Identifying Enrichment Candidates in Textbooks

Rakesh Agrawal Sreenivas Gollapudi Anitha Kannan Krishnaram Kenthapadi
Search Labs, Microsoft Research
Mountain View, CA, USA
{rakesha, sreenig, ankannan, krisken}@microsoft.com

ABSTRACT

Many textbooks written in emerging countries lack clear and adequate coverage of important concepts. We propose a technological solution for algorithmically identifying those sections of a book that are not well written and could benefit from better exposition. We provide a decision model based on the syntactic complexity of writing and the dispersion of key concepts. The model parameters are learned using a tune set which is algorithmically generated using a versioned authoritative web resource as a proxy. We evaluate the proposed methodology over a corpus of Indian textbooks which demonstrates its effectiveness in identifying enrichment candidates.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Education, Textbooks, Readability, Concepts, Dispersion

1. INTRODUCTION

Education is acknowledged to be the primary vehicle for improving the economic well-being of people living in emerging regions and for helping them become productive members of society [1, 24]. While the problem of providing high-quality education is multi-faceted and complex [8, 41, 43], studies during last twenty five years have highlighted the positive impact on student achievement of relevant, good-quality textbooks. Particularly in emerging economies, evidence suggests that textbooks are one of the most cost-effective means of improving the educational quality [13, 18, 25, 34]. Textbooks are also indispensable for fostering teacher learning and for their ongoing professional development [19, 40].

Unfortunately, many textbooks in emerging countries suffer from the lack of clarity of language as well as the inadequacy of information provided in the textbook [38]. Because of cost considerations, textbooks are often compressed into fewer pages resulting in poor exposition of subject matter [2]. Recognizing the role of education in development and the importance of textbooks in creating a high quality education system, a data mining based approach has been recently proposed for enhancing the quality of textbooks [3].

The essential idea is to enrich textbooks by algorithmically augmenting different sections with links to authoritative content from the Web. Their implementation uses Wikipedia as the source for mining authoritative content. They first identify the set of key concept phrases contained in a section. Using these phrases, they find Wikipedia articles that represent the central concepts presented in the section and augment the section with links to them.

A tacit assumption in [3] is that every section of a book needs augmentation. In this paper, we study the complementary problem of determining whether a section should be a candidate for enrichment since indiscriminate augmentations may put undue cognitive burden on the reader. We propose a decision model that uses the following variables:

1. *Syntactic complexity* of writing. Abstracting from the readability research [15], we employ two variables to model the syntactic complexity of a section: *average sentence length* in number of words, and ii) *average word length* in number of syllables. The higher the complexity, the greater is the need for augmentation.

2. *Dispersion* of key concepts. This variable originates from the observation that a section that contains widely dispersed concepts is harder to grasp than one that describes closely related concepts. The more dispersed are the concepts, the greater is the need for augmentation. We formalize the notion of dispersion and provide an algorithm for computing it.

Weights given to decision variables are learned using a tune set. The readability literature discourages using human ratings for creating tune set because of the difficulty of assembling a sufficiently large group of qualified judges [6, 48]. We therefore generate the tune set algorithmically in a novel way. The tune set consists of sections with different *maturity*, the intuition being that the more immature is a section, the greater the need for enriching it. In the absence of availability of update history, our implementation infers the maturity of a section by mapping it to the closest version of a Wikipedia article that has been updated multiple times and then using its maturity as a proxy for the maturity of the book section. If a section maps to multiple Wikipedia articles, maturities of the closest versions of

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

each of the Wikipedia articles are aggregated to arrive at the maturity of the section.

The paper is organized as follows. We discuss related work in §2. The rationale and definitions of model variables are presented in §3. We describe the procedure for generating tune set and how it is used to combine the above variables into the decision model in §4. We present in §5 the results of applying our model to a corpus of textbooks published by the Indian National Council of Educational Research and Training (NCERT). We conclude with a summary and directions for future work in §6.

2. RELATED WORK

The question of what factors influence understandability of a reading material has intrigued researchers for a long time. An early comprehensive investigation, dating back to 1935 [20], identified two principal sets of factors. The first set pertains to individual differences, such as levels of intellectual capacity, reading skills, attitudes and goals, previous experiences, and personal interests and tastes. The second set relates to the readability of the material, which in turn depends on format (page layout, appearance, etc.), organization (headings, indexes, 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 [22].

We refer the reader to the survey in [15] 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 per 100 periods, he showed how sentence-length averages had shortened over time [45]. The first readability formula, a weighted index of vocabulary complexity, is attributed to the work of Lively and Pressley in 1923 [33]. Since then, hundreds of linguistic variables have been used for the development of over two hundred readability formulas. Four principal types of linguistic variables have been studied : vocabulary load, sentence structure, idea density, and human interest. The most common measures of vocabulary load are rarity or difficulty of words, diversity of words, and word length. The most common measure of sentence difficulty is the average sentence length, though the percentages of indeterminate clauses and prepositional phrases have also been used. Idea density has been assessed by counts of propositional phrases or percentages of content words. Variables used for approximating human interest include the number of personal pronouns and nouns, proper names, and colorful words [6].

Readability formulas have been developed by first assessing the reading difficulty of a collection of texts and then applying regression analysis to the values of the chosen linguistic variables. Lorge's 1939 study [35] first used scores associated with the McCall-Crabbs Standard Test Lessons in Reading [36], which were subsequently used in developing many readability formulas. The cloze procedure, introduced by Taylor in 1953 [47], replaces every n^{th} word from a passage with a marker and then computes the difficulty of the passage as a fraction of deleted words that can be correctly guessed by a reader. Human judgments are rarely used for assigning reading difficulty because of the complexity of selecting an appropriate, large group of judges and questions concerning reliability and generalizability of results [6, 48].

The readability formulas generally compute the grade level (1 to 12), or a score from 0 (hard) to 100 (easy). Some widely used formulas include Flesch Reading Ease Score [17], Flesch-Kincaid Grade Level [31], Dale-Chall Grade Level [14], Gunning Fog Index [23], SMOG Index [37], Coleman-Liau Index [10], and Automated Readability Index [46]. Notwithstanding their criticism because of their purported low validity from the perspective of psycholinguistic theories [5, 42] and efforts to develop new approaches [11, 29, 30, 51], the usefulness of readability formulas has been documented in a large number of papers and they remain in wide use in a variety of settings [15]. They have also been found to be valid for English as foreign/second language use [21].

We do not directly use readability formulas as decision variables in our model. However, it is only through a careful analysis of various formulas that we arrive at our decision variables.

Work related to our notion of the dispersion of key concepts includes research on idea density, cohesion, and coherence. The psycholinguistic research, stemming from the work of Kintsch and Keenan [32], considers the proposition to be the basic unit of understanding, and defines idea density to be the fraction of propositions in a text. Since a proposition requires certain amount of processing effort, the high idea density makes for slower processing. Propositions correspond roughly to verbs, adjectives, adverbs, prepositions, and subordinating conjunctions, but not nouns or pronouns [12]. On the other hand, the building blocks of our dispersion computation are concepts, which correspond to terminological noun phrases [28]. The consensus in the readability research is that idea density explains little variance beyond what is already accounted by vocabulary overload and sentence structure [22].

In linguistics, coherence refers to the connectedness of the ideas in a piece of writing, whereas cohesion refers to connections between sentences [50]. 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. The topic of a sentence is thought to be among the first few words of a sentence [39]. Thus, our notion of dispersion of concepts has conceptual similarity with the notion of coherence. However, coherence is still in the process of being defined and there is considerable fuzziness in the use of the terms coherence and cohesion; the author of the classic work, *The grammar of coherence* [49] now says that it should have been entitled *The grammar of cohesion* [27]. We give a rigorous definition of dispersion and provide an algorithm for computing it.

Related work also includes the proposal in [2] to create an education network to harness the collective efforts of educators, parents, and students to collaboratively enhance the quality of educational material. Some websites (e.g. Notemonk.con) allow students to download textbooks, ask questions on a topic, and annotate books for quick reference. Several institutions are making the videos of the course lectures available through Internet and there are websites (e.g. EducationPortal.com) that aggregate links to them. Another noteworthy effort is the Digital StudyHall project [44]. They digitally record live classes, collect them in a large distributed database, and distribute them on DVDs to poor

Flesch Reading Ease Score [17]	206.835	−	84.6	×	S/W	−	1.015	×	W/T
Flesch-Kincaid Grade Level [31]	−15.59	+	11.8	×	S/W	+	0.39	×	W/T
Dale-Chall Grade Level [14]	14.862	−	11.42	×	D/W	+	0.0512	×	W/T
Gunning Fog Index [23]			40	×	C/W	+	0.4	×	W/T
SMOG Index [37]	3.0	+	$\sqrt{30}$	×	$\sqrt{C/T}$				
Coleman-Liau Index [10]	−15.8	+	5.88	×	L/W	−	29.59	×	T/W
Automated Readability Index [46]	−21.43	+	4.71	×	L/W	+	0.50	×	W/T

C	=	Number of words with three syllables or more
D	=	Number of words on the Dale Long List
L	=	Number of letters
S	=	Number of syllables
T	=	Number of sentences
W	=	Number of words

Table 1: Popular readability formulas and their variables

rural and slum schools. We view these efforts as complementary approaches to improving the quality of textbooks.

3. DECISION VARIABLES

Our decision model for enriching a section is based on the syntactic complexity of the writing and the dispersion of key concepts mentioned in the section. We discuss here the rationale for choosing these decision variables and formally define them.

3.1 Syntactic Complexity

Table 1 summarizes some of the popular readability formulas and the variables they use. We observe that all formulas base their calculations on two classes of variables. First, they all use a sentence structure measure, generally sentence length, the underlying intuition being that longer sentences are harder to read and comprehend. The sentence length can be in terms of the number of letters or the number of words, though the empirical evidence from past studies overwhelmingly favors the number of words.

The second measure they use captures the difficulty of the vocabulary at word level in terms of word familiarity or word length. The Dale long list [14] is frequently used for computing word familiarity. We do not employ word familiarity because of potential vocabulary mismatch between textbooks written in local variants of English and the Dale list. The word length can be defined in terms of the number of syllables or the number of letters. Both Coleman-Liau Index and Automated Readability Index calculate word length in number of letters, but their primary consideration was data processing efficiency and the effectiveness of this approach is suspect [15]. Another approach is to compute word length in terms of the number of syllables, the intuition being that the words with more syllables are more complex.

We also note that different readability formulas combine the above two measures differently and the combinations are learned with respect to specific datasets (often McCall-Crabbs Standard Test Lessons in Reading [36]). As a result, these formulas are highly correlated, a fact we confirmed in our experiments. We find it unnatural to use the regression equations (with their specific intercepts and coefficients) underlying these formulas directly as variables in the decision model.

After considerable experimentation, we settled on the following two variables as measures for the complexity of writing:

1. *Average sentence length*: average number of words per sentence in the section.

2. *Average word length*: average number of syllables per word in the section.

Algorithm 1 ComputeDispersion

Input: A textbook section s; Grammatical pattern R for detecting terminological noun phrases; An authoritative structured external source of concepts that also contains relationship between them (e.g. Wikipedia).
Output: Dispersion value for section s.

1: Compute set C of candidate concepts present in section s using the linguistic pattern R.
2: Determine set V of nodes corresponding to concepts in C that match an article title from the external source.
3: Let W denote the set of all links in the external source. Define $E = \{(v_1, v_2) | v_1, v_2 \in V \wedge (v_1, v_2) \in W\}$. Compute the directed graph $G = (V, E)$ thus induced by links in W.
4: $dispersion(s) := 1 - \frac{|E|}{|V|(|V|-1)}$.

See [9, 16] for algorithms for computing the number of syllables per word. The number of syllables in a word can also be approximated by counting consonant-separated vowels. Each group of adjacent vowels counts as one syllable (for example, 'ea' in 'real' contributes one syllable, whereas 'e...a' in 'regal' contributes two syllables), but an 'e' occurring at the end of a word does not contribute to syllable count. Each word has at least one syllable.

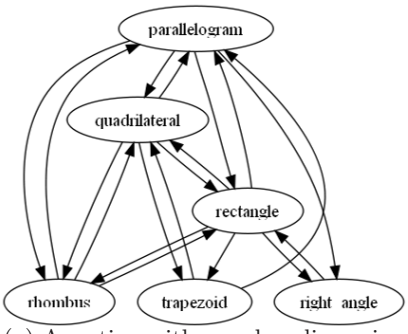
3.2 Dispersion

We next consider a semantic notion of the quality of a book section. After going through several textbooks, we observed that a section that discussed concepts related to each other had better quality than one that discussed many unrelated concepts. We formally capture this intuition by defining a measure of dispersion over key concepts.

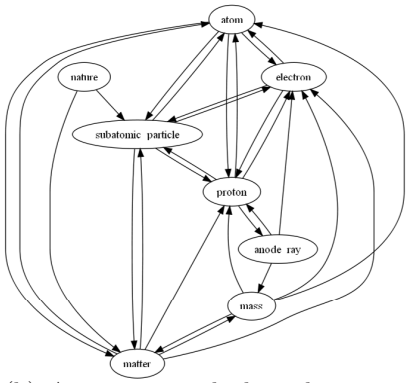
Let V represent the set of key concepts present in a section s. Let rel be a binary relation that determines whether a concept in V is related to another concept in V, that is, $rel(x, y)$ is *true* if concept x is related to concept y and *false* otherwise. We define *dispersion* of a section as the fraction of ordered key concept pairs that are not related, that is,

$$dispersion(s) := \frac{|\{(x, y) | x, y \in V \wedge x \neq y \wedge \neg rel(x, y)\}|}{|V|(|V| - 1)}. \quad (1)$$

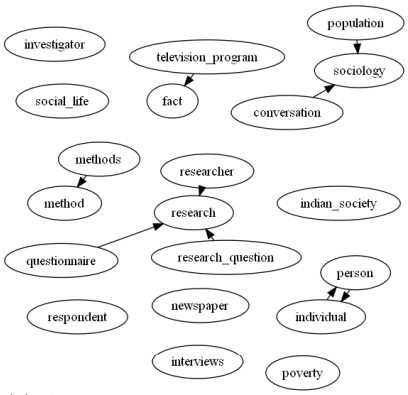
We note that *dispersion* takes values between 0 and 1, with 0 corresponding to a section where all key concepts are mutually related and 1 corresponding to a section with mutually unrelated key concepts. We next describe how we compute the set of key concepts and the *rel* relation (Algorithm 1).



(a) A section with very low dispersion (Grade IX Mathematics book)



(b) A section with low dispersion (Grade IX Science book)



(c) A section with very large dispersion (Grade XII Sociology book)

Figure 1: Concept graphs illustrating dispersion

3.2.1 Computing dispersion

The concepts of interest in our application typically consist of terminological noun phrases containing adjectives, nouns, and sometimes prepositions [28]. It is rare for concepts to contain other parts of speech such as verbs, adverbs, or conjunctions. We identify candidate concepts using the techniques described in [3], employing the linguistic pattern A^*N^+, where A is an adjective and N a noun. Examples of concepts satisfying this pattern include "cumulative distribution function", "fiscal policy", and "electromagnetic radiation". This pattern has been shown to perform somewhat better than the patterns C^*N and $(C^*NP)^?(C^*N)$, where P refers to a preposition and $C = A|N$.

Having determined the set of concepts, a straightforward approach to derive *rel* relation would be to manually label the concept pairs. However, labeling is a laborious and subjective task. We instead consider an authoritative structured external source of concepts that also contains relationship between them and map concepts in textbooks to this source.

Our implementation maps textbook concepts to Wikipedia articles and treats a concept c_1 to be related to another concept c_2 if the Wikipedia article corresponding to c_1 has a link to the Wikipedia article corresponding to c_2. We only consider concept phrases that match the title of a Wikipedia article exactly. If any Wikipedia article is redirected to another article, we follow the redirect link till an article is found. We then consider the directed graph induced by these mapping articles and the Wikipedia links between them. We remove the isolated nodes in this graph, and compute dispersion as one minus the edge density in the resulting concept graph (using Eqn. 1).

We illustrate our notion of dispersion through some examples from the NCERT textbooks. Figure 1(a) and 1(b) show the concept graphs for two sections with small dispersion. The first section titled "Types of Quadrilaterals" from Grade IX Mathematics book has 19 directed edges over 6 nodes with dispersion 0.37 and the second section titled "Charged Particles in Matter" from Grade IX Science book has 29 directed edges over 8 nodes with dispersion 0.48. Indeed we observe that the concepts within each of these sections are quite related to each other, contributing to many links between the corresponding Wikipedia articles and thus the low dispersion values. Figure 1(c) shows the concept graph for a section with large dispersion (with some isolated nodes also shown). This section titled "Variety of Methods" from Grade XII Sociology book has 9 edges over 13 non-isolated nodes, contributing to a dispersion value of 0.94. We see that the section discusses rather unrelated concepts, contributing to fewer links between the corresponding Wikipedia article and thus large dispersion.

4. DECISION MODEL

We take a learning approach to arrive at the model for deciding whether a book section requires enrichment. Our proposed model is probabilistic which learns its parameters using a tune set. A seemingly obvious way of generating the tune set would be to have human judges. However, it is difficult to assemble a sufficiently large group of qualified judges who can provide consistent ratings. We, therefore, generate the tune set algorithmically. The tune set consists of sections with different maturity, the intuition being that the more immature is a section, the greater the need for its enrichment. We discuss the decision model and the generation of tune set next.

4.1 Model

Our goal is to learn a decision model that can provide a probabilistic score of whether a textbook section requires enrichment based on the values of decision variables for that section. We would also like such a decision model to automatically learn the relative importance between the decision variables. The binary logistic regression eminently lends itself to this desiderata.

Let $\mathbf{z}$ represent a section's decision variables: a three dimensional vector whose components are the average sentence length, average word length, and dispersion. Given $\mathbf{z}$, the binary logistic regression predicts the probabilistic score that a section needs enrichment (i.e., label $y = 1$) through the logistic function:

$$P(y = 1|\mathbf{z}, \mathbf{w}) = \frac{1}{1 + \exp\{-(b + \mathbf{z}^T\mathbf{w})\}}. \quad (2)$$

The parameter $\mathbf{w}$ is the weight vector of the function, with each component w_j measuring the relative importance of the decision variable z_j for predicting the label y.

Algorithm 2 GENERATETUNESET

Input: A corpus of textbooks divided into sections; A collection of versioned documents from an authoritative web resource such as Wikipedia; Threshold parameters θ_1 and θ_2.

Output: A tune set consisting of a subset of sections, each labeled either 1 (Enrich) or 0 (Don't).

1: **for** each section s **do**
2: Map section s to a set $W(s)$ of most similar versioned documents from the web resource, along with their similarity scores $sim(s, v)$ $\forall v \in W(s)$. (§4.2.1)
3: Compute immaturity score $\tilde{m}(v)$ for each versioned document $v \in W(s)$. (§4.2.2)
4: Compute immaturity score $m(s)$ for section s by aggregating immaturity scores $\tilde{m}(v)$ for $v \in W(s)$, weighted by their similarity scores $sim(s, v)$.
5: $Label(s) := 1$ if $m(s) > \theta_1$; 0 if $m(s) < \theta_2$; $undefined$ otherwise.
6: **end for**
7: Output $\langle s, Label(s)\rangle$ for sections s where $Label(s)$ is either 0 or 1.

The weight vector $\mathbf{w}$ is learned from a tune set consisting of N textbook sections: $\{\mathbf{Z}, \mathbf{y}\} = \{(\mathbf{z}_1, y_1), \ldots, (\mathbf{z}_N, y_N)\}$, with $(\mathbf{z}_i, y_i)$ representing the decision variable vector $\mathbf{z}_i$ and the label y_i for the i^{th} textbook section. The optimal $\mathbf{w}$ is the one that maximizes the conditional log-likelihood of the labels in the tune set:

$$\arg\max_{\mathbf{w}} \log P(\mathbf{y}|\mathbf{Z}, \mathbf{w}) = \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log P(y_i|\mathbf{z}_i, \mathbf{w}). \quad (3)$$

During the application phase, the probability computed by the model for a given textbook section is multiplied by 100 to yield the *Enrichment Score* which is then used to decide whether to enrich the section or not.

4.2 Generating Tune Set

Given the difficulty of obtaining manual judgments, we propose using meta data associated with textbooks to obtain labels. One such meta data is the immaturity level of a section; an immature section hinders the positive learning experience of a student, and therefore calls for enrichment. However, immaturity computation requires access to rich data such as extent and timing of the revisions, which is typically not available for textbooks. We, therefore, resort to an indirect device for estimating the maturity of a section.

We note that authoritative information resources on the Web, such as Wikipedia, are created through collective efforts of multiple authors. The content gets repeatedly updated until writers expressing opinions on the subject come to a consensus. Hence, we map a textbook section to the most similar version of a similar article in a web resource and use the immaturity of that version as the proxy for the immaturity of the textbook section.

The tune set generation is outlined in Algorithm 2. We sample a subset of textbook sections across all subjects and classes. For each section, we find a small set of closest matching versions in the web resource that are similar in content. The matches are found using the technique described in §4.2.1. We then compute the immaturity for these versions using the technique given in §4.2.2. The immaturity scores are then aggregated through a weighted combination (weights are the normalized similarity scores) to produce the maturity score for the textbook section. This score is then converted into a decision on what label should be assigned to this book section.

We note that the immaturity computation is reliable only at the extreme ends: very high values or very low values of scores. Fortunately, we only need a few labeled sections in the tune set. The parameters θ_1 and θ_2 allow us to achieve this goal. Their values are empirically determined, balancing the need for high precision with the need for having sufficient labeled data.

4.2.1 Computing similarity

In a document model where each document is treated as a set of words, a well-known measure of similarity between documents A and B is the Jaccard index, defined as $sim(A, B) = |A \cap B|/|A \cup B|$. However, documents in our application can have large sizes that can make this computation expensive. Documents can also have very different lengths and not all terms in a document contribute equally to the identity of the document. We describe next the method that incorporates these concerns.

First, a few preliminaries. Given a set $A \subseteq U$ of elements, its min-wise independent permutation, $MH(A) := \arg\min_x\{R(x)|x \in A\}$, where $R : U \rightarrow [0, 1]$ is a consistent hash function that maps elements from U uniformly and randomly in the interval $[0, 1]$. Thus, $MH(A)$ denotes the leftmost element of A in the permutation. Now, for any two sets A and B of elements in U, $|A \cap B|/|A \cup B| = \Pr[MH(A) = MH(B)]$. This result extends to multisets as well [7].

Hence represent documents as bags of words wherein frequencies (or other weights such as *tf-idf*) are associated with each word. Frequency normalize the weight for each document $x \in A$ to give w_x. Now, using a consistent hash function $R(x)$ that maps elements of A to the interval $[0, 1]$, compute $\tilde{A} = \{x \in A|x \in A \bigwedge R(x) \leq w_x\}$. Next, compute $MH(\tilde{A})$ using the min-wise independent permutations of $\tilde{A}$. Finally, compute H min-hashes to yield the sketch of A, $S(A) = \{MH_1(\tilde{A}), MH_2(\tilde{A}), \ldots, MH_H(\tilde{A})\}$. Repeat for B. Now, $|S(A) \cap S(B)|/|S(A) \cup S(B)|$ gives the estimate for $sim(A, B)$.

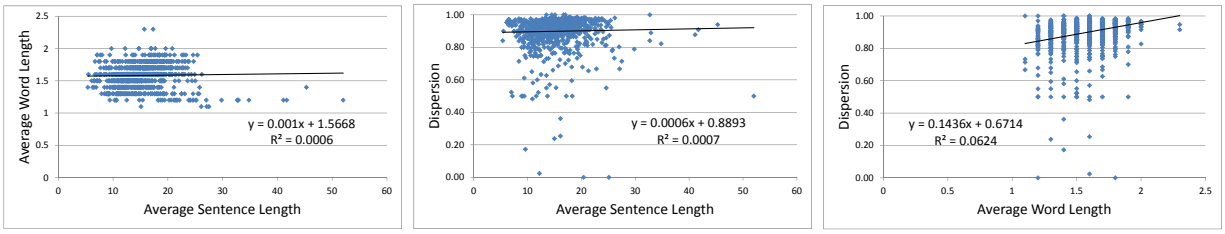
4.2.2 Computing immaturity

Consider a web repository in which a new version of a document is created at the end of the day, ignoring multiple updates to the document within a day. Older versions of a document are saved when a new version is created. We observe the following:

- Paraphrasing, additions or deletions indicate the amount of revision. Thus, the relative change in the size of the document is an indicator of the maturity of a version (the smaller the change, the higher the maturity).
- The number of days for which a version remains the latest version is also an indicator of the maturity of the version (the longer the duration, the higher the maturity).
- People tend to consult nearby versions when creating a revision. Thus, maturity is a local phenomenon driven by local context.

	Sciences	Social Sciences	Commerce	Mathematics
Grade IX	Science	History, Political Science		Mathematics
Grade X	Science	History		Mathematics
Grade XI		Political Science	Accountancy, Economics, Business Studies	
Grade XII	Physics	History, Sociology	Accountancy, Economics	Mathematics

Table 2: NCERT textbooks by grade and subject



(a) Average word length vs. Average sentence length (b) Dispersion vs. Average sentence length (c) Dispersion vs. Average word length

Figure 2: Correlation between the decision variables

Assume days are numbered from 1 to the current day T. Consider a document whose initial version v_1 was created on day 1. Let $\mathbf{L}$ be a vector of length T whose i^{th} component L_i is equal to the size of the document (in number of words) on day i. Define a vector $\delta(\mathbf{L})$ whose i^{th} component is the relative change in document size between neighboring days i and $i-1$:

$$\delta(L)_i = |L_i - L_{i-1}|/L_i. \quad (4)$$

For a particular version v created on day d, we define its *immaturity* $\tilde{m}(v)$ to be the value of convolution between $\delta(L)$ and a smooth filter $\mathbf{h}$ on day d:

$$\tilde{m}(v) := (\delta(L) * \mathbf{h})_d = \sum_{j=\max(-K/2,1-d)}^{\min(K/2,T-d)} h_j \delta(L)_{d+j}, \quad (5)$$

where K is a parameter of the filter used in the convolution.

The convolution with a smooth filter allows for modeling immaturity as a smooth continuous process, and the use of local neighborhood enables incorporating local context. We employ the frequently used Hann Filter [4]

$$h_j = 0.5(1 + \cos(2\pi j/K)) \quad (6)$$

that has K days spatial support with a smooth fall off in the chosen K sized neighborhood.

We note that there have been efforts (e.g. [26]) to assign quality index to Wikipedia articles taking into account edit history of the article such as the frequency and size of edits and the type and reputation of the authors. However, we are not aware of any work targeted at computing the maturity of an arbitrary version of a Wikipedia article.

5. EXPERIMENTS

We performed a large number of experiments to evaluate the effectiveness of the proposed methodology. We next present some of the key results.

5.1 Experimental Setup

We used a corpus of high school textbooks published by the Indian National Council of Educational Research and Training (NCERT). We considered seventeen books from grades IX–XII, covering four broad subject areas: Sciences, Social Sciences, Commerce, and Mathematics. Table 2 provides a breakup of the books by grade and subject. The same dataset has been used in the study in [3]. There are a total of 191 chapters and 1313 sections in these books. We limit our experiments to sections with at least 20 sentences and 300 words so that each section is large enough to contribute to meaningful decision variable computations.

The decision model is trained over normalized values of the variables. We use the normalization, $x \rightarrow (x-\mu)/(2 \times \sigma)$. For generating tune set, we use Wikipedia as the web resource and consider all versions of a Wikipedia article from year 2008 to 2010. For computing the immaturity score of a version, we use the local neighborhood of one month ($K = 30$ in Eqn. 5). To obtain the thresholds for converting the immaturity score to labels (Algorithm 2), we look at the histogram of immaturity scores, and pick thresholds that yield balanced dataset. We thus use thresholds of $\theta_1 = 0.3$ and $\theta_2 = 0.1$.

5.2 Analysis of Decision Variables

We first investigated whether our decision variables have independent signals. Figure 2 shows the mutual correlation between the three variables. We see that the variables have almost no correlation (R^2 values are low: 0.0006 between average sentence length and average word length; 0.0007 between dispersion and average sentence length; 0.0624 between dispersion and average word length). Hence, we choose to include them all in the decision model.

5.3 Decision Model

Decision variable	Weight
Average sentence length	0.268
Average word length	0.068
Dispersion	0.284
Intercept (b)	−0.806

Table 3: Weights of the decision model

Table 3 shows the learned weights of our decision model. The weights capture the relative importance of the decision

variables; in fact, a positive value indicates that the corresponding variable positively contributes towards the need for enrichment. We can see that all the weights have the right directionality; they all have positive coefficients suggesting that sections having disperse concepts, long sentences, or long words are good candidates for enrichment. While both average sentence length and dispersion have similar weights, average word length plays a smaller role.

Decision variable	Avg. weight	Variance
Average sentence length	0.274	0.007
Average word length	0.076	0.009
Dispersion	0.290	0.012
Intercept (b)	−0.807	0.002

Table 4: Stability of the weights

We performed sensitivity analysis to validate the learned weights. In particular, we relearned the model using random subsets of the tune set, each time removing 100 random samples from it. The mean and variance of the weights over 100 learned models are shown in Table 4. This analysis indicates that the learned model is fairly stable.

We use the learned model to make three kinds of prediction: **Enrich**, **Don't**, and **Examine**. We compute the enrichment score predicted by the decision model for all the sections and sort the scores in decreasing order. Sections in the first quartile of this sorted list are candidates for **Enrich**ment. Those in the fourth quartile do not require enrichment and are tagged with **Don't**. Sections in the remaining two quartiles need human intervention in order to decide whether they need enrichment, and hence the tag **Examine**. This quartile binning put a threshold of 34 and above for **Enrich**, and 29 and below for **Don't**.

5.4 Quality of Results

We applied our decision model to the NCERT corpus and then manually examined the quality of results for some sample sections.

5.4.1 Sections Needing Enrichment

Table 5 shows four sections with high predicted scores for the need for enrichment, along with the values of the decision variables. These sections have relatively long sentences (up to six standard deviations to the right of the mean) and large dispersion values (close to unity). Consider, for example, the section titled "Choice of Form of Business Organization". There are many long sentences such as the one below, making the comprehension harder.

Factors like capital contribution and risk vary with the size and nature of business, and hence a form of business organisation that is suitable from the point of view of the risks for a given business when run on a small scale might not be appropriate when the same business is carried on a large scale.

Furthermore, this section presents a number of new concepts, many of which are not directly related to each other. Here are some examples from the 56 concepts identified: "limited liability", "assets", "companies law", "functional areas" and "karta". Given the broad range of concepts discussed in this section, a reader is likely to benefit from enrichment.

Similarly, the section titled "Forms of Organizing Public Sector Enterprises" lists a taxonomy of public and private sector enterprises and discusses three forms of public enterprises. This section presents many concepts that are not directly related (e.g. "legislature", "private sector", "national security", "globalisation", "statutory corporation", "capital market"). Moreover, this section contains a large number of complex words (e.g. 'liberalisation' (6 syllables), 'proprietorship' (5 syllables), 'globalization' (5 syllables), 'privatisation' (5 syllables), 'statutory' (4 syllables), 'legislature' (4 syllables)). The section also contains long sentences such as the one below.

According to the Indian Companies Act 1956, a government company means any company in which not less than 51 percent of the paid up capital is held by the central government, or by any state government or partly by central government and partly by one or more state governments.

Business Studies (XI, Business Studies) Chapter 2: Forms of Business Organisation *Section 7: Choice of Form of Business Organization* ⟨**53**, 45.3, 1.4, 0.94⟩
Business Studies (XI, Business Studies) Chapter 3: Private, Public and Global Enterprises *Section 3: Forms of Organizing Public Sector Enterprises* ⟨**41**, 25.4, 1.9, 0.97⟩
Themes In Indian History (XII, History) Chapter 11: Rebels and the Raj *Section 3: What the rebels wanted* ⟨**41**, 26.2, 1.6, 0.98⟩
Indian Society (XII, Sociology) Chapter 4: The Market as a Social Institution *Section 1: Sociological Perspective on Markets and the Economy* ⟨**40**, 23.8, 1.8, 0.98⟩

Table 5: Sample sections needing enrichment (Tuple below the section title gives ⟨Enrichment score, Average sentence length, Average word length, Dispersion⟩ values)

Science (X, Science) Chapter 13: Magnetic Effects of Electric Current *Section 6: Electric Generator* ⟨**11**, 12.2, 1.6, 0.02⟩
Science (IX, Science) Chapter 11: Work and Energy *Section 3: Rate of Doing Work* ⟨**12**, 9.6, 1.4, 0.17⟩
Mathematics (IX, Mathematics) Chapter 8: Quadrilaterals *Section 3: Types of Quadrilaterals* ⟨**14**, 7, 1.5, 0.37⟩
Democratic Politics (IX, Political Science) Chapter 6: Democratic Rights *Section 0: Overview* ⟨**25**, 12.5, 1.5, 0.75⟩

Table 6: Sample sections not needing enrichment

5.4.2 Sections Not Needing Enrichment

Table 6 shows four sections with low scores for the need for enrichment. We notice that these sections typically have low dispersion values (up to eight standard deviations left of the mean), suggesting that they discuss concepts that are related to each other. For example, consider the section titled "Electric Generators". The key concepts identified from this section are "electric current", "magnetic field", "electric power", "electricity", "direct current", "electrical gener-

Physics (XII, Physics)	
Chapter 10: Wave Optics *Section 1: Introduction* **Enrich:**⟨**34**, 19.7, 1.7, 0.89⟩	Chapter 1: Electric Charges and Fields *Section 7: Forces between Multiple Charges* **Don't:**⟨**14**, 15.0, 1.3, 0.24⟩
Chapter 10: Wave Optics *Section 4: Coherent and Incoherent Addition of Waves* **Enrich:**⟨**34**, 19.1, 1.5, 0.93⟩	Chapter 13: Nuclei *Section 7: Nuclear Energy* **Don't:**⟨**16**, 16.1, 1.6, 0.25⟩
Introductory Macroeconomics (XII, Economics)	
Chapter 6: Open Economy Macroeconomics *Section 2: The Foreign Exchange Market* **Enrich:**⟨**37**, 21.5, 1.6, 0.97⟩	Chapter 4: Income Determination *Section 2: Movement along a Curve versus Shift of a Curve* **Don't:**⟨**27**, 12.0, 1.4, 0.87⟩
Chapter 2: National Income Accounting *Section 1: Some Basic Concepts of Macroeconomics* **Enrich:**⟨**37**, 20.5, 1.6, 0.98⟩	Chapter 6: Open Economy Macroeconomics *Section 4: Trade Deficits, Savings and Investment* **Examine:**⟨**30**, 13.1, 1.6, 0.90⟩
Themes In Indian History (XII, History)	
Chapter 11: Rebels and the Raj *Section 3: What the rebels wanted* **Enrich:**⟨**41**, 26.2, 1.6, 0.98⟩	Chapter 1: Bricks, Beads and Bones *Section 7: Seals, Script, Weights* **Don't:**⟨**25**, 11.4, 1.5, 0.77⟩
Chapter 15: Framing the Constitution *Section 2: The Vision of the Constitution* **Enrich:**⟨**39**, 24.3, 1.6, 0.96⟩	Chapter 12: Colonial Cities *Section 1: Towns and Cities in Pre-colonial Times* **Don't:**⟨**26**, 15.7, 1.7, 0.69⟩
Indian Society (XII, Sociology)	
Chapter 4: The Market as a Social Institution *Section 1: Sociological Perspective on Markets and the Economy* **Enrich:**⟨**40**, 23.8, 1.8, 0.98⟩	Chapter 2: The Demographic Structure of the Indian Society *Section 5: Literacy* **Don't:**⟨**28**, 9.9, 1.8, 0.89⟩
Chapter 3: Social Institutions: Continuity and Change *Section 2: Tribal Communities* **Enrich:**⟨**40**, 23.8, 1.8, 0.97⟩	Chapter 2: The Demographic Structure of the Indian Society *Section 6: Rural-Urban Differences* **Don't:**⟨**27**, 14.6, 1.6, 0.76⟩
Mathematics (XII, Mathematics)	
Chapter 9: Differential Equations *Section 5: Methods of Solving First Order, First Degree Differential Equations* **Enrich:**⟨**37**, 25.5, 1.3, 0.92⟩	Chapter 6: Application of Derivatives *Section 5: Approximations* **Don't:**⟨**21**, 10.8, 1.3, 0.65⟩
Chapter 13: Probability *Section 6: Random Variables and its Probability Distributions* **Examine:**⟨**33**, 18.6, 1.3, 0.95⟩	Chapter 9: Differential Equations *Section 3: General and Particular Solutions of a Differential Equation* **Don't:**⟨**21**, 11.8, 1.5, 0.60⟩

Table 7: Two sections each with high enrichment scores (left column) and low enrichment scores (right column) for all the books for which key concepts were identified in [3]

ator" and "magnet", all of which are very related to each other. By going through this section, we verified that the section is written clearly and cogently. We observed similar pattern for other sections. For example, the section titled "Overview" provides a two paragraph introduction to the chapter "Democratic Rights" and discusses related concepts using short sentences. We have already discussed dispersion for the section titled "Types of Quadrilaterals" in §3.2.

5.4.3 Enrichment Scores of Books / Sections Discussed in [3]

For reference, we provide in Table 7 two sections each that have been predicted to need enrichment most and least for all the books for which key concepts were identified in [3]. Trends discussed earlier also hold for these books. For example, for Grade XII Physics book, we observe that the large difference in the enrichment scores between sections needing enrichment and sections not needing enrichment arises due to the large difference in dispersion scores. For Grade XII Introductory Macroeconomics book, the difference arises due to the significant difference in average sentence length. For Grade XII books on History (Themes in Indian History), Sociology (Indian Society) and Mathematics, the significant differences in both average sentence length and dispersion contribute to the difference in enrichment scores.

We provide in Table 8 the enrichment scores for the specific sections for which key concepts were identified in [3]. We see that some of these sections do need enrichment, while some don't. Thus, by identifying sections that are not well-written and could benefit from enrichment, our results complement the work in [3] where the focus is on identifying key concepts that should be augmented with links to web content.

6. CONCLUSIONS AND FUTURE WORK

Given the centrality of education for improving the lives of people in emerging regions and the role of textbooks in a high quality education system, we set out to devise technolgies for enriching textbooks. We presented a diagnostic tool for identifying those sections of a book that are not well-written and hence should be candidates for enrichment. We propose a probabilistic decision model for this purpose, which is based on syntactic complexity of the writing and the newly introduced notion of the dispersion of key concepts mentioned in the section. The model is learned using a tune set which is automatically generated in a novel way. This procedure maps sampled text book sections to the closest versions of Wikipedia articles having similar content and uses the maturity of those versions to assign need-

Physics (XII, Physics)
Chapter 8: Electromagnetic Waves *Section: Introduction* **Don't:**⟨**29**, 16.1, 1.8, 0.80⟩
Chapter 9: Ray Optics and Optical Instruments *Section: Refraction* **Don't:**⟨**29**, 12.8, 1.5, 0.92⟩
Chapter 15: Communication Systems *Section: Modulation and its Necessity* **Examine:**⟨**31**, 15.7, 1.7, 0.87⟩
Introductory Macroeconomics (XII, Economics)
Chapter 1: Introduction *Section: Emergence of Macroeconomics* *Excluded due to small section length*
Chapter 5: The Government: Functions and Scope *Section: Fiscal Policy* **Examine:**⟨**33**, 15.5, 1.6, 0.96⟩
Chapter 6: Open Economy Macroeconomics *Section: Trade Deficits, Savings and Investment* **Examine:**⟨**30**, 13.1, 1.6, 0.90⟩
Themes In Indian History (XII, History)
Chapter 7: An Imperial Capital Vijayanagara *Section: Rayas, Nayaks and Sultans* **Enrich:**⟨**35**, 17.9, 1.8, 0.96⟩
Chapter 11: Rebels and the Raj *Section: What the rebels wanted* **Enrich:**⟨**41**, 26.2, 1.6, 0.98⟩
Indian Society (XII, Sociology)
Chapter 4: The Market as a Social Institution *Section: Understanding Capitalism as a Social System* **Enrich:**⟨**36**, 19.1, 1.7, 0.96⟩
Mathematics (XII, Mathematics)
Chapter 7: Integrals *Section: Introduction* **Don't:**⟨**29**, 14.9, 1.8, 0.81⟩
Chapter 13: Probability *Section: Random Variables and its Probability Distributions* **Examine:**⟨**33**, 18.6, 1.3, 0.95⟩
Chapter 9: Differential Equations *Section: Methods of Solving First Order, First Degree Differential Equations* **Enrich:**⟨**37**, 25.5, 1.3, 0.92⟩

Table 8: Enrichment scores for sections for which key concepts were identified in [3]

for-enrichment labels. The maturity of a version is computed by considering the revision history of the corresponding Wikipedia article and convolving the changes in size with a smoothing filter.

We applied the model on a corpus of Indian textbooks published by the National Council of Educational Research and Training. The empirical evaluation demonstrates that the proposed techniques are able to identify enrichment candidates across various subjects and grades. We remark that though we use Indian textbooks in the experiments, our methodology has broad applicability as there is nothing country specific in our methodology.

Some of our results could be of general interest. For example, our notion of maturity could possibly be used to assess the soundness of the latest versions of Wikipedia articles and identify candidates for improvement. Similarly, the readability literature has not paid attention to the idea of the dispersion of key concepts, possibly because of the difficulty of operationalizing this idea. We have not only provided formal definition of dispersion, but also provided an algorithmic procedure for estimating its value. Our study found this feature to be a strong predictor of the need for enrichment of a section. Exploring new applications of the ideas introduced in this paper and generalizing them would be an interesting direction for future work.

Another interesting direction would be to vet the results of this study by polling the current users of the textbooks. Ideally, the students would provide their judgments soon after they have finished studying a section. A beneficial by product of this exercise would be the creation of useful labeled data for further refining the model for diagnosing deficient sections. Finally, it would be fruitful to investigate the extensions needed for applying the techniques from this paper to textbooks written in non-English languages.

7. REFERENCES

[1] Knowledge for development: World development report 1998/99. Technical report, World Bank, 1998.

[2] Improving India's education system through information technology. IBM, 2005.

[3] R. Agrawal, S. Gollapudi, K. Kenthapadi, N. Srivastava, and R. Velu. Enriching textbooks through data mining. In *First Annual ACM Symposium on Computing for Development (ACM DEV)*, 2010.

[4] R. Blackman and J. Tukey. *The measurement of power spectra from the point of view of communications engineering*. Dover, 1959.

[5] B. Bruce, A. Rubin, and K. Starr. Why readability formulas fail. *IEEE Transactions on Professional Communication*, PC-24:50–52, 1981.

[6] J. Chall. *Readability: An appraisal of research and application*. Ohio State University Press, 1958.

[7] M. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC*, pages 380–388, 2002.

[8] J. P. G. Chimombo. Issues in basic education in developing countries: An exploration of policy options for improved delivery. *Journal of International Cooperation in Education*, 8(1), 2005.

[9] E. Coke and E. Rothkopf. Note on a simple algorithm for a computer-produced reading ease score. *Journal of Applied Psychology*, 54(3):208–210, 1970.

[10] M. Coleman and T. L. Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284, 1975.

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

[12] M. Covington. Idea density – A potentially informative characteristic of retrieved documents. In *IEEE SoutheastCon*, 2009.

[13] M. Crossley and M. Murby. Textbook provision and the quality of the school curriculum in developing countries: Issues and policy options. *Comparative Education*, 30(2):99–114, 1994.

[14] E. Dale and J. Chall. A formula for predicting readability. *Educational research bulletin*, 27(1):11–20, 1948.

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

[16] I. Fang. By computer: Flesch's reading ease score and a syllable counter. *Behavioral Science*, 13(3):249–251, 1968.

[17] R. Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221–233, 1948.

[18] B. Fuller. What school factors raise achievement in the Third World? *Review of educational research*, 57(3):255–292, 1987.

[19] 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.

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

[21] J. Greenfield. Readability formulas for EFL. *Japan Association for Language Teaching*, 26(1), 2004.

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

[23] R. Gunning. *The technique of clear writing*. McGraw-Hill, 1952.

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

[25] S. Heyneman, J. Farrell, and M. Sepulveda-Stuardo. Textbooks and achievement in developing countries: What we know. *Journal of Curriculum Studies*, 13(3), 1981.

[26] M. Hu, E. Lim, A. Sun, H. Lauw, and B. Vuong. Measuring article quality in Wikipedia: models and evaluation. In *CIKM*, 2007.

[27] S. Hwang and W. Merrifield. *Language in context: Essays for Robert E. Longacre*. Summer Institute of Linguistics, 1992.

[28] 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.

[29] T. Kanungo and D. Orr. Predicting the readability of short web summaries. In *WSDM*, 2009.

[30] 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 *International Conference on Computational Linguistics (Coling)*, 2010.

[31] J. Kincaid, R. Fishburne, R. Rodgers, and B. Chissom. Derivation of new readability formulas for navy enlisted personnel. 1975. Research Branch Report 8-75, Naval Air Station Memphis, Millington, Tennessee.

[32] W. Kintsch and J. Keenan. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive psychology*, 5(3):257–274, 1973.

[33] B. Lively and S. Pressey. A method for measuring the vocabulary burden of textbooks. *Educational Administration and Supervision*, 9(389-398):73, 1923.

[34] M. Lockheed and E. Hanushek. Improving educational efficiency in developing countries: What do we know? *Compare: A Journal of Comparative and International Education*, 18(1):21–38, 1988.

[35] I. Lorge. Predicting reading difficulty of selections for children. *Elementary English Review*, 16(6):229–33, 1939.

[36] W. McCall and L. Crabbs. *Standard test lessons in reading*. Columbia University Teachers College Press, 1926.

[37] G. McLaughlin. SMOG grading: A new readability formula. *Journal of reading*, 12(8):639–646, 1969.

[38] R. Mohammad and R. Kumari. Effective Use of Textbooks: A Neglected Aspect of Education in Pakistan. *Journal of Education for International Development*, 3(1), 2007.

[39] B. Mukherjee and N. Puketza. Style in technical writing, UC Davis, 2004. Notes based on the book, J.M.Williams, Style: Ten lessons in clarity and grace.

[40] J. Oakes and M. Saunders. Education's most basic tools: Access to textbooks and instructional materials in California's public schools. *Teachers College Record*, 106(10), 2004.

[41] D. Pennycuick. School Effectiveness in Developing Countries - A Summary of the Research Evidence. *Education Research Papers*, 1993.

[42] J. Redish and J. Selzer. The place of readability formulas in technical communication. *Technical communication*, 32(4):46–52, 1985.

[43] A. Riddell. *Factors influencing educational quality and effectiveness in developing countries: A review of research*. Deutsche Gesellschaft fur Technische Zusammenarbeit (GTZ), Germany, 2008.

[44] A. Saxena, R. Anderson, N. Linnell, U. Sahni, A. Arora, and R. Gupta. Evaluating facilitated video instruction for primary schools in rural india. In *ICTD*, 2010.

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

[46] E. Smith and J. Kincaid. Derivation and validation of the automated readability index for use with technical materials. *Human Factors*, 12(5):457–464, 1970.

[47] W. Taylor. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30:415–433, 1953.

[48] A. Williams, A. Siegel, and J. Burkett. Readability of textual material – A survey of the literature. Technical report, Air Force Human Resources Labortory, 1974.

[49] W. Winterowd. The grammar of coherence. *College English*, 31(8):828–835, 1970.

[50] S. Witte and L. Faigley. Coherence, cohesion, and writing quality. *College Composition and Communication*, 32(2):189–204, 1981.

[51] J. Zhao and M. Kan. Domain-specific iterative readability computation. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 205–214. ACM, 2010.