Empowering Authors to Diagnose Comprehension Burden in Textbooks

Rakesh Agrawal Sunandan Chakraborty* Sreenivas Gollapudi Anitha Kannan
Krishnaram Kenthapadi
Search Labs, Microsoft Research
Mountain View, CA, USA

ABSTRACT

Good textbooks are organized in a systematically progressive fashion so that students acquire new knowledge and learn new concepts based on known items of information. We provide a diagnostic tool for quantitatively assessing the comprehension burden that a textbook imposes on the reader due to non-sequential presentation of concepts. We present a formal definition of comprehension burden and propose an algorithmic approach for computing it. We apply the tool to a corpus of high school textbooks from India and empirically examine its effectiveness in helping authors identify sections of textbooks that can benefit from reorganizing the material presented.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Data Mining, Knowledge Discovery, Education, Textbooks, Diagnostic Tool for Authors, Comprehension Burden, Concepts

1. INTRODUCTION

Education is known to be a key determinant of economic growth and prosperity [19, 41]. While the issues in devising a high-quality educational system are multi-faceted and complex, textbooks are acknowledged to be the educational input most consistently associated with gains in student learning [36]. Textbooks are the primary conduits for delivering content knowledge to the students and the teachers base their lesson plans primarily on the material given in textbooks [14].

Considerable research has gone into investigating what makes for good textbooks [11, 15, 20, 27, 40]. There has also been work on designing ideal textbook and guidelines and checklists have been proposed for assessing the quality of a textbook [5, 8, 31]. Most education researchers concur that the good textbooks are organized in a systematically progressive fashion so that students acquire new knowledge and learn new concepts based on known items of information [4, 23, 31]. Unfortunately, many textbooks suffer from what Harriet Tyson-Bernstein calls the "mentioning" problem [35] that causes concepts to be encountered before they have been adequately explained and forces students to randomly "knock around" in the book [8]. Indeed, the extensive survey in [6] reports a number of empirical studies in which learning from textbooks was successfully improved by rewriting the text to enhance comprehension.

We propose a diagnostic tool for authors to enable them to mine the content of a textbook to quantitatively assess the comprehension burden that a particular organization of the textbook can impose on the reader due to non-sequential presentation of concepts. A textbook with large comprehension burden makes it harder for the reader to understand the textbook material. We introduce the notion of comprehension burden, present a formal definition, and provide an algorithmic approach for computing it. We evaluate the proposed methodology over a corpus of Indian textbooks that demonstrates its effectiveness for identifying sections of textbooks that can benefit from reorganizing the material presented.

The layout of the rest of the paper is as follows. We begin with a discussion of related work in §2. We present key properties followed in well-written textbooks and provide a formal definition of comprehension burden based on these properties in §3. We also describe the algorithm for computing the comprehension burden in this section. We present the experimental results in §4 and conclude with a summary and directions for future work in §5.

2. RELATED WORK

The question of what factors influence understandability of a text has intrigued researchers for a long time. An early comprehensive investigation, dating back to 1935 [16], identified two principal sets of factors. The first set pertains to individual differences amongst readers, such as levels of intellectual capacity, reading skills, attitudes and goals, previous experiences, and personal interests and tastes. The

*Currently at New York University; work done during internship at Microsoft Research.

KDD'12, August 12–16, 2012, Beijing, China.

second set relates to the readability of the material, which in turn depends on format (page layout, appearance, etc.), organization (headings, indexes, flow, etc.), style (linguistic structural elements, tone of the writer, etc.), and content (theme, nature of the subject matter, etc.). Much of the readability research has focused on the style category because of perceived relative importance of stylistic variables and the fact that stylistic variables are easier to operationalize [18].

We refer the reader to the survey in [12] for overview of readability research. Sherman is considered to be the first to use statistical analysis for analyzing readability in 1890's. By counting average sentence length, he showed how sentence-length averages had shortened over time [32]. The first readability formula, a weighted index of vocabulary complexity, is attributed to the work of Lively and Pressley in 1923 [24]. Since then over two hundred formulas have been developed for measuring the difficulty of reading. Some popular formulas include Flesch Reading Ease Score, Flesch-Kincaid Grade Level, Dale-Chall Grade Level, Gunning Fog Index, SMOG Index, Coleman-Liau Index, and Automated Readability Index. The readability formulas have come under criticism because of their purported low validity from the perspective of psycholinguistic theories [7] and there have been efforts to develop new approaches for predicting reading difficulty, for example, by using statistical language modeling techniques and linguistic features [10, 22] and by devising domain-specific readability measures [42]. While this body of readability research can be used to assess the reading complexity of a piece of text in isolation, our focus is on studying the organization and presentation of concepts in the entire textbook.

In linguistics, cohesion refers to connections between sentences, whereas coherence refers to the connectedness of the ideas [39]. Cohesion provides a sense of flow from sentence to sentence and the principle of cohesion states that one must start a sentence with old information and end it with new information. The principle of coherence states that to make a series of individual sentences into a coherent passage, one must focus the topics of those sentences on a limited number of concepts. Discourse parsing techniques are used to build a structural representation of a text which reflects the semantic relationships among basic textual units [17, 25, 29, 33]. We build upon these ideas in our definition of comprehension burden, where we focus on flow from concept to concept.

The cognitive load theory proposes that poorly designed instructional materials place cognitive overload on learners as working memory is limited. This cognitive overload impairs schema acquisition, later resulting in a lower performance [26, 28]. The theory distinguishes between three types of load: intrinsic, extraneous and germane. Intrinsic load is a function of the complexity of the content rather than instructional design. Extraneous load on the other hand is not inherent within the content, but depends on how the instructional designer has structured and presented information. Intrinsic and extraneous loads are additive. Germane load is the remaining free capacity in working memory that may be directed toward schema acquisition. Thus, a well-designed book increases the learning capacity of a learner. We make use of these ideas in abstracting the properties of well-written books.

In [2], a probabilistic decision model has been proposed for identifying those sections of a textbook that are not well-written. The decision model is based on the syntactic complexity of the writing and the notion of the dispersion of key concepts mentioned in the section. However, each section is treated independently and the flow of writing across different sections is not taken into account. By contrast, the focus of the present paper is on determining whether the entire textbook is organized in a systematically progressive fashion.

3. COMPREHENSION BURDEN

We begin by enunciating properties of well-written textbooks abstracted from education literature. We then describe our model of a textbook and how a poorly written book can impose comprehension burden on the reader. Finally, after introducing some notations used in the paper, we formally define comprehension burden of a textbook.

3.1 Properties of Well-written Textbooks

3.1.1 Organization into Focused Sections

PROPERTY 3.1 (FOCUS). *Each section in a well-written textbook explains very few concepts.*

3.1.2 A Key Section for Every Concept

PROPERTY 3.2 (UNITY). *For each concept in a well-written textbook, there is a unique section that best explains the concept.*

The properties of FOCUS and UNITY are rooted in the following notion: a text that is more unified and addresses a single purpose results in better comprehension [4, 8, 16, 23, 39]. The cognitive load theory also proposes that intrinsic load may be reduced by dividing an instruction into smaller pieces and that if learners process the individual elements of instruction serially, rather than simultaneously, they are able to process the instruction more efficiently [9, 30]. In fact, Alfred North Whitehead admonished way back in 1917: "What you teach, teach thoroughly, seizing on the few general ideas which illuminate the whole, and persistently marshalling subsidiary facts around them" [38].

3.1.3 Sequential Presentation of Concepts

PROPERTY 3.3 (SEQUENTIALITY). *Concepts in a well-written textbook are discussed in a sequential fashion, that is, a concept is adequately explained prior to occurrences of this concept or any related concept.*

This property stems from the following insights: (a) good textbooks should be organized in a systematically progressive fashion [4, 31] (b) a well-designed book can reduce extraneous cognitive load and thus increase the learning capacity of a student [26, 28].

3.1.4 Prioritization of Concepts

PROPERTY 3.4 (PRIORITIZATION). *In a well-written textbook, the tie for precedence in presentation between two mutually related concepts is broken in favor of the more significant of the two.*

Although a well-written textbook strives to present the concepts in a sequential fashion, two mutually related concepts may vie to have precedence in the order of their presentation. A good book will break the tie in favor of the concept that is more significant [23].

3.2 Origin of Comprehension Burden

Although textbooks are typically organized into chapters which are subdivided into logical sections, for the purposes of this paper, we find it sufficient to model a textbook simply as a sequence of sections. Assume that the textbook has been written in such a way that each section *explains* one or very few concepts, that is, each of these concepts is discussed in depth (Property 3.1). But a section can also *mention* other concepts, that is, each of these concepts is referenced but not explained in detail. For each concept, there is a key section in the book that best explains it (Property 3.2). Thus a reader who has gone through the key section corresponding to a concept has comprehended the concept, while a reader who has not yet read the key section has a vague comprehension of the concept. A concept could be related to other concepts, and hence significant for understanding other concepts.

A textbook imposes comprehension burden on the reader if the concepts are not presented in a sequential fashion. In particular, for a concept c, an occurrence of a related concept d in a section preceding the key section for c imposes comprehension burden on the reader since the reader has not comprehended c yet and hence faces difficulty in understanding d. Note that d could be same as c, in which case the comprehension burden is imposed on the reader who encounters a mention of c in an earlier section whereas c has been explained properly in a later section (Property 3.3).

Suppose the reader encounters d in section i prior to explanation of c in its key section k_c. The more significant the explanation of c in the section k_c is for understanding other concepts in the book, the greater chance the reader cannot understand d in section i without thoroughly understanding c and hence larger the comprehension burden on the reader while reading section i. Further, the more significant d in section i is for understanding other concepts in the book, the larger is the comprehension burden since the reader may not be able to follow subsequent material that is based on the discussion of d in section i (Property 3.4).

For illustration, consider a book consisting of three sections, each of which contains three mutually related concepts, c_1, c_2 and c_3. UNITY demands that each concept has a unique key section where it is explained. FOCUS requires that a section explains very few concepts. Assume that a section can be the key section for only one concept. Suppose that c_1 is the most significant and c_3 is the least significant of the three. PRIORITIZATION implies that the author will explain c_1 in section 1, followed by c_2 in section 2, and finally c_3 in section 3. Concept c_1 obeys SEQUENTIALITY but not the other two. Its mention in any of the later sections does not incur comprehension burden as the reader has by now comprehended it. A mention of c_2 in section 1 will incur comprehension burden as it is explained only in section 2. However, its mention in section 3 will not incur comprehension burden. Similarly, any mention of c_3 in a section earlier than section 3 will incur comprehension burden.

As the above example demonstrates, there is a trade-off to be made since the properties interact with each other. In the extreme, a book that consists of a single long section violates FOCUS completely while satisfying the other three. Clearly, this organization is not acceptable. Typically the author faces the task of explaining certain mutually related concepts in a book. As all of them cannot be presented in the same section due to the requirement of FOCUS, SEQUENTIALITY needs to be violated and hence comprehension burden cannot be entirely avoided. However, different organizations can result in different amounts of comprehension burden, which points to the necessity of being able to quantify the comprehension burden due to a given organization of a textbook. We assume that the author first tries to satisfy FOCUS and UNITY as much as possible.

S	Set of sections in a given textbook
C	Set of concepts in a given textbook
$R(c)$	Set of concepts related to concept c
k_c	Key section for concept c
$\lambda(c, i)$	Significance score of concept c in section i
$\psi((d, i) \leftarrow c)$	Comprehension burden for concept d in section i attributed to concept c
$\psi(d \leftarrow c)$	Comprehension burden for concept d attributed to concept c
$\Psi(c)$	Comprehension burden attributed to concept c
$\tilde{\Psi}(i)$	Comprehension burden of section i
$\mathcal{B}$	Comprehension burden of the textbook

Table 1: Notations

3.3 Notations

Before formally defining comprehension burden, we introduce some notations. Let $S = \{1, 2, \ldots, n\}$ denote the set of sections in a given textbook. Let C denote the set of concepts in the book. For each concept $c \in C$, let k_c denote the key section for understanding the concept. For each concept $c \in C$, denote the set of concepts related to it by $R(c)$. Note that $R(c)$ includes c. Let $\lambda(c, i)$ denote the significance score of concept c in section i for understanding other concepts in the entire book.

Let $d \in R(c)$ be a concept related to c. We denote by $\psi((d, i) \leftarrow c)$ the comprehension burden imposed on the reader while reading about d in section i due to c being necessary for understanding d but c being explained in a later section in the book. We say that $\psi((d, i) \leftarrow c)$ is the comprehension burden for concept d in section i attributed to concept c. Similarly, denote the comprehension burden for concept d attributed to concept c over all sections by $\psi(d \leftarrow c)$, the comprehension burden on all concepts attributed to concept c by $\Psi(c)$, the comprehension burden of section i by $\tilde{\Psi}(i)$, and the total comprehension burden of the textbook by $\mathcal{B}$. Table 1 summarizes key notations.

3.4 Definition of Comprehension Burden

We now present the formal definition of comprehension burden for a given textbook, assuming that the book satisfies the FOCUS and UNITY properties.

DEFINITION 3.5. *1. Given a concept $c \in C$ with key section k_c, and a concept $d \in R(c)$ occurring in section i, define the* comprehension burden for concept d in section i attributed to concept c *as*

$$\psi((d, i) \leftarrow c) := \begin{cases} f(\lambda(d, i), \lambda(c, k_c)) & \text{if } i < k_c \\ 0 & \text{if } i \geq k_c, \end{cases}$$

where f is a monotonically increasing function in two variables satisfying $f(x,y) < f(y,x)$ whenever $x > y$.

2. *Given a concept $c \in C$ and a concept $d \in R(c)$, define the* comprehension burden for concept d attributed to concept c *as*

$$\psi(d \leftarrow c) := \sum_{i \in S} \psi((d,i) \leftarrow c).$$

3. *Define the* comprehension burden attributed to concept $c \in C$ *as*

$$\Psi(c) := \sum_{d \in R(c)} \psi(d \leftarrow c).$$

4. *Define the* comprehension burden of section $i \in S$ *as*

$$\tilde{\Psi}(i) := \sum_{d \text{ occurring in section } i} \; \sum_{c:\ d \in R(c)} \psi((d,i) \leftarrow c).$$

5. *Define the* comprehension burden of the textbook *as*

$$\mathcal{B} := \sum_{c \in C} \Psi(c) = \sum_{i \in S} \tilde{\Psi}(i).$$

3.4.1 Explanation of the Definition

The key aspect of Definition 3.5 is to quantify $\psi((d,i) \leftarrow c)$. Occurrence of d in k_c or a section following it does not impose any comprehension burden on the reader since the reader would have understood c before encountering d, and hence $\psi((d,i) \leftarrow c)$ is non-zero only when $i < k_c$. See Fig. 1 for an illustration.

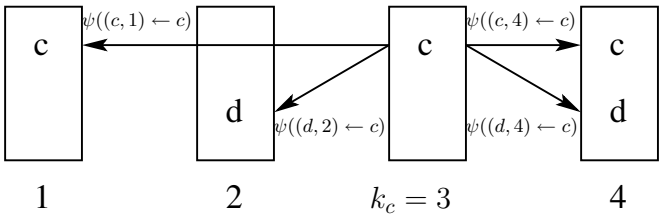


Figure 1: Illustration of comprehension burden: Concept c is explained in section 3 ($k_c = 3$) and is also mentioned in sections 1 and 4. A related concept d occurs in sections 2 and 4. As c is explained in section 3, the reader incurs comprehension burden when reading about c in section 1 ($\psi((c,1) \leftarrow c) > 0$) and about d in section 2 ($\psi((d,2) \leftarrow c) > 0$), but not in section 4 when encountering c and d ($\psi((c,4) \leftarrow c) = \psi((d,4) \leftarrow c) = 0$).

Burden $\psi((d,i) \leftarrow c)$ depends monotonically on the significance score of concept d in section i and the significance score of concept c in section k_c, and has been expressed as a monotonically increasing function $f(\lambda(d,i), \lambda(c,k_c))$. As the reader incurs comprehension burden every time she encounters an occurrence of d in a section prior to k_c, we sum $\psi((d,i) \leftarrow c)$ over all sections to obtain $\psi(d \leftarrow c)$. The comprehension burden attributed to concept c, $\Psi(c)$ is obtained by summing $\psi(d \leftarrow c)$ over all concepts related to c. The comprehension burden of section i, $\tilde{\Psi}(i)$ is obtained by summing $\psi((d,i) \leftarrow c)$ over all concepts that d is related to and then over all concepts in section i. The comprehension burden of the textbook is obtained by summing the comprehension burden attributed to each concept, or equivalently by summing the comprehension burden of each section.

3.4.2 Characterization of f

The requirement on f in Definition 3.5 stems from PRIORITIZATION property.

CLAIM 3.6. PRIORITIZATION *implies that $f(x,y) < f(y,x)$ whenever $x > y$.*

PROOF. Consider two organizations of a book differing in just the order of presentation of two mutually related concepts c and d. In the first version, c is explained in section i with significance score λ_c and d is explained in a later section j with significance score λ_d, where $\lambda_c > \lambda_d$. In the second version, the positions of c and d are interchanged. The first version obeys PRIORITIZATION while the second violates it. Hence the second version should impose larger comprehension burden on the reader. For the first version, $\psi(d \leftarrow c) = 0$ and $\psi(c \leftarrow d) = f(\lambda_c, \lambda_d)$ so that the total comprehension burden is $f(\lambda_c, \lambda_d)$. Similarly the total comprehension burden for the second version is $f(\lambda_d, \lambda_c)$. Thus, the requirement is that $f(\lambda_c, \lambda_d) < f(\lambda_d, \lambda_c)$ whenever $\lambda_c > \lambda_d$. In other words, f should satisfy $f(x,y) < f(y,x)$ whenever $x > y$. □

We next show that a broad category of simple functions can satisfy the above requirement.

CLAIM 3.7. *Suppose f is defined as the product of two univariate functions on either input: $f(x,y) = g(x) \cdot h(y)$. Then the characterization of f in Definition 3.5 is satisfied if $g(.)$ and $h(.)$ are monotonically increasing functions and $h(x)/g(x)$ is also a monotonically increasing function.*

The claim follows since (a) g and h are monotonically increasing $\Rightarrow$ f is monotonically increasing (b) $x > y \Rightarrow h(y)/g(y) < h(x)/g(x) \Rightarrow f(x,y) < f(y,x)$. Stated differently, the requirement above is that $h(.)$ grows faster than $g(.)$, or equivalently the significance score of concept c in section k_c is weighed more than the significance score of concept d in the earlier section i in the computation of $\psi((d,i) \leftarrow c)$. For example, $g(x) = 1$ and $h(x) = x$ is a valid choice that takes only $\lambda(c,k_c)$ into account, ignoring $\lambda(d,i)$.

3.4.3 Generalization of Comprehension Burden Definition

1. We use summation to define $\Psi(c)$ and $\mathcal{B}$, but in general, any function that is dominated by summation can be used instead since different concepts may be related and the marginal comprehension burden due to an additional concept could be less than the comprehension burden due to this concept in isolation.

2. Suppose the book does not satisfy UNITY, so that the explanation of a concept is spread across multiple sections. Then the burden $\psi((d,i) \leftarrow c)$ depends on the explanation of c not covered in sections up to i, and can be obtained by taking into account the significance score of concept d in section i and the significance scores of concept c in sections following i, that is, $\psi((d,i) \leftarrow c) := \sum_{j>i} f(\lambda(d,i), \lambda(c,j))$.

3. FOCUS does not figure in the definition of comprehension burden. The extent to which a textbook follows FOCUS can be measured by computing the average number of concepts to be explained in a section (closer to 1 is better) and the deviation of the number of concepts explained in each section from this average.

3.5 Computing Comprehension Burden

Computation of comprehension burden of a textbook using Definition 3.5 requires the following inputs: (i) concepts in the book, (ii) relationship between concepts, (iii) the significance score for each concept in each section, and (iv) the key section for every concept. We describe next the computation of each of these inputs.

3.5.1 Concepts

Following [3, 21], we define concept phrases to be terminological noun phrases. We first form a candidate set of phrases using linguistic patterns, with the help of a part-of-speech tagger [34]. We adopt the pattern A^*N^+, where A refers to an adjective and N a noun, which was found to be particularly effective in identifying concept phrases. Examples of phrases satisfying this pattern include 'cumulative distribution function', 'fiscal policy', and 'electromagnetic radiation'. The initial set of phrases is further refined by exploiting complementary signals from different sources. First, WordNet [13], a lexical database is used to correct errors made by the part-of-speech tagger. Next, both malformed phrases and very common phrases are eliminated, based on the probabilities of occurrences of these phrases on the Web, obtained using Microsoft Web N-gram Service [37]. The reason for eliminating common phrases is that they would be already well understood.

3.5.2 Relationship between Concepts

We first attempted to induce relationships between concepts by mapping concept phrases to Wikipedia articles and use the link structure between the Wikipedia articles to infer relationship between concepts. We discovered the following issues. Many Wikipedia articles have asymmetric hyperlink structure, plausibly due the encyclopedic nature of Wikipedia: there are relatively less links from articles on specialized topics to articles on more general topics. For instance, the Wikipedia article titled 'Gaussian surface' mentions 'electric field' 11 times but does not have a link to the latter. Furthermore, while Wikipedia provides good coverage for universal subjects like Physics and Mathematics, it has inadequate coverage for concepts related to locale-dependent subjects such as History.

We, therefore, derive the relationship between concepts directly from textbooks using co-occurrence. More precisely, we defined $R(c)$ to be the set of concepts (including c) that co-occur with c in at least e sections such that both c and the co-occurring concept c' occur within a window of size l in each of these e sections. The requirements of co-occurrence in multiple sections and co-occurrence within a window size ensure that we only consider concept pairs that are significantly related to each other. Our implementation sets $e = 2$ and $l = 500$ words.

3.5.3 Significance Scores

The significance score $\lambda(c, i)$ is a measure of how significant is the description of concept c in section i for understanding other concepts in the book. One possible approach would be to define the significance score in terms of the relative frequency of the concept phrase in the section, for example, $\lambda(c, i) := (freq(c, i))/(\sum_{1 \le i \le n} \sum_{c \in C} freq(c, i))$. A problem with such a definition is that it does not differentiate between two concept phrases c_1 and c_2 with the same frequency in a given section. However, concept c_1 may be related to many concepts that occur in other sections in the book and hence more significant for understanding the entire book, while c_2 may be relevant only for the current section. We, therefore, define the significance score of a concept phrase in a section taking into account: (a) how frequent is the concept in the section, and (b) how many concepts are related to it.

DEFINITION 3.8. *The significance score of a concept c in section i is defined as*

$$\lambda(c, i) := \pi(freq(c, i), |R(c)|),$$

where π is a monotonically increasing function in two variables.

Our implementation instantiates function π as $\pi(freq(c, i), |R(c)|) := freq(c, i) \cdot |R(c)|$. This choice has the interpretation that the significance score of a concept in a section is proportional to its frequency in the section as well as the number of concepts related to it.

3.5.4 Key Sections

We use the intuition that the key section for a concept will have high significance score for that concept. Thus, we algorithmically obtain the key section for a concept by comparing its significance scores in different sections. For each concept $c \in C$, we set the key section for c to be

$$k_c := \arg \max_{1 \le i \le n} \lambda(c, i).$$

With the significance score defined as in Definition 3.8, this choice is equivalent to selecting the section where the concept is most frequent.

We remark that multiple alternatives exist for computing the above inputs and our model of comprehension burden can admit multiple such choices. For example, the significance score computation can benefit by including anaphoric references to a concept when computing its frequency in a section.

4. FROM THEORY TO PRACTICE

We next present a diagnostic tool we have built based on the notions just introduced that allows an author to understand the sources of comprehension burden in a textbook. This tool helps authors accomplish the following goals:
(1) identify and investigate sections with large burden, and
(2) identify and probe concepts that impose large burden.

4.1 Corpus

We studied the characteristics of our tool over a corpus of Indian high school textbooks published by the National Council of Educational Research and Training (NCERT). We selected this corpus because millions of students study from these books every year and these books were readily available online. We applied the tool to eleven books from grades IX–XII, covering four broad subject areas: Sciences, Social Sciences, Commerce, and Mathematics. In [1], we

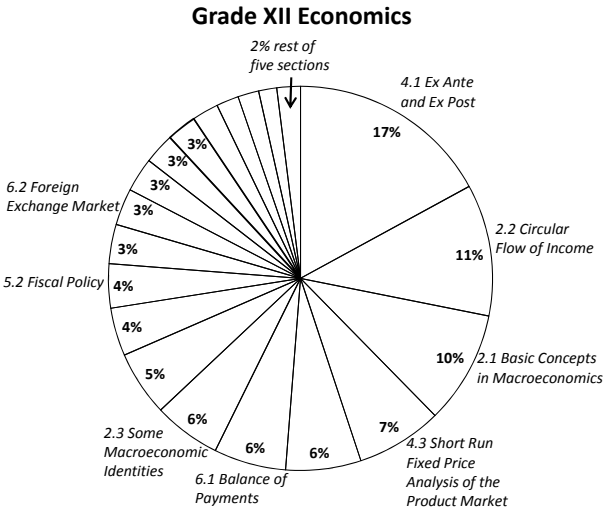


Figure 2: Textbook overview page, with sections listed in the decreasing order of burden. The number within a slice indicates percentage of total burden for the corresponding section. Names of some of the sections have been omitted for clarity.

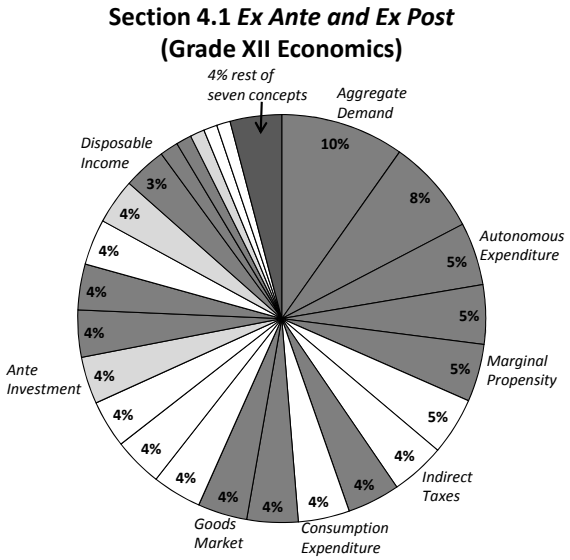


Figure 3: Section page. Concepts occurring in the section are shown in the decreasing order of their burden. Further, concepts that are explained in the section have light gray background color, concepts that are mentioned but explained in a later section have dark gray background, and concepts that are mentioned but have been explained earlier have white background. Names of some of the concepts have been omitted for clarity.

described how to quantify the properties in §3.1 and measured the extent to which these properties are followed in the above books. Here our focus is on measuring the comprehension burden in these books.

In applying the tool, we instantiate function f as $f(\lambda(d,i),\lambda(c,k_c)) := \lambda(c,k_c)$ for computing comprehension burden. This choice satisfies the characterization of f in Definition 3.5 and has the interpretation that the comprehension burden for concept d in section i attributed to c is proportional to the significance of c in its key section.

4.2 Diagnostic Tool

4.2.1 Burden of Sections

The author of a textbook might like to first determine sections where the readers incur large comprehension burden, so that she can prioritize such sections for revision. Hence our tool presents to the author an overview of the book with the sections ordered in the decreasing order of burden. Figure 2 illustrates this page for Grade XII Economics book. From this page, the author can observe that though this book consists of twenty five sections, 75% of the burden arises from just ten sections. Further, Section 4.1 titled "Ex Ante and Ex Post" has the largest burden.

Having identified a section with large burden, the author might next like to understand why the reader incurs large burden in the section. For this purpose, our tool allows navigation to a page created for each section. Figure 3 shows this page for the section, "Ex Ante and Ex Post". This page lists the concepts occurring in the section in the decreasing order of their burden. This page also helps the author to determine whether each of these concepts is *explained* or *mentioned* in the section, and for a concept that is mentioned, whether it has been explained in an earlier section or a later section. Thus, the author can identify those concepts that are mentioned in the section prior to their explanation in the book. For such a concept, the author may choose to include greater explanation in the current section or an earlier section where it occurs. The author can observe that the section "Ex Ante and Ex Post" contains a large number of distinct concepts, of which only a small fraction is explained in this section. About half of the remaining have been explained in earlier sections but the remaining half are explained in later sections. Moreover, 75% of the burden for the section arises from half of the concepts.

Next the author may want to drill down to understand what causes burden for each of these concepts. For this purpose, our tool permits navigation to a separate page dedicated to each concept occurring in the section. This page provides the distribution of the significance score of the concept across different sections in the book. It also shows the related concepts needed for understanding this concept, in the decreasing order of burden they impose on this concept. Figure 4 shows the page for the concept 'aggregate demand', which has the largest burden in the section, "Ex Ante and Ex Post". The author learns that this concept is mentioned in this section, prior to its explanation in a later section. Furthermore, the author sees that the large burden for 'aggregate demand' is caused by the presence of many related concepts (such as 'exchange rate', 'final goods' and 'government spending') that are explained only in later sections.

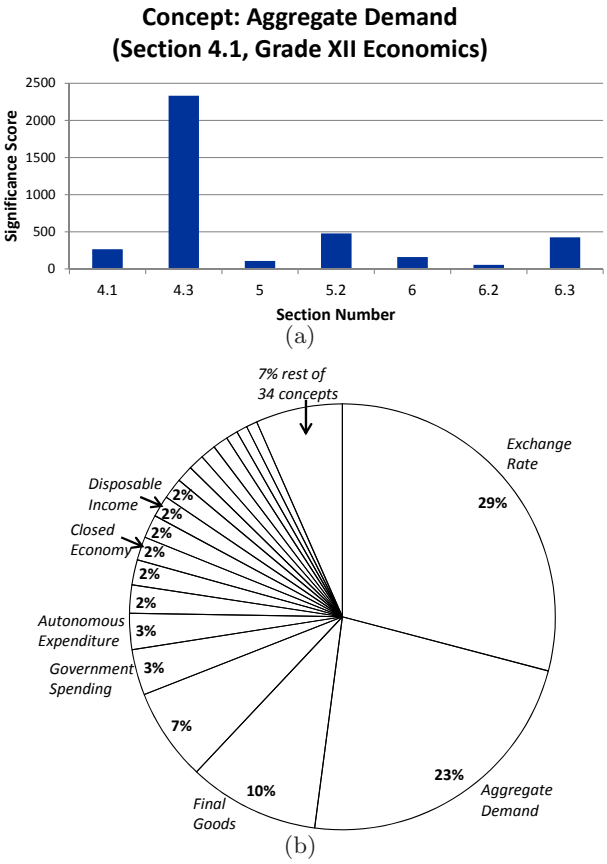


Figure 4: Concept page for a section. (a) The distribution of significance scores of the concept 'aggregate demand' across different sections is shown. (b) Related concepts are shown in the decreasing order of burden they impose on the concept 'aggregate demand' in the section "Ex Ante and Ex Post".

4.2.2 *Burden Attributed to Concepts*

The author may also examine the book along an orthogonal dimension by investigating concepts that impose large burden across the book. Such diagnosis is beneficial since the author may choose to include a glossary of such concepts. For this purpose, our tool presents an analysis of the book with respect to concepts present in the book, ordered in the decreasing order of burden imposed by them. Figure 5 shows this analysis for Grade XII Economics book. The author can see that concepts such as 'aggregate demand', 'exchange rate' and 'final goods' impose the largest burden on the reader, accounting for nearly 55% of the total burden in the book. Moreover, 80% of the burden is attributed to just 10% of the concepts.

For a high burden concept, the author might like to understand its impact on related concepts. This information might help the author decide whether this concept should be explained in an early section in the book. Hence our tool lets the author navigate to a page for each concept, which lists the burden imposed by this concept on its related concepts. Figure 6 illustrates this page for the concept 'aggregate demand'. By probing into concepts related to this concept and their distribution across sections, the author can infer that the burden can be reduced by explaining 'aggregate demand' in a section prior to occurrences of many of its related concepts.

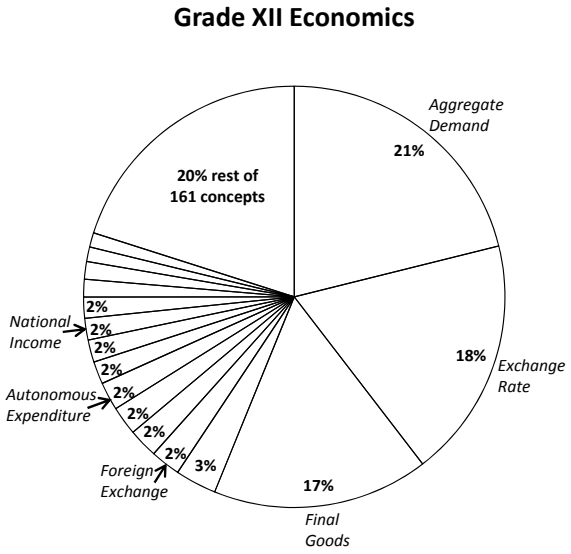


Figure 5: Textbook overview page, with concepts listed in the decreasing order of burden imposed.

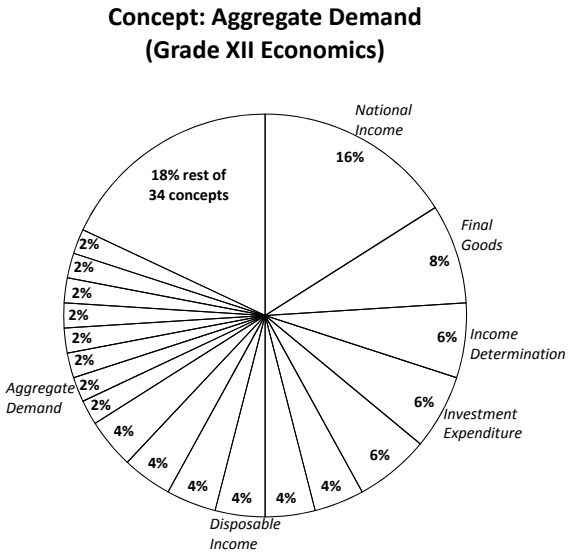
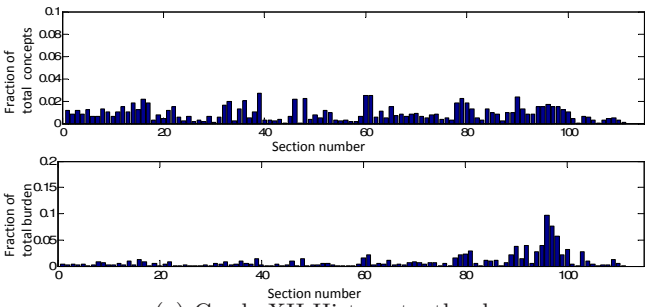


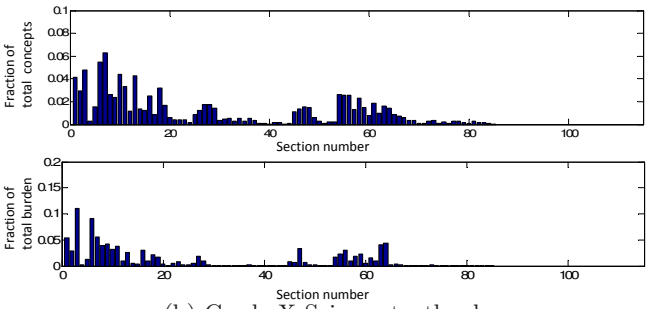
Figure 6: Concept page for the book. Burden imposed by the concept 'aggregate demand' on its related concepts is shown.

4.2.3 *Differences between Books*

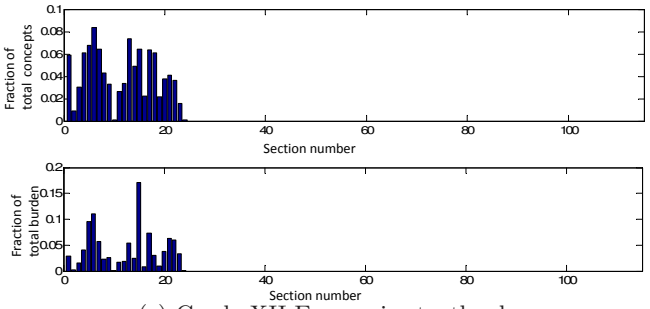
We next present some comparative observations from applying our tool to books belonging to three different subjects: Grade XII Economics, Grade X Science, and Grade XII History. These books also had different organizational structure. Figure 7 gives a statistical overview of the sec-



(a) Grade XII History textbook



(b) Grade X Science textbook



(c) Grade XII Economics textbook

Figure 7: Characteristics of different books.

tions in these books. It shows the distribution of burden across different sections in a book, as well as the distribution of the number of concepts across different book sections. Note that the X-axis refers to the section number (ordered over the entire book) and that these books have different number of sections.

Grade XII History book consists of chapters covering different periods of Indian history, ranging from 3000BC to the 20th century. Our manual inspection revealed that the last few chapters in this book pertain to interrelated topics such as British rule and Indian freedom struggle, while the chapters in the initial two thirds of the book mostly discuss disjoint time periods, and hence disjoint concepts. As a result, the burden for this book arises from very few sections occurring in the last few chapters.

Grade X Science book consists of independent modules, corresponding to Biology, Chemistry, Physics, and Environmental Sciences in that order. However certain concepts such as 'electric current' and 'carbon dioxide' are shared across modules, and are mentioned in the Biology module prior to their explanation in later parts of the book. Moreover, sections in the Biology module contain a large number of concepts. Consequently, the burden is concentrated in the initial part of the book.

By contrast, Grade XII Economics book pertains to the single theme of macroeconomics, with different sections sharing related concepts, and hence the burden is spread out across more sections. We also observe that sections with large burden tend to have a large number of concepts.

5. CONCLUSIONS AND FUTURE WORK

This paper represents our attempt to expand the scope of data mining by considering a new application area – mining textbooks for identifying sections and concepts that can benefit from reorganizing the material presented. Towards this goal, we introduced the notion of comprehension burden, presented a rigorous definition as well as an algorithm for computing it, and provided a diagnostic tool for authors to quantitatively assess the burden that a textbook imposes on the reader due to non-sequential presentation of concepts. We applied the tool to a corpus of high school textbooks, currently in active use in India. Using the tool, we were able to isolate high-burden sections and concepts.

The work presented here can be extended along several dimensions. One particularly important direction is to incorporate the background knowledge of the reader in the definition and computation of comprehension burden. For instance, consider a textbook designed for a course in Physics for which an introductory course covering basic ideas from different sciences is prerequisite. In such a book, it is reasonable for the author to presume that a mention of Newton's laws without explaining them will not impose comprehension burden, but this assumption is invalid for Einstein's theory of relativity. Similarly, authors may correctly mention concepts perceived to be part of the common knowledge of the target readers without explaining them. How should comprehension burden factor in such general knowledge? Another important direction is to explore the use of deeper text analysis to differentiate between bad mentions and mentions where the author has provided key idea and supplemented it with a reference to the section where the idea has been discussed in depth. It will also be interesting to investigate connections between comprehension burden and notions from cognitive psychology such as cognitive load and schema acquisition.

An obstacle we faced in our work was the lack of an established evaluation methodology for studying the performance of the proposed techniques. Ensuring objectivity and consistency across judges in the user studies are some challenges to be addressed in designing a direct measurement. Discounting externality and removing bias are some challenges to be addressed in designing an indirect measurement such as comparison of performance scores of students using good and bad versions of the same book. Designing sound evaluation methodology is fodder for rich future research.

6. REFERENCES

[1] R. Agrawal, S. Chakraborty, S. Gollapudi, A. Kannan, and K. Kenthapadi. Quality of textbooks: An empirical study. In *ACM DEV*, 2012.

[2] R. Agrawal, S. Gollapudi, A. Kannan, and K. Kenthapadi. Identifying enrichment candidates in textbooks. In *WWW*, 2011.

[3] R. Agrawal, S. Gollapudi, K. Kenthapadi, N. Srivastava, and R. Velu. Enriching textbooks through data mining. In *ACM DEV*, 2010.

[4] B. B. Armbruster and T. H. Anderson. Content area textbooks. Technical report, Reading Education Report No. 23, BBN, 1981.

[5] B. B. Armbruster and T. H. Anderson. Producing considerate expository text: Or easy reading is damned hard writing. Technical report, Reading Education Report No. 46, BBN, 1984.

[6] B. K. Britton, A. Woodward, and M. R. Binkley. *Learning from Textbooks: Theory and Practice*. Routledge, 1993.

[7] B. Bruce, A. Rubin, and K. Starr. Why readability formulas fail. *IEEE Transactions on Professional Communication*, PC-24, 1981.

[8] M. Chambliss and R. Calfee. *Textbooks for Learning: Nurturing Children's Minds*. Wiley-Blackwell, 1998.

[9] R. Clark, F. Nguyen, and J. Sweller. *Efficiency in learning: Evidence-based guidelines to manage cognitive load*. Pfeiffer, 2006.

[10] K. Collins-Thompson and J. P. Callan. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, 2004.

[11] E. K. Dishner. *Reading in the Content Areas: Improving Classroom Instruction*. Kendall/Hunt, 1992.

[12] W. DuBay. *The principles of readability*. Impact Information, 2004.

[13] C. Fellbaum. *WordNet: An electronic lexical database*. MIT Press, 1998.

[14] J. Gillies and J. Quijada. Opportunity to learn: A high impact strategy for improving educational outcomes in developing countries. *USAID Educational Quality Improvement Program (EQUIP2)*, 2008.

[15] S. R. Goldman. Learning from text: Reflections on the past and suggestions for the future. *Discourse Processes*, 23, 1997.

[16] W. Gray and B. Leary. *What makes a book readable*. University of Chicago Press, 1935.

[17] B. Grosz and C. Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3), 1986.

[18] R. Guillemette. Predicting readability of data processing written materials. *ACM SIGMIS Database*, 18(4), 1987.

[19] E. A. Hanushek and L. Woessmann. The role of education quality for economic growth. *Policy Research Department Working Paper 4122, World Bank*, 2007.

[20] E. B. Johnsen. *Textbooks in the Kaleidoscope: A Critical Survey of Literature and Research on Educational Texts*. Scandinavian University Press, 1992.

[21] J. S. Justeson and S. M. Katz. Technical terminology: Some linguistic properties and an algorithm for indentification in text. *Natural Language Engineering*, 1(1), 1995.

[22] R. Kate, X. Luo, S. Patwardhan, M. Franz, R. Florian, R. Mooney, S. Roukos, and C. Welty. Learning to predict readability using diverse linguistic features. In *COLING*, 2010.

[23] D. Kieras and C. Dechert. Rules for comprehensible technical prose: A survey of the psycholinguistic literature. Technical Report TR-85/ONR-21, University of Michigan, 1985.

[24] B. Lively and S. Pressey. A method for measuring the vocabulary burden of textbooks. *Educational Administration and Supervision*, 9(73), 1923.

[25] D. Marcu. *The theory and practice of discourse parsing and summarization*. MIT Press, 2000.

[26] F. Paas, A. Renkl, and J. Sweller. Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38(1), 2003.

[27] J. M. Pawlowski, K. C. Barker, and T. Okamoto. Quality research for learning, education, and training. *Reading & Writing Quarterly*, 10(2), 2007.

[28] J. Plass, R. Moreno, and R. Brünken. *Cognitive load theory*. Cambridge University Press, 2010.

[29] L. Polanyi and R. Scha. A syntactic approach to discourse semantics. In *COLING*, 1984.

[30] E. Pollock, P. Chandler, and J. Sweller. Assimilating complex information. *Learning and Instruction*, 12(1), 2002.

[31] R. Seguin. The elaboration of school textbooks. Technical report, ED-90/WS-24, UNESCO, 1989.

[32] L. Sherman. *Analytics of literature: A manual for the objective study of English prose and poetry*. Ginn and Company, 1893.

[33] S. A. Thompson and W. C. Mann. Rhetorical structure theory: A framework for the analysis of texts. *IPRA Papers in Pragmatics*, 1(1), 1987.

[34] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL-HLT*, 2003.

[35] H. Tyson-Bernstein. A conspiracy of good intentions: America's textbook fiasco. Technical report, Council for Basic Education, Washington, 1989.

[36] A. Verspoor and K. B. Wu. Textbooks and educational development. Technical report, World Bank, 1990.

[37] K. Wang, C. Thrasher, E. Viegas, X. Li, and P. Hsu. An overview of Microsoft Web N-gram corpus and applications. In *NAACL-HLT*, 2010.

[38] A. N. Whitehead. The organisation of thought. *Proceedings of the Aristotelian Society*, 17, 1916–17.

[39] S. Witte and L. Faigley. Coherence, cohesion, and writing quality. *College Composition and Communication*, 32(2), 1981.

[40] A. Woodward, D. L. Elliott, and C. Nagel. *Textbooks in School and Society: An Annotated Bibliography and Guide to Research*. Garland, 1988.

[41] World-Bank. *Knowledge for Development: World Development Report: 1998/99*. Oxford University Press, 1999.

[42] J. Zhao and M. Kan. Domain-specific iterative readability computation. In *ACM JCDL*, 2010.