

Dimensional Standard Alignment in K-12 Digital Libraries: Assessment of Self-found vs. Recommended Curriculum

Byron Marshall
College of Business
Oregon State University
Corvallis, OR

byron.marshall@bus.oregonstate.edu

René Reitsma
College of Business
Oregon State University
Corvallis, OR

reitsmar@bus.oregonstate.edu

Malinda Zarske
College of Engineering
University of Colorado at Boulder
Boulder, CO

Malinda.zarske@colorado.edu

ABSTRACT

Enhancing the experience of digital library users depends, in part, on recognizing and understanding user tasks. In the context of K-12 educational libraries this means that we must understand how K-12 teachers interact with such libraries and how they assess the relevance of documents found or encountered. This paper presents the results of an experiment in which K-12 teachers scored the relevance of curriculum they found themselves and the relevance of documents their colleagues found and recommended. We found that teachers apply a significantly more detailed notion of relevance, both qualitatively and quantitatively, when searching for as compared to evaluating recommended curricula. Differences were observed in both relevance judgments and system interaction logs. These variations may be useful in identifying user intent and in dynamically adapting the behavior of digital libraries of educational material.

Categories and Subject Descriptors

K.3.1 [Computer Uses in Education]

General Terms

Measurement, Reliability, Experimentation, Human Factors.

Keywords

Curriculum-standard alignment, Inter-rater reliability, Relevance, Digital library, Context-specific measurement.

1. INTRODUCTION

Technical and institutional changes are causing a proliferation of educational materials on the Web. Recent U.S. congressional and National Science Foundation (NSF) initiatives such as the National Science Digital Library [14] and GK-12 [8] have contributed to a rapidly growing community of digital libraries of K-12 math and science materials. Engineering is Elementary, TeachEngineering, Teacher's Domain, Middle School Portal, Curriki, NetTrekker, and NSDL reference collections ranging from a few hundred to several hundred thousand learning objects.

'Educational' or 'teaching standards' drive much of the use of

these resources as teachers must teach to standards in their locality [6, 9, 12]. Achievement Standard Network [5] estimates that there are about 65,000 periodically revised K-12 mathematics and science standards. The combination of growing collections and expanding and changing lists of standards has created an important challenge for these collections; namely, to establish and maintain 'alignments' between objects and standards.

Following a famous statement by L. Wittgenstein [13]: "*to understand a proposition is to know what is the case when it is true,*" we ask "*under which circumstances are curriculum and standard aligned?*" If part or much of the value of a K-12 digital library is derived from the ability of teachers to find 'aligned' curriculum then we should carefully consider the meaning of 'alignment.' An enhanced understanding may be useful both in improving document retrieval and in creating user interfaces that recognize and adapt to important patterns of user behavior.

We asked 43 K-12 teachers and teachers in training to assess the alignment of documents to standards on nine different alignment aspects, each of them anchored in the everyday life of K-12 teaching. The experiment primarily explored how improved levels of inter-rater reliability can be achieved and was intended to generate a set of 'correctly' aligned documents (a gold standard set) for testing and improving lesson plan retrieval. However, the experimental design also created a context for studying and comparing the behavior and judgment of users as they themselves search for aligned documents vs. how they act when evaluating the alignment of document/standard pairs suggested by others.

We found that the notion of 'alignment' used by the teachers was both qualitatively and quantitatively different when they were searching for aligned curriculum themselves vs. judging alignment suggestions identified by others. These differences were present in both the alignment judgments made by teachers, and in the interactions the teachers had with the digital library as documented in system interaction logs.

This article describes our experiment, discusses the differences between the two modes of assessment, and looks at how they are reflected in the system log. We conclude by discussing the implications of our findings for the definition and operationalization of 'alignment' in K-12 digital libraries.

2. 'ALIGNMENT' EXPERIMENT

Reported low inter-rater reliability in K-12 alignment [4] and other relevance experiments, e.g., [1, 7], prompted us to consider the alignment concept. Taking the work of Saracevic and others to

heart [2, 3, 10], we operationalized the vague and ambiguous notion of ‘alignment’ into nine specific aspects or dimensions in the context of K-12 teaching (Table 1).

Table 1: Relevance clues operationalized in K-12 math and science classroom teaching dimensions.

Alignment clue	Alignment dimension	Statement
Affective Match	Motivation	The document contains materials that are motivational or stimulating (interesting, appealing, or engaging) for students.
Content Match	Concepts	The document includes concepts, keywords, terms, and definitions from the standard.
Content Match	Background	The document provides interesting/important background material related to the standard.
Object Match	Grade level	The grade level of this material is appropriate for this task or else I can easily adapt the materials in this document to my grade level.
Situational Match	Non-textuals	I can use a non-textual component(s); e.g., figures, tables, images, videos or graphics, etc.
Situational Match	Examples	I can use the real-world examples provided in the document in class.
Situational Match	Hands-on	I can use one or more of the hands-on, active engineering activities.
Situational Match	Attachments	I can use some of the attachments; e.g., score sheets, rubrics, test questions, etc.
Situational Match	References	I can use references or Internet links to relevant materials elsewhere.
Overall alignment		Overall, I consider this document relevant for this teaching assignment.

Based on pilot results [9], we hypothesized that clarifying and explicating the meaning of ‘alignment’ into several aspects would reduce the latitude of raters in interpreting the concept, raising the inter-rater reliability of alignment judgments, and that ‘overall’ alignment could be modeled as a function of these aspects. Notably only two of the nine dimensions of Table 1 address lexicographic content; the other seven dimensions address other aspects that are likely to factor into the alignment decision.

Forty-three K-12 math and science teachers and teacher trainees from several parts of the U.S. spent three hours in which they:

1. Received basic training in how to navigate a K-12 math and science digital library (TeachEngineering [11]).
2. Assumed hypothetical teaching tasks based on an existing standard; e.g., Colorado, Grade 6: “*Physical Science: Students know and understand common properties, forms, and changes in matter and energy. (Focus: Physics and Chemistry).*”
3. Searched the TeachEngineering collection for curriculum that they considered supportive in teaching this task.
4. Rated the curriculum they found on ten alignment scales (Table 1). These results were recorded as ‘*search phase*’ results.
5. After a break in the session, scored the alignment of standard-curriculum pairs found by others. We will refer to these as ‘*recommended phase*’ results.

Document/teaching-task pairs were assessed by participants using six-point Likert agreement scales (strongly agree – agree –

somewhat agree – somewhat disagree – disagree – strongly disagree) and a ‘Not Applicable’ category. In addition, a control scale rating the teacher’s self-assessed ability to rate the alignment was offered. Assessments associated with a poor score on this control scale were eliminated from analysis.

Data were collected in sessions at Oregon State University (5 subjects), University of Colorado (16), Worcester Polytechnic Institute (15), and Duke University (7). In total, 247 scalings (a documents/standard pair scored by a person) were obtained during the search phase and 708 during recommended phase.

Participants’ online requests to the TeachEngineering library were recorded during the experiment. A ‘document,’ as used here, refers to a lesson plan or activity stored and indexed by TeachEngineering. Documents often refer to other resources such as worksheets or handouts (attachments) or to other Web resources through links. Available logs list access to documents, searches run, access to other site pages, and views of attachments.

3. DOES IT MATTER WHO FOUND IT?

Figure 1 shows the 45 pairwise correlations between the various alignment measures (nine alignment aspects and ‘overall’ alignment), contrasting search (solid line) with recommended (dashed line) phase scalings. (Due to space limitations, the figure might not print well; however it renders nicely in a PDF viewer) Two patterns stand out. First, on average, the correlations for the search scalings are lower than for the recommended ones. Hence, when searching, the teachers’ alignment dimensions were less correlated to each other than when assessing recommendations. The means—Fisher Z-transformed because the variables are correlations—were shown to be significantly different in a t-test. Secondly, ‘overall’ assessments did not fit this pattern; four of the five largest exceptions (between one and three o’clock in the figure) involved ‘overall alignment.’ That is, in search phase responses ‘overall alignment’ showed a stronger correlation with attachments, reference links, activities and non-textuals.

Table 2 also suggests an increased influence of activities and attachments on overall alignment during the search phase. The table contains several multiple linear regression models with ‘overall’ alignment as the dependent variable and statistically significant independent aspect variables. Note that when we once again separate the search phase from the recommended phase scalings, two different models emerge. Both models have a reasonable fit (78.8% explained variance ($R^2 = .788$) for the search phase model; and 73.3% for the recommended phase model) and both have an important role reserved for the alignment between concepts, background material and grade adjustability. However, the models are different where it concerns the role of hands-on exercises and activities (‘activities’) and the presence of ready-to-use worksheets or ‘attachments.’ Whereas these alignment aspects function significantly in the search phase—together they explain 13.5% of the variance in ‘overall’ alignment—they do not factor into the recommended phase.

So far, we have argued that despite close correspondence between the scalings of alignments from the search and recommended phases of the experiment, respondents exhibited a different notion of ‘overall alignment’ while conducting those scalings. One might wonder if that difference is also reflected in how respondents interacted with the digital library. Did they behave differently as well? And if so, is that behavior consistent with their judgments?

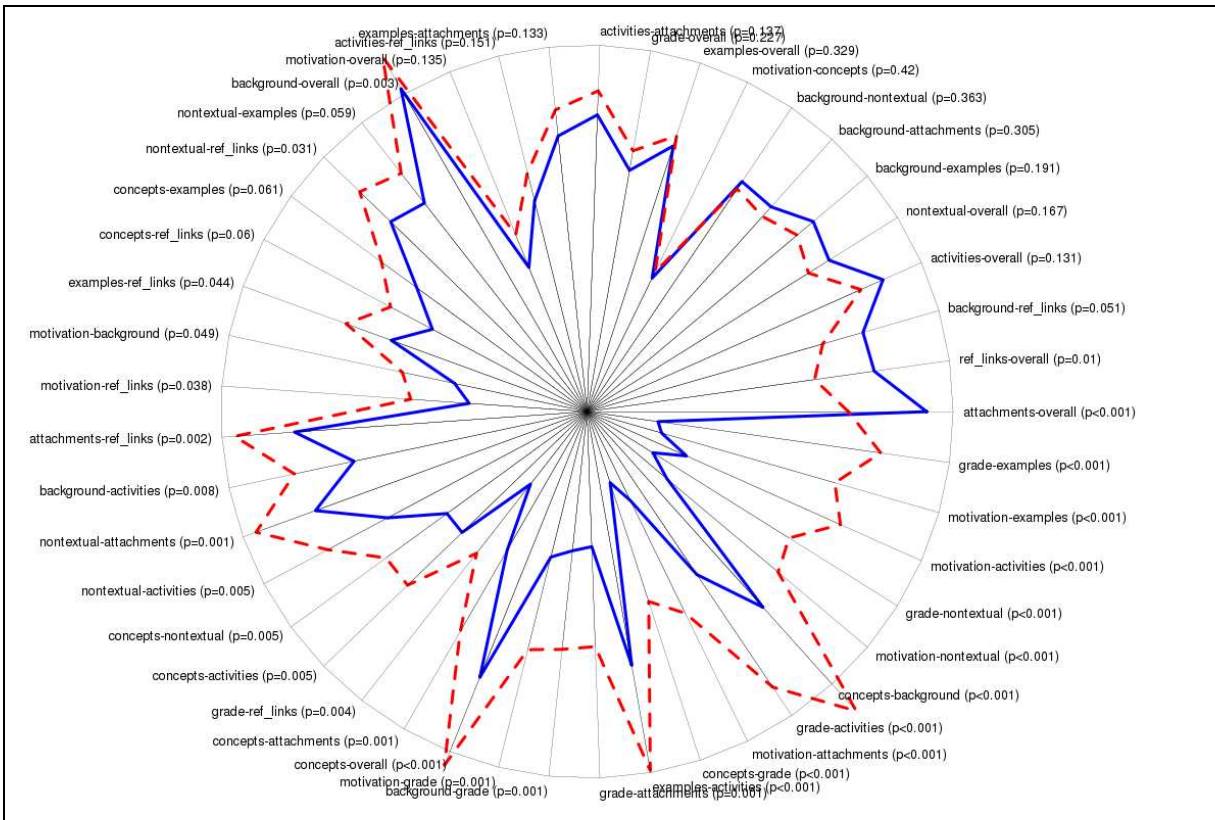


Figure 1: Correlations between alignment dimensions: search (solid line) phase vs. recommended (dashed line) phase. Correlations are plotted starting at three o'clock, clockwise in descending order of difference ($r_{\text{search}} - r_{\text{recommended}}$). p values give the probability that associated correlations are identical.

Table 2: Multiple linear regression models of ‘overall’ alignment. All coefficients shown are significant at $p < 0.01$.

	Search		Recommended	
	Coefficient	R ² cum.	Coefficient	R ² cum.
Intercept	-.559		-.064	
Concepts	.227	.338	.343	.608
Background	.346	.564	.474	.703
Grade level	.295	.653	.159	.733
Activities	.199	.705		
Attachments	.788	.788		
Examples			.123	.734

Table 3 enumerates actions found in the system interaction logs. Because respondents rated the alignment of a single document to several (similar) teaching tasks, we use the number of unique respondent-document combinations as a unit of analysis. Of these, 175 were recorded during the search phase of the experiment; 210 were recorded during the recommended phase. Not surprisingly, respondents visited more pages when searching: 2,531/175=14.47 Web page requests per respondent-document combination when searching vs. 1,225/210=5.8 when evaluating recommendations. Searchers must ‘home in’ on documents by running queries and rendering various documents whereas evaluating recommendations requires much less interaction. Similarly,

respondents conducted more document renderings when searching (978/175=5.59) than when evaluating recommendations (795/210=3.79). The unexpectedly high 3.79 recommended phase renderings resulted from participants who frequently viewed a similarly-named document before finding the correct one.

Table 3: Frequency of library requests by phase.

Request Frequency	Search phase	Recommended phase
Web pages	2,531	1,225
Document views	978	795
*.doc or *.pdf	243	197
worksheet	100	91

Much more interesting, however, are the access frequencies for ‘attachments’ (ready-to-use worksheets that teachers hand out to their students during class). Recall that ‘attachments’ made a significant contribution to ‘overall’ alignment among searchers (Table 2). Hence, it would be interesting to see if this would correspond with, for instance, a higher access frequency for those documents during the search phase of the experiment.

Table 3 lists two operationalizations of requests for such attachments: 1) requests for *.doc and *.pdf files and 2) requests with the string “worksheet” in the filename. Either of these measures corresponds well with the frequency of these documents in the TeachEngineering collection. The data in Table 3 show a

clear difference between the access frequencies of these documents for the two phases. Respondents logged significantly higher request frequencies for these documents when searching than when evaluating recommendations (243/175=1.40 vs. 197/210=.94 and 100/175=.57 vs. 91/210=.43).

4. DISCUSSION & CONCLUSION

Our experiment aimed primarily to devise a multidimensional definition of 'alignment.' However, the experimental design allows us to also explore differences between alignment assessment and behavior. On the face of it, one might not expect much of a difference; after all, why would teachers use different criteria or weigh identical criteria differently depending on whether they are evaluating curricula they are searching for themselves or evaluating curricula recommended by others. The data from our experiment, however, show clear and statistically significant differences between these two evaluation contexts, even when these tasks are conducted by the same people, under identical experimental conditions, scoring the same alignment scales. These differences are also reflected in the interactions these users have with the system.

Digital libraries frequently provide both search and recommender services. These services often rely on machine learning to identify promising items. Our results suggest that users have different models of alignment for search and recommendation modalities. Such differences can inform the development of reference collections used to train recommendation services and provide real-time behavioral clues from which user intent can be inferred. It may be possible to improve the selection and/or presentation of query results based on such clues.

Alignment and retrieval mechanisms that focus on the 'concepts' and 'background' dimensions of standard/curriculum alignment leverage the two strongest contributors to 'overall alignment': 70% when evaluating recommendations and 56% when searching. Nevertheless, such a focus ignores factors that explain 44% of the variance shown in our search phase model. Therefore, we suggest that alignment methods and retrieval mechanisms include other more practical, or as Saracevic might say, 'situational,' aspects.

Clues for these additional alignment indicators may be available from a variety of sources. In our results the number of supplemental documents viewed indicated a user's task with implications for their decision model. Folksonomy tagging where users assess documents may provide qualitatively different clues as contrasted with term-based approaches. And feedback mechanisms where users can interact with additional data such as the number of attachments may be useful in helping teachers find documents they would adopt. While it is not yet clear how such indicators can be integrated, our results point out the need for continued exploration of methodologies that combine multiple task and relevance clues.

5. ACKNOWLEDGMENTS

This work was sponsored, in part, by the U.S. National Science Foundation, Grant #0532709.

6. REFERENCES

[1] Bar-Ilan, J., Keenoy, K., Yaari, E., Levene, M. (2007) User Rankings of Search Engine Results. *Journal of the American*

Society of Information Science and Technology. 58. 1254-1266.

[2] Borlund, P. (2003) The Concept of Relevance in IR. *Journal of the American Society of Information Science and Technology*. 54. 913-925.

[3] Cosijn, E., Ingwersen, P. (2000) Dimensions of Relevance. *Information Processing and Management*. 36. 533-550.

[4] Devaul, H., Diekema, A.R., Ostwald, J. (2007) Computer-assisted Assignment of Educational Standards Using Natural Language Processing. Annual Meeting of the National Science Digital Library (NSDL). Arlington, VA.

[5] Gateway (2007). NSDL:ASN Achievement Standards Network. Available: <http://www.thegateway.org/asn>.

[6] Jay, M., Longdon, D. (2003) Death, Taxes and Correlations: A Primer on the State of Correlation in the K-12 Education. *Upgrade, SIIA*. 20-21.

[7] Kim, G. (2006) Relationship Between Index Term Specificity and Relevance Judgment. *Information Processing and Management*. 42. 1218-1229.

[8] NSF (National Science Foundation) (2008) Graduate Teaching Fellows in K-12 Education Program. Available at <http://www.nsfkg12.org>. Accessed 12/03/2008.

[9] Reitsma, R., Marshall, B., Dalton, M., Cyr, M. (2008) Exploring Educational Standard Alignment. In Search of 'Relevance.' Proceedings of the Joint Conference on Digital Libraries (JCDL'08), Pittsburgh, PA. IEEE-ACM.

[10] Saracevic, T. (2007) Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part III: Behavior and Effects of Relevance. *Journal of the American Society of Information Science and Technology*. 58. 2126-2144.

[11] Sullivan, J.F., Cyr, M.N., Mooney, M.A., Reitsma, R.F., Shaw, N.C., Zarske, M.S., Klenk, P.A. (2005) The TeachEngineering Digital Library: Engineering Comes Alive for K-12 Youth. Proceedings of the ASEE Annual Conference; Portland, OR. ASEE, Washington D.C. Available: http://www.teachengineering.org/documents/ASEE%202005-851_TE_Final.pdf. Accessed: 12/03/2008.

[12] Sumner, T. Ahmad, F., Bhushan, S., Gu, Q., Molina, F., Stedman, W., Wright, M., Davis, L., Janée, G. (2005) Linking Learning Goals and Educational Resources Through Interactive Concept Map Visualizations. *International Journal on Digital Libraries*. 5. 18-24.

[13] Wittgenstein, L. (1922) *Tractatus Logico Philosophicus*. Harcourt, Brace & Company, Inc., New York, NY. K. Paul, Trench, Trubner & Co., Ltd., London, UK.

[14] Zia, L.L. (2002) The NSF National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) Program. *New Projects in Fiscal Year 2002. D-Lib Magazine*. 8. Available: <http://www.dlib.org/dlib/november02/zia/11zia.html>. Accessed 12/03/2008.