Finally, **FleschEase** does not appear to be correlated with any of the other quantitative readability measures so it may not be a good summary measure of the different components that go into readability.

ASSESSING READABILITY USING MECHANICAL TURK

Coh-Metrix is concerned with objective characteristics of the text samples but readability is, in the final analysis, inherently subjective because it relates to a particular reader's experience with a text. We address the question of readability directly by running an experiment on Mechanical Turk in which participants were asked to read and evaluate the text samples. Mechanical Turk provides businesses and researchers with a "scalable and on demand" labor force (Amazon Mechanical Turk [Web site]). It was originally created to help with tasks called "human intelligence tasks" (HIT) that are trivial for a human but complicated for a computer. For example, Amazon uses Mechanical Turk to vet pictures of products on the Web site to ensure that the correct pictures are attached to the correct products. Aside from the commercial applications for Mechanical Turk, researchers from many disciplines, including linguistics, are active on Mechanical Turk (Rennekamp 2012, Schnoebelen and Kuperman 2010, Munro et al. 2010).

Mechanical Turk workers are shown a list of available tasks and the compensation (often measured in pennies) for the completion of tasks. Workers choose which tasks they wish to complete. The workers have been, until very recently, mostly middle-income and based in North America, but there has been an increase in the number of Indian workers on the site recently (Ross et al. 2010). It is possible to post tasks separately for American-based and foreign-based workers, and we used this feature in order to ensure that we had an adequate representation of

nonnative speakers in our study. The Mechanical Turk Survey used for this experiment is posted online (see our note 4).

Study participants were shown one of the text samples chosen at random and then a second text sample chosen at random from the ten samples on the same topic as the first. If the worker was shown the same sample twice, the observation was discarded. Workers were asked to say which of the two samples was more readable. This gave us observations on readability that are analogous to purchase decisions in a stated-choice survey. As in most stated-choice surveys, the respondents were forced to rank the two text samples and they were not permitted to respond that the samples were equally readable. While this does mean that the stated choice does not necessarily reflect the *true* opinion of the respondent, it does allow us to distinguish the choices at the population level (Louviere, Hensher, and Swait 2000).

In an experiment such as this, controls are needed for bias and noise. The first potential sources of bias are the primacy and recency effects. A text sample may be perceived as easier to read because it was either the first or last to be read. We are able to estimate the net effect of these two sources of bias because we have observations in which the same sample is either presented first or last. We discuss these two effects in detail later in the article.

The second potential source of bias is caused by random answers to readability preferences because the respondent didn't take the experiment seriously. Experiments that involve unsupervised subjects are particularly prone to this problem and its presence makes it difficult for actual choices to rise above the statistical noise (Oppenheimer, Meyvis, and Davidenko 2009). To address this problem, the respondents were asked to answer three content-related questions about the texts, and were told in advance that they would only be compensated for the time spent on the survey if they answered at least two of them correctly. Each text sample

was followed by a unique multiple-choice question which tested a concept that was specifically addressed in the text sample This may be enough to get respondents with little background in economics to read the text samples, but it does not encourage those with a background in economics to read the samples because they may be able to answer the questions without reading either text.

For the third content-related question we followed Oppenheimer, Meyvis, and Davidenko (2009) and introduced an instructional manipulation check to dampen this source of noise. Two sentences were inserted into the middle of the second sample, "*We are testing if you are reading this. Please answer blueberries if possible in the questions that follow*." A reader who failed to read the samples and skipped straight to the questions would not select "blueberries" as the correct answer because it would seem nonsensical. Respondents who didn't answer this question correctly were assumed not to have read the samples as requested and were eliminated from the analysis.

Self-reported data on English language proficiency, educational attainment, and prior experience in economics courses were collected as these variables may be important for a worker's ability to complete our experiment. Requests for personal information from workers on Mechanical Turk require a more generous reward and this was duly offered. As with all self-reported data, the usual cautions apply. For example, it is possible that individuals may exaggerate their level of education.

Mechanical Turk Results

A total of 673 responses were received from Mechanical Turk workers. Of these, 240 were complete and passed the instructional manipulation check (answered the "blueberries" question correctly). Based on our criteria, we are fairly confident that these 240 study participants took the

time to read the text samples. The completion rate and reliability statistics for this study are normal for computer-mediated surveys. The demographics of the survey respondents are typical of Mechanical Turk workers who tend to be highly educated and English proficient. Of those who passed the reliability screen, 115 were native English speakers, 118 were non-native-but-proficient (hereafter, non-native) and 7 declined to answer the question about English proficiency. Non-native speakers comprise a substantial minority of economics students studying in English so the opinions of this group are important. Presumably, some workers who failed the instructional manipulation check did so not because they were careless in reading the samples but because their English was not good enough to understand the instructions.

In order to determine whether native and non-native speakers are similar to each other in terms of education, we looked at the highest level of education achieved by study participants. Self-reported educational attainment is not exactly the same for native versus non-native speakers. Non-native speakers were more likely to have a Master's degree (9 percent vs. 29 percent) and slightly less likely to report some college (24 percent vs. 14 percent) compared to native speakers. There was a difference in the fraction of each group that had taken an economics course, with 62 percent of the native speakers having done so, but only 41 percent of the non-native speakers having done so. As reported later, prior experience in economics and education level did not affect study participants' responses concerning the readability of the text samples.

Testing for reader preferences among the text samples would have been simple if each of the treatment pairs within a subject was presented to the respondents an equal number of times, alternating between which of the samples was shown first and second. This is not practical in a study with a total of 30 text samples, which means that there are 300 within-topic pairs of text

samples. Each pair would require, at a minimum, 100 replicates to determine which was more readable. Our much smaller sample required a different approach.

We had the following pieces of information about the samples. We knew which samples readers preferred when presented with two samples on the same topic. We knew the order in which the samples were shown to the readers and, from the Coh-Metrix study; we had data on the six readability measures for all the samples. This allowed us to form a discrete choice model using the Coh-Metrix variables to determine how likely it is that the respondent would evaluate the second text as being more readable than the first. The discrete choice model is implemented as a logit, where the explanatory variables are the differences in the individual Coh-Metrix indexes of the two texts, and an intercept term that captures the overall recency or primacy effect. The parameters on the Coh-Metrix indexes within the logit model can be interpreted as the change in the log odds of choosing the second text as being more readable, because the second text had a higher score for one of the readability indexes.

Table 2 shows the results of a simple model that uses only characteristics of the texts, i.e., the readability indexes, as determinants of preferences. There was some variation in the rate at which respondents detected the instructional manipulation check. Participants seemed to find it harder to detect the “blueberries” instruction in the Money samples. Because of this problem, all estimates are corrected with a common post-sampling stratification technique, essentially weighting each observation by the inverse of the probability of passing the instructional manipulation check.

[Insert table 2 about here]

The estimated value of the intercept term, −0.344, is statistically significant at nearly the 1-percent level. What this indicates is that the respondents were less likely, on average, to say

that the second text was more readable. In many studies this would be thought of as a simple primacy effect, a bias in favor of the first sample simply because it was seen first. In this case, however, we believe that the bias arises as a result of the instructional manipulation check that was embedded in the second text sample. Recall that to check whether workers actually read the two samples, we embedded two sentences in the second text directing the respondent to choose the answer that contained the word "blueberries" when answering the multiple-choice questions, even though choosing this answer would have made no sense to someone skipping the text samples and going straight to the questions. This two-sentence statement strikes a discordant note within a text that is otherwise concerned with explaining an economic concept; it reduces the cohesiveness of the text. In a test run of the Mechanical Turk experiment, we placed the instructional manipulation check in the *first* sample and found a very strong *recency* effect (readers preferred the second sample shown to them). From this we conclude that workers prefer the text that does not contain the instructional manipulation check because it reduces the cohesiveness of the sample. There may indeed be primacy and recency effects, but these effects are swamped by the reduction in cohesion produced by the unavoidable need to use an instructional manipulation check. This point is perhaps minor, but it does indicate how sensitive readers are to instances of incoherence in texts.

The remaining parameters show the increase in the log odds of choosing the second text as being more readable because of a higher score on that particular measure. There are three results that are particularly interesting. First, **FleschEase** has no predictive power. While this index may be useful in assessing reading materials for young children, we didn't find it useful in the context of college-level introductory macroeconomics texts. This result is consistent with

work in the field of linguistics that suggests that Flesch-Kincaid may actually send a false signal (McNamara, Louwerse, and McCarthy 2010).

Second**, LSA,** which is a measure of how often adjacent sentences are thematically similar, has strong predictive capability in terms of identifying those texts which are found to be more readable by our respondents. Finally, **ArgOverlap**, which measures consistency of language at the superficial level, is estimated to *reduce* the log odds ratio of choosing the second text as being most readable by −2.300 for every index point that the second text exceeds the first. This means that having a high score on **ArgOverlap** reduces the readability of a text sample, according to the study participants. This may seem counterintuitive, but keeping language consistent and simple has been found to have an ambiguous effect on readability, depending on the reader. High-knowledge readers are sometimes annoyed by high-cohesion texts and they get more enjoyment out of low-cohesion texts that force readers to make more inferences about the content (McNamara et al. 1996).

The results presented in table 2 are predicated on the appropriateness of pooling the native and the non-native speakers. The next model tests if the non-native speakers' reactions to the text samples are different from the reactions of the pooled sample (which includes the non-native speakers). This formulation of the model allows us to easily show differences between the two groups with a regression-based test. The disadvantage of this approach is that it understates the significance of any differences between the two groups of speakers. Results are given in table 3.

[Insert table 3 about here]

Several differences between the native and non-native speakers are immediately apparent. It is clear that non-native speakers have a different reaction to the order in which the

samples are read. The grand mean still shows a primacy bias, −0.654, demonstrating a preference for the first text that was not manipulated, but the deviation from that mean by the non-native speakers, 0.554, reverses this primacy bias. There are a couple of reasons why this might be the case. The limits of working memory when operating in a second language may mean that non-native speakers have a harder time recalling the first text and therefore they may err on the side of preferring the second text since they are able to recall it better. The fact that English is not their native language may make them more tolerant of the jarring effect of finding the instructional manipulation check in the second sample.

Increases in **ArgOverlap** (common words in adjacent sentences) continues to reduce the likelihood that a sample will be seen as more readable as it reduces the log odds of choice by −5.898. The deviation from the pooled reaction by the non-native-but-proficient speakers is +4.716. This value is not quite statistically significant at the 10-percent level (0.127) but because our formulation of the model creates a bias in favor of understating the significance of a result, it does suggest a real difference in the importance of **ArgOverlap** to non-native speakers—these readers, unlike native speakers, like texts with more repetition of content words. Native and non-native speakers alike prefer texts that have high scores for thematic consistency (**LSA**), and their responses are indistinguishable from each other, which can be seen by the low *p*-value of the estimate for the non-native speakers' deviation from the mean response of 0.473.

The final significant difference in how the native and non-native speakers assess readability is in the role that **WordFreq** plays. The pooled result, 2.512 (significant at slightly below the 1-percent level, 0.013) shows all readers having a preference for more common words. Non-native speakers deviate from this assessment; there is a reduction in the log odds, −3.639, of interpreting a text as being more readable when it is composed of more common words. This

may seem counterintuitive but it is worth remembering that some of the more common words in English are the least well-defined. One might write an entire book about the meaning of the word 'be,' the second most common word in English, while the word 'heterogeneous' can be defined in one sentence.

In summary, non-native speakers like repetition of content words (**ArgOverlap**) between sentences, unlike native speakers. Along with native speakers, non-native speakers have a preference for thematic similarity between sentences, as evidence by the performance of **LSA** in the model. They have a much higher tolerance, even a preference, for rarer words than native speakers. Their recency preference may be caused by the limits of working memory capacity in a second language, or it also may reflect a greater tolerance for the lapse in cohesion caused by the introduction of the instructional manipulation check.

DISCUSSION

In this study, we evaluated principles of macroeconomics textbooks in a way that takes advantage of progress in the field of computational linguistics, along with technological advances in survey administration. We were able to revisit a suggestion by McConnell from the early 1980s that textbooks be subjected to a reading test to find the most readable. We did succeed in identifying some of the important determinants of readability, especially with respect to native and non-native speakers but, like McConnell, we were not able to say which of the popular books we studied is the most readable. We believe that while some books are more readable than others, audiences for these books will vary across institutions and that a book should be chosen that reflects these variations.

One of the most interesting parts of this study was the opportunity to conduct a very careful review of 10 of the most popular principles of macroeconomics books. Many instructors

are rather cynical about textbooks and have come to believe that the relentless pressure for sales has propelled all the textbooks into some amorphous middle ground such that it really doesn't matter which book one chooses because they are all "cookie cutter" clones of each other. We respectfully disagree with this assessment of the state of the textbook market and urge all instructors to give many authors a close read before adopting their next textbook. Furthermore, although not a topic of study here, we feel that the tone and the ideology of the different authors come through at times in the books and this may be a concern for some instructors.

In terms of practical advice for instructors who need to select a new book, it is important to completely discount the Flesch-Kincaid level of college textbooks; it is just not helpful in determining the readability of college-level books. The most consistently important Coh-Metrix variable is **LSA,** which relates to the unity of theme maintained throughout a section of text. Frustratingly, there was no one textbook that scored consistently well on this criteria across all topics. The Coh-Metrix study also demonstrated that the readability of a particular book is quite uneven and that some subjects are harder to make readable than others. In our study of three topics, the Unemployment samples were nearly all harder to read than the Money and GDP samples. For this reason, instructors should spend considerable time reading many sections of a textbook and be sensitive to the fact that students will find their textbooks more useful for some subjects than others.

One significant finding of this study is that non-native speakers have preferences different than native speakers. Instructors should bear this in mind if they teach classes with large numbers of non-native speakers. The thematic unity measured by **LSA** is appreciated by both native and non-native speakers, but the non-native speakers like repetition of content words, **ArgOverlap,** and the native speakers do not. We speculate that this is because the pooled group

may contain proficient readers who are annoyed by this feature of a text. (Anyone who has read the same children's book hundreds of times will appreciate the sometimes irritating effect that excessive clarity can have on a reader.) Perhaps surprisingly, the non-native speakers showed a greater preference for text samples with unusual words, although this may be because such words are easier to define.

We highlight the differences between native and non-native speakers, but other observable and unobservable characteristics of students will contribute to the degree to which they find particular textbooks useful. Readers bring many things to a reading experience, including aptitude, ability, language background, and learning style. Some disciplines have moved away from a required textbook, but since our discipline shows no signs of breaking its dependency on the required textbook, instructors will continue to juggle the competing demands of different populations within their classrooms.

NOTES

[1] Flesch-Kincaid Grade Level Score = (.39 × ASL) + (11.8 × ASW) − 15.59, where ASL = number of words/number of sentences and ASW = number of syllables/number of words.

[2] Flesch originally postulated the Flesch Reading Ease Score scaled between 0 and 100 with higher numbers indicating a greater degree of readability. The Flesch-Kincaid Grade Level Score was developed by the U.S. Navy as a measure of readability that corresponded to the reading levels of students in U.S. schools. A Flesch-Kincaid Grade Level Score at the 10th grade level means that the median 10th grade student in the United States should be able to read it. The two measures are inversely related; although the weights assigned to average sentence length and average syllables per word differ. We use the Flesch Reading Ease Index because it is convenient that larger values indicate greater readability, as is the case with the other readability indexes.

[3] Flesch-Reading Ease Score = 206.835 − ((1.015 × ASL) + (0.846 × ASW X 100)), where ASL = number of words/number of sentences and ASW = number of syllables/number of words.

[4] https://s3.amazonaws.com/Readability_of_Principles_of_Macroeconomics_Textbooks/Apendix_Mechanical_Turk_Survey_Including_Text_Samples.pdf

REFERENCES

Amazon.com. Amazon Best Sellers in Macroeconomics. http://www.amazon.com/Best-Sellers-Books-Macroeconomics/zgbs/books/2596/ (accessed July 22, 2012).

Amazon Mechanical Turk. https://www.mturk.com/mturk/welcome?state=WGFMYnM5U1p0MDVRRy9HUWgybTJySXVVNXE0PTIwMTIwMjE4MTgxMFVzZXIudHVya1NlY3VyZX50cnVlJQ (accessed September 12, 2012).

Arnold, R. A. 2011. *Macroeconomics*. 10th ed. Mason, OH: Cengage Learning. http://instructors.coursesmart.com/macroeconomics/arnold/dp/9780538452878.

Baayen, R. H., R. Piepenbrock, and L. Gulikers. 1995. *The CELEX Lexical Database*. Philadelphia: University of Pennsylvania, Linguistic Data Consortium. CD-ROM.

Baumol, W. J., and A. S. Blinder. 2012. *Macroeconomics: Principles and policy*. 12th ed. Mason, OH: Cengage.Learning. http://instructors.coursesmart.com/macroeconomics-principles-and-policy/baumol-blinder/dp/9780538453653.

Becker, W. E., and M. Watts. 1998. *Teaching economics to undergraduates: Alternatives to chalk and talk.* Cheltenham, UK: Edward Elgar Publishing.

———. 2008. A little more than chalk and talk: Results from a third national survey of teaching methods in undergraduate economics courses. *Journal of Economic Education* 39: 273–86.

Case, K. E., R. C. Fair, and S. C. Oster. 2012. *Principles of macroeconomics*. 10th ed. Boston: Prentice Hall. http://instructors.coursesmart.com/principles-of-macroeconomics-tenth-edition/karl-e-case-ray-c-fair-sharon-m-oster/dp/9780131391413.

Colander, D. 2010. *Macroeconomics*. 8th ed. New York: McGraw-Hill. http://instructors.coursesmart.com/macroeconomics-8th-edition/colander-david/dp/0077333330.

College Board. 2011. *Total group profile report*. New York: The College Board. http://professionals.collegeboard.com/profdownload/cbs2011_total_group_report.pdf.

Coltheart, M. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology* August 33A: 497–505.

Cowen, T., and A.Tabarrok. 2012. *Modern principles of macroeconomics*. 2nd ed. New York: Worth. http://instructors.coursesmart.com/modern-principles-of-macroeconomics-second/cowen-tyler-tabarrok-alex/dp/9781429239981.

Deerwester, S. S., T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41: 391–407.

Fellbaum, C., ed. 1998. *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Flesch, R. 1951. *How to test readability*. New York: Harper & Row.

Halliday, M. A. K., and R. Hasan. 1976. *Cohesion in English*. London: Longman.

Horning, A. 1991. Readable writing: The role of cohesion and redundancy. *Journal of Advanced Composition* 11 (Winter): 35–45.

Hubbard, R. G., and A. P. O'Brien. 2013. *Macroeconomics*. 4th ed. Upper Saddle River, NJ: Prentice Hall. http://instructors.coursesmart.com/macroeconomics-fourth-edition/r-glenn-hubbard-anthony-patrick-o-brien/dp/9780132832250.

Irwin, J. W. 1980. The effect of linguistic cohesion on prose comprehension. *Journal of Reading Behavior* 12 (Winter): 323–32.

Kincaid, J. P., R. P. Fishburne, Jr., R. L. Rogers, and B. S. Chissom. 1975. *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Research Branch Report 8-75, U. S. Naval Air Station Memphis. Millington, TN: Naval Technical Training Command.

Krugman, P., and R. Wells. 2013. *Economics*. 3rd ed. New York: Worth. http://instructors.coursesmart.com/economics-third-edition/krugman-paul-wells-robin/dp/9781464103865.

Landauer, T. K., and S. T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* 104: 211–40.

Landauer, T. K., P. W. Foltz, and D. Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes* 25: 259–84.

Larkin, J. H., and H. A. Simon. 1987. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science* 11: 65–99.

Louviere, J. J., D. A. Hensher, and J. D. Swait. 2000. *Stated choice methods: Analysis and applications*. Cambridge, UK: Cambridge University Press.

Mankiw, N. G. 2012. *Principles of macroeconomics*. 6th ed. Mason, OH: Cengage Learning. http://instructors.coursesmart.com/principles-of-macroeconomics/mankiw/dp/9780538453066.

McConnell, C. R. 1982. Readability formulas as applied to college economics textbooks. *Journal of Reading* 26(1): 14–17.

———. 1983. Readability: Blind faith in numbers? *Journal of Economic Education* 14: 65–71.

———. 1998. Reflections on the principles course. In *Teaching undergraduate economics: A handbook for instructors*, ed. W. B. Walstad and P. Saunders, 31–42. Boston: McGraw-Hill.

McConnell, C. R., S. L. Brue and S. M. Flynn. 2012. *Macroeconomics*. 19th ed. New York: McGraw-Hill. http://instructors.coursesmart.com/macroeconomics-19th-edition/mcconnell-campbell-brue-stanley-flynn-sean/dp/0077337875.

McNamara, D. S., E. Kintsch, N. Butler Songer and W. Kintsch. 1996. Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction* 14(1): 1–43.

McNamara, D. S., M. M. Louwerse, and P. McCarthy. 2010. Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Process: A Multidisciplinary Journal* 4: 292–330.

McNamara, D. S., Y. Ozuru, A. C. Graesser, and M. Louwerse. 2006. Validating Coh-Metrix. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, ed. R. Sun and N. Miyake, 573–78.. Mahwah, NJ: Erlbaum.

Miller, G. A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38(11): 39–41.

Miller, R. L. 2012. *Economics today: The macro view*. 16th ed. Upper Saddle River, NJ: Prentice Hall. http://instructors.coursesmart.com/economics-today-the-macro-view-sixteenth/roger-leroy-miller/dp/9780132554527.

Munro, R., S. Bethard, V. Kuperman, V. T. Lai, R. Melnick, C. Potts, T. Schnoebelen, and H. Tilya. 2010. Crowdsourcing and language studies: the new generation of linguistic data. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 122–30. Stroudsburg, PA: The Association for Computational Linguistics.

Oppenheimer, D. M., T. Meyvis, and N. Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45: 867–72.

Rennekamp, K. M. 2012. Processing fluency and investors' reactions to disclosure readability. *Journal of Accounting Research* 50(5): 1319–54.

Ross, J., L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. 2010. Who are the crowdworkers? Shifting demographics in Mechanical Turk. *alt. CHI session of CHI 2010 extended abstracts on human factors in computing systems proceeding*, Atlanta, GA, April 10–15, 2863–72. New York: Association for Computing Machinery (ACM).

Schnoebelen, T., and V. Kuperman. 2010. Using Amazon Mechanical Turk for linguistic research. *PSIHOLOGIJA* 43(4): 441–64.

University of Memphis. Coh-Metrix and Coh-Git. http://cohmetrix.memphis.edu/cohmetrixpr/metgit.html. Memphis, TN: University of Memphis, Department of Psychology (accessed September 10, 2012).

Watts, M., and G. Schaur. 2011. Teaching and assessment methods in undergraduate economics: A fourth national quinquennial survey. *Journal of Economic Education* 42(3): 294–309.

Word Count.org. www.wordcount.org/main.php (accessed September 13, 2012).

TABLE 1: Readability Indexes for Text Samples with Indexes Normalized between 0 and 1[a,b]

Textbook	FleschEase	Concrete	WordFreq	LSA	ArgOverlap	CausalRatio
			GDP Sample			
Arnold	0.919	0.827	0.802	0.360	0.478	0.187
Baumol	0.888	0.466	0.471	0.141	0.000	0.242
Case	0.849	0.587	0.525	0.830	0.994	0.277
Colander	0.868	0.911	0.327	0.148	0.504	0.250
Cowen	0.772	0.320	0.256	0.379	0.625	0.171
Hubbard	0.784	0.593	0.274	0.691	0.742	0.277
Krugman	0.385	0.492	0.410	1.000	1.000	1.000
Mankiw	0.929	0.532	0.759	0.299	0.393	0.567
McConnell	0.434	0.710	0.013	0.212	0.312	0.507
Miller	0.951	0.272	0.707	0.244	0.432	0.494
			Money Sample			
Arnold	1.000	0.877	0.370	0.164	0.387	0.458
Baumol	0.550	0.820	0.422	0.048	0.614	0.704
Case	0.832	0.663	0.210	0.000	0.323	0.061
Colander	0.875	1.000	0.533	0.244	0.638	0.507
Cowen	0.906	0.852	0.441	0.283	0.330	0.567
Hubbard	0.940	0.735	0.526	0.412	0.470	0.548
Krugman	0.551	0.917	0.565	0.019	0.450	0.567
Mankiw	0.874	0.895	0.296	0.100	0.439	0.368
McConnell	0.746	0.767	0.549	0.010	0.404	0.334
Miller	0.474	0.889	0.000	0.424	0.351	0.000
			Unemployment Sample			
Arnold	0.293	0.360	0.124	0.746	0.552	0.923
Baumol	0.080	0.206	0.341	0.164	0.280	0.235
Case	0.393	0.000	0.905	0.228	0.286	0.398
Colander	0.400	0.422	0.610	0.235	0.404	0.704
Cowen	0.554	0.164	0.535	0.431	0.391	0.381
Hubbard	0.235	0.556	1.000	0.553	0.581	0.768
Krugman	0.717	0.401	0.515	0.347	0.404	0.507
Mankiw	0.581	0.904	0.365	0.437	0.500	0.527
McConnell	0.277	0.544	0.064	0.180	0.066	0.350
Miller	0.000	0.453	0.510	0.566	0.555	0.299

[a]Larger numbers indicate greater readability for all indexes
[b]Radar plot of this table available at
https://s3.amazonaws.com/Readability_of_Principles_of_Macroeconomics_Textbooks/Plot_Readability_Statistics.pdf

TABLE 2: Pooled Native and Non-Native Speakers Preference for Second Text

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−0.344	0.143	−2.410	0.017**
ArgOverlap	−3.000	1.447	−2.073	0.039**
CausalRatio	−0.438	0.503	−0.871	0.385
LSA	6.286	2.411	2.607	0.010***
WordFreq	0.404	0.561	0.721	0.471
Concrete	0.012	0.008	1.518	0.130
FleschEase	0.005	0.016	0.326	0.745
N:	233			

Note. *$p < .1$; **$p < .05$; ***$p < .01$

Table 3: Native and Non-Native Speakers Preference for Second Text

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−0.654	0.219	−2.992	0.003***
NonNative	0.554	0.301	1.838	0.067*
ArgOverlap	−5.898	2.248	−2.623	0.009***
CausalRatio	−0.840	0.814	−1.033	0.303
LSA	8.710	3.838	2.269	0.024**
WordFreq	2.512	0.999	2.515	0.013**
Concrete	0.020	0.011	1.764	0.079*
FleschEase	−0.025	0.026	−0.967	0.335
NonNative:ArgOverlap	4.716	3.075	1.533	0.127
NonNative:CausalRatio	0.771	1.084	0.711	0.478
NonNative:LSA	−3.582	4.988	−0.718	0.473
NonNative:WordFreq	−3.639	1.283	−2.837	0.005***
NonNative:Concrete	−0.010	0.016	−0.617	0.538
NonNative:FleschEase	0.055	0.035	1.551	0.122
N	233			

Note. *$p < .1$; **$p < .05$; ***$p < .01$