Exploring Educational Standard Alignment: In Search of 'Relevance'

René Reitsma
College of Business
Oregon State University
Corvallis, OR

reitsmar@bus.oregegonstate.edu

Byron Marshall
College of Business
Oregon State University
Corvallis, O OR

byron.marshall@bus.oregegonstate.edu

Michael Dalton
College of Education
Oregon State University
Corvallis, OR

michael.dalton@oregonstate.edu

Martha Cyr
K-12 Outreach
Worcester Pol. Institute
Worcester, MA

mcyr@wpi.edu

ABSTRACT

The growing availability of online K-12 curriculum is increasing the need for meaningful alignment of this curriculum with state-specific standards. Promising automated and semi-automated alignment tools have recently become available. Unfortunately, recent alignment evaluation studies report low inter-rater reliability, *e.g.*, 32% with two raters and 35 documents. While these results are in line with studies in other domains, low reliability makes it difficult to accurately train automatic systems and complicates comparison of different services. We propose that inter-rater reliability of broadly defined, abstract concepts such as 'alignment' or 'relevance' must be expected to be low due to the real-world complexity of teaching and the multidimensional nature of the curricular documents. Hence, we suggest decomposing these concepts into less abstract, more precise measures anchored in the daily practice of teaching.

This article reports on the integration of automatic alignment results into the interface of the TeachEngineering collection and on an evaluation methodology intended to produce more consistent document relevance ratings. Our results (based on 14 raters x 6 documents) show high inter-rater reliability (61 - 95%) on less abstract relevance dimensions while scores on the overall 'relevance' concept are (as expected) lower (64%). Despite a relatively small sample size, regression analysis of our data resulted in an explanatory (R^2 = .75) and statistically stable (p-values < .05) model for overall relevance as indicated by matching concepts, related background material, adaptability to grade level, and anticipated usefulness of exercises. Our results suggest that more detailed relevance evaluation which includes several dimensions of relevance would produce better data for comparing and training alignment tools.

Categories and Subject Descriptors

K.3.1 [**Computer Uses in Education**]

General Terms

Measurement, Design, Reliability, Experimentation, Human Factors, Theory.

Keywords

Curriculum-standard alignment, Inter-rater reliability, Relevance, Digital library, Social science theory, Reification, Context-specific measurement.

1. INTRODUCTION

Growth in the online availability of K-12 curriculum, for instance through digital libraries such as the National Science Digital Library or NSDL [6, 7], is driving an increased need for alignment of these curricula with the educational standards to which K-12 teachers must teach. Standard alignment, however, is a nontrivial problem for several reasons. First, many standards exist—there are over 90,000 mathematics and science standards in the United States—and they frequently change. Approximately 10 US states change their educational standards each year, so a given standard has an average life expectancy of about five years. To help manage this dynamic environment, several organizations and projects have created standards databases and automated standards correlation and alignment services. For instance, the Achievement Standards Network project made available by the nonprofit JES&Co [4] manages an XML version of all US K-12 standards. Similarly, the Curriculum Alignment Tool (CAT) and Standard Alignment Tool (SAT) made available by the Center for Natural Language Processing at Syracuse University [3] are Web services that provide on-the-fly document-to-standard alignment (CAT) and standard-to-standard alignment (SAT).

Although the sheer massiveness and dynamism of the standards alignment problem has been addressed by services such as ASN, CAT and SAT, creating consistent and accurate alignments is still a challenging problem. Devaul *et al*. [2], for instance, report that inter-rater reliability (the rate of agreement between different evaluators) for curriculum-standards alignments tends to be low, and hence, that these alignments might or might not be a good resource for standards-based searching. Similarly, Bar-Ilan *et al*. [1] report on low inter-rater reliability of generic search engine results.

Low inter-rater reliability causes two problems. First, it puts in doubt the validity of the curriculum-standard alignments generated by some of the automated tools. Second, it bodes poorly for the development of a gold standard or agreed-upon set of curriculum-standard matches against which alignment mechanisms can be tested or with which automated algorithms can be trained.

In this article, we first present an example of how the new automated curriculum-standard alignment tools are used to create elegant and flexible systems. We follow this by taking a more

detailed look at inter-rater reliability of curriculum-standard alignments. We propose that at least some of the poor scores are caused by imprecise measurement methods; *i.e.*, given the way in which people were asked to judge the validity of the alignments, it is not surprising that the results were poor.

Next, we propose a different method for validating curriculum-standard alignments. In this method, we separate the 'validity,' 'appropriateness,' or 'relevance' of a document given a standard into a number of sub components, or aspects, based on a typology of 'relevance clues' offered by Saracevic [14].

We then present the results of a preliminary study in which we asked 14 people to judge these more detailed relevance aspects and observe that, as expected, the inter-rater reliability of overall 'relevance' judgment was lower than that of most of the more precise relevance aspects.

Next, we present the results of a regression analysis which attempts to model overall relevance as a linear combination of one or more of the more specific relevance aspects. Although the model explains 75% of the total variation in overall relevance, only four of the nine relevance aspects have statistically significant coefficients, despite having very high inter-rater reliability. We interpret this result to support a strong need for more detailed and more differentiated ways of measuring the quality or validity of curriculum-standard alignments.

2. TEACHENGINEERING, ASN, CAT & SAT

Automated curriculum-standard alignment services such as CAT and SAT provide attractive means for educational digital libraries to align their contents with standards. For instance, www.teachengineering.org (Figure 1) is a digital library of K-12 mathematics and science engineering curriculum. At the time of this writing (Jan. 2008) it contains 429 hands-on curricular activities and 235 lessons.

All curricular items in *TeachEngineering* have the same structure, with mandatory and optional components, and the same look and feel. All items are indexed against a search engine that can retrieve lesson material based on criteria such as keyword, target grade, available lesson time, cost, and most important, alignment with educational standards.

The alignments are based on four sources of information:

1. The *Alignment Standard Network (ASN)* is an XML–based database of all K-12 educational standards in the US. The database is maintained by the nonprofit JES&Co. As of the time of this writing, the ASN contains over 90,000 mathematics and science standards. These standards are grouped into 55 separate sources or 'authors,;' the 50 states and some 'national' standard bodies such as AAAS, McREL, NSES, or ITEA. Each of these forms a hierarchy, with paths through them representing educational standards. The ASN contains over 80,000 such paths. Figure 2 shows the path for a third grade Maryland physics standard.
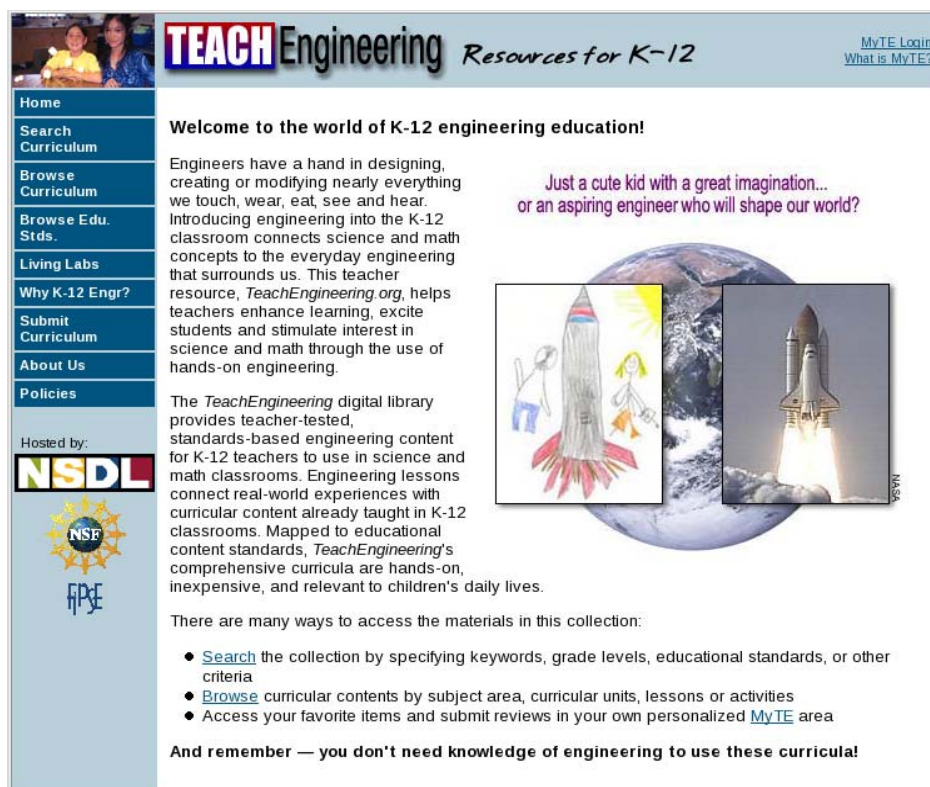

Figure 1. **www.teachengineering.org** home page.

Maryland: Science [2005]

[i] 5.0 Physics - Students will use scientific skills and processes to explain the interactions of matter and energy and the energy transformations that occur
(Grades K - 8)

[i] A. Mechanics
(Grades K - 8)

[i] 2. Explain that changes in the ways objects move are caused by force
(Grades 3 - 3)

Currrent Standard

• c. Observe and describe that objects fall to the ground unless something holds them up (gravity).
(Grades 3 - 3)

Figure 2. ASN representation of a Maryland grade 3 physics standard.

2. *TeachEngineering* curriculum authors align each individual item with one or more standards. Most often they select standards from their home state, but sometimes they select standards from one or more national standard bodies.

3. The Center for Natural Language Processing (CNLP)'s *Curriculum Alignment Tool (CAT)* provides a Web service interface to which curricular items can be submitted along with a request to align them with ASN standards from one or more standard bodies. TeachEngineering's entire curriculum was submitted to CAT with the request to find at most five applicable mathematics and science standards from each ASN standard source (*e.g.*, Alaska).

4. CNLP also offers a web service interface to its *Standard Alignment Tool (SAT)* to which educational standards can be submitted along with a request to find aligned ASN educational standards from one or more standard bodies. All standards in the ASN are in the process of being submitted to SAT with the request to align them with at most five standards from all of the standard sources in the ASN. At the time of this writing this has amounted to 13,555,647 alignments with about 60% of all standards processed.

5. With this information, a new, soon to be released version of *TeachEngineering* is being built through which users can submit requests to find curriculum that supports a specific educational standard. The goodness-of-fit between the standard and the curricular document is computed as follows:

$$(\sum_{s=0}^{S} (Alignments_s \times W_s)) \times S$$

Where:
• $Alignments_s$ = number of alignments through source s,
• W_s = weight or priority of source s,
• S = number of different sources involved in the alignment.

The index allows for differential weighting of alignment sources and favors multiple-source alignments (*e.g.*, CAT alignments that match up with manual alignments) over single-source ones. Figure 3 shows an example of results returned for a search for curriculum aligned with a 2000 Maryland standard concerning the Pythagorean theorem.

It is worth noting that the alignments used in the new TE standard-based retrieval interface (both manual and computer generated) were performed from the curriculum developer's perspective. That is, we record which standards are relevant to a document. But then, the stored relations are used in reverse for retrieval: which documents relate to this standard? The system infers that if two standards are aligned, then documents relevant to the one standard will also tend to be relevant to the other. This approach is promising because it effectively utilizes the time of the curriculum developer, the utility of existing ranking systems, and established standard-to-standard mappings. Some of the implications of this approach are discussed in [10]. It is not clear that the relationships are symmetric enough for this to be an optimal solution. While it seems this that approach adds value, these concerns may affect the accuracy of results.

3. WHAT IS A GOOD ALIGNMENT?

With automated curriculum-standard alignment tools being so new, little evidence on the quality of the found alignments is available. However, Devaul *et al*. [2] present the results of an inter-rater reliability study of automated curriculum-standard alignment assessments. Thirty-five learning resources were independently aligned by two human experts with the help of the CAT system mentioned above. Despite the fact that these people were experts and were assisted by a computerized tool that suggests aligned standards given a learning resource, the inter-rater reliability between the two subjects was low. On average only 32% of the alignments were shared by the two experts. Agreement was higher (40%) for standards addressing more abstract subjects and lower (18%) for more precise, application-oriented standards. The authors thus conclude that "*Assignment of standards is an extremely subjective task ... even with significant calibration efforts, especially when the standards being assigned are numerous and wide ranging in scope, ...strong inter-rater reliability can be difficult to achieve.*"

Moreover, the authors specifically state that tools such as CAT are not meant to be used in an unsupervised fashion. Indeed, CAT's alignment engine is meant to be trained through feedback from its users so that it can learn how to better align standards and curriculum for that user.

Maryland: Math [2000]

ℹ️☑ 2.0 Knowledge of Geometry -- Students will apply the properties of one-, two-, and three-dimensional geometric figures to describe, reason, and solve problems about shape, size, position, and motion of objects.
(Grades K - 12)

ℹ️☑ 2.8.2a use the properties of angles and triangles
(Grades 6 - 8)

Currrent Standard

• use the Pythagorean theorem to solve problems by determining the missing side of a right triangle (MLO 2.3)
(Grades 6 - 8)

The following curriculum was selected by the CAT service as correlated to this standard.
Correlated Activities   **Correlated Lessons**
  ◆ Stay in Shape   ◆ Navigating by the Numbers
  ◆ Trig River

The following curriculum was selected by the SAT service as correlated to this standard.
Correlated Activities
  ◆ Groundwater Detectives
  ◆ Too Much Pressure!

Figure 3. CAT and SAT TeachEngineering matches for a 2000 Maryland standard addressing the Pythagorean theorem.

Although in a different domain, similar low levels of interreliability were found by Bar-Ilan *et al*. [1] when they asked people to rank the quality of the results returned by a set of well-known search engines such as Google, Yahoo, and MSN Search. Here too, "*the similarities between the users' choices and the rankings by the search engines are low.*"

We propose that these results are not really surprising. Of course, we all would like to make available search engines which find universally 'good' matches retrieved with procedures which accurately and reliably map the 'true' meaning of a concept or question to an objective set of document attributes. Indeed, this is what search engine providers hope to find when they conduct their inter-rater reliability tests: that our machines and programs interpret the world more or less the way we do when we enter a search query. Moreover, that we agree with each other and with the machines on what it all means and how it all corresponds to the neutrally observable world surrounding us.

However, this objective notion of meaning, that we can reliably and accurately map the meaning of words, constructs, sentences and concepts to observable quantities, has long been questioned by social theorists. So-called relationalist thinkers in particular have argued against concepts having true and independent meanings. Following Van der Smagt [15], we quote Noble [11]: "*Objects are realized, ...they do not exist in their own right, but rather become existent in virtue of the organismic agency in relation to them.*" Similarly, Greenwood [5] points out that "*social ...phenomena are relational in nature; their identity is determined by their relation to other social phenomena.*"

According to these theorists, meaning is not inherent in the intersubjectively observable attributes of things, but is instead dynamically created from the context and situations in which that meaning is used and addressed. If the context and/or the actors change, so does the meaning of the concept. Hence, a curriculum-standard alignment might be perfectly acceptable in one context yet unacceptable in another. In the words of Bar-Ilan *et al*.: "*there is no average user, and even if the users have the same basic knowledge of a topic, they evaluate information in their own context…*"

The latter statement poignantly illustrates Van der Smagt's warnings against the temptation to reify; *i.e.*, to consider abstract concepts to refer to a true and objective reality. He quotes Levins and Lewontin [9]: "*Abstraction becomes destructive when the abstract becomes reified and when the historical process of abstraction has been forgotten so that the abstract descriptions are taken for descriptions of the actual objects.*" In this view, Bar-Ilan *et al*.'s statement that "*there is no average user*" should be taken quite literally.

We do not, however, even really have to consider ontological issues such as whether or not average users exist, because earlier, more pragmatic arguments apply here as well. For instance, the 1954 work of Lazarsfeld [8] pointed out that the same concept, measured in different contexts, must be operationalized differently; *i.e.*, the observable variables by means of which a concept is measured differ in different contexts. The separation of the structural model from the measurement model, a common approach in social science research, is a direct representation of this notion of context-specific measurement.

Following these arguments, we submit that the poor inter-rater reliability reported in [2] and [1] could well be a methodological artifact; *i.e.*, by not referring or appealing to the different and context-specific dimensions of 'relevance,' subjects partaking in the measurement are asked to formulate their own interpretation of these concepts. That many of them do so differently can be expected.

3.1 The Quest for Relevance

In a three-part series of papers, Saracevic [12, 13, 14] explores similar problems and theories associated with the concept of 'relevance' as it applies to information retrieval and information science. He suggests that information retrieval evaluations are often built on five postulates which imply that relevance is:

- Topical – (based on the concepts presented),
- Binary – (learning objects are either relevant or not),
- Independent – (objects can be judged independently),
- Stable – (judgments do not change over time), and
- Consistent – (raters agree on relevance).

We observe that these postulates are usually built into standards alignment evaluation. Indeed, it is difficult to evaluate the accuracy of a retrieval system without making some of these assumptions. While Saracevic's literature survey notes that a general theory of relevance is ellusive, we employed two frameworks drawn from his analysis which we believe can improve the inter-rater reliability (and hopefully thereby the usefulness) of standard alignment results.

A "stratified model of relevance interactions" [13] depicts how stored information is usefully connected to users through several computerized strata (content, processing, and engineering), a surface level interface, and user strata labeled query, cognitive, affective, and situational. These interactions occur within a social and cultural context. The model highlights the importance of including a variety of user concerns beyond conceptual similarity in retrieving learning objects that relate to a given educational standard. It also emphasizes the importance of looking at the retrieval process as a whole. For example, the processing (software, algorithms, *etc*.) used would ideally adapt the cognitive, affective, and situational context a user brings to a query. While previous efforts in the standards alignment literature focus on selected parts of this process, we propose that a more comprehensive view is needed in measuring relevance if consistent ratings, useful training data, and improved matching systems are to be obtained and developed.

Saracevic [14] distilled a list of 'clues' from previous studies used to make relevance inferences. His list includes content, object, validity, use or situational match, cognitive match, and belief match. In developing the survey instrument for this work, we sought to collect clues in several of these areas. 'Content' issues are most commonly considered in standards alignment work when the topics or concepts from a standard are compared to the subject and/or details of a learning object. Educational digital library 'object' issues include cost of implementation, learning object type, and formatting. The 'validity' (trustworthiness) of an educational resource is very important to educational practitioners. We propose that 'situational match' (value in use) issues may be an important source of the inconsistency between people who rate the alignment of educational materials and standards. 'Affective match' concerns involve emotional response to the identified items. For example, teachers want to choose learning objects that will engage their students. Documents which do not align with the 'beliefs' of an instructor are likely to be judged as inappropriate for use in the classroom.

4. EXPERIMENT: TEACHENGINEERING INTER-RATER RELIABILITY

With the general goal of improving the standard-to-document search functionality of the TeachEngineering system, we developed a relevance evaluation survey which we hoped would provide reasonably consistent and useful results. Accurate assessments of relevance could be employed directly in retrieval and might also be used to train or guide relevance inference algorithms. Following the theoretical arguments against abstract and reified notions of relevance, we hypothesize that inter-rater reliability of abstract notions of the quality of curriculum-standard alignments will be worse than that for more precise, part-worth aspects of relevance. Similarly, asking subjects about curriculum-standard alignments *in the context of an actual teaching task* should remove a significant amount of the ambiguity and randomness of the 'quality' or 'relevance' concept. After all, being asked to teach a lesson for a sixth grade standard puts some definite restrictions on the types and level of materials that one would include in such a lesson.

To explore this, we devised a small experiment and asked 14 people familiar with the *TeachEngineering* system to rate the quality of curriculum-standard matches using two educational standards with three curricular items for each standard: a total of six curriculum-standard alignments.

Rather than asking these subjects for a simple relevance or quality judgment, the rating was solicited in the context of a hypothetical teaching task; *i.e.*, subjects were asked to imagine that they, as K-12 science teachers, were asked to teach the learning contents of two different educational standards. For each of these standards the subjects were given three learning objects from the *TeachEngineering* collection, the contents of which could possibly be used to teach to the standard.

The two standards and six documents (listed below) were chosen such that the quality of the document-standard match was neither obviously high, nor obviously low; *i.e.*, although each of the documents had at least some conceptual overlap with the standard to which it was aligned, grade levels of the teaching task, hands-on activities, examples, *etc*., varied sufficiently to allow for ample inter-rater unreliability in assessment of overall alignment quality. For instance, several 9th-grade level documents were combined with a 3rd grade teaching tasks.

Teaching Task 1:

As a 3rd-grade Massachusetts teacher you are assigned to teach material related to the standard "Relate earthquakes, volcanic activity, mountain building, and tectonic uplift to plate movements." You have two hours of class time to spend on instruction.

Task 1 Learning Objects:

Earthquakes Rock: http://www.teachengineering.org/view_lesson .php?url=http://www.teachengineering.com/collection/cub_/lessons/ cub_natdis/cub_natdis_lesson03.xml

Soapy Stress: http://www.teachengineering.com/view_activity.php? url=http://www.teachengineering.com/collection/cub_/activities/cub _rock/cub_rock_lesson01_activity1.xml

Tsunami Attack!: http://www.teachengineering.org/view_lesson .php?url=http://www.teachengineering.com/collection/cub_/lessons/ cub_natdis/cub_natdis_lesson06.xml

Teaching Task 2:

As a 10th-grade Colorado teacher you are assigned to teach material related to the standard "Students know and understand how organisms change over time in terms of biological evolution and genetics." You have four hours of class time to spend on instruction.

Task 2 Learning Objects:

Mice Rule (Or not): http://www.teachengineering.org/view_lesson .php?url=http://www.teachengineering.com/collection/duk_/lessons/ duk_evolution_mary_less/duk_evolution_mary_less.xml

Fantastic Fossils: http://www.teachengineering.org/view_lesson.php? url=http://www.teachengineering.com/collection/cub_/lessons/cub_r ock/cub_rock_lesson03.xml

Population Density; How much Space do you Have?!: http://www.teachengineering.org/view_activity.php?url=http://www. teachengineering.com/collection/cub_/activities/cub_bio/cub_bio_le sson01_activity1.xml

The subjects were then asked to express their level of agreement with a series of nine statements, addressing different aspects of the document's 'relevance.' The dimensions were chosen to correspond with the various relevance clues recognized by Saracevic [14]. For instance, Saracevic's 'affective match' clue was operationalized with the statement "*The document contains materials that are motivational or stimulating (interesting, appealing, or engaging) for students*." In addition to these nine aspect dimensions, users were asked to express the overall relevance of the document for the teaching task. Table 1 shows the mappings between Saracevic's clue types, curriculum-standard mapping dimensions and their operationalizations in our survey instrument. We did not include items for 'belief match' or 'validity' because our test bed of TeachEngineering documents is carefully vetted for consistency on these dimensions. These components, however, might be important in other contexts.

Level of agreement was expressed as a six-point Likert scale: 'strongly agree,' 'agree,' 'somewhat agree,' 'somewhat disagree,' 'disagree,' 'strongly disagree,' with a 'not applicable' answer category added.

We defined inter-rater reliability as the percentage of 'same' scores between any two subjects, where 'same' was defined in three different ways:

1. IRR-1: both subjects score on the same side of the scale; *i.e.*, both score either 'strongly agree,' 'agree,' or 'somewhat agree,' or both score 'somewhat disagree,' 'disagree,' or 'strongly disagree.'

2. IRR-2: same as IRR-1 except that answers may not differ with more than one scale point.

3. IRR-3: both subjects answer the question identically.

'Not applicable' answers were excluded from the computations.

With 14 subjects participating, this resulted in 91 pairwise comparisons per relevance aspect for each of the six curriculum-standard alignments. However, in five cases raters expressed doubts about their ability to teach to one of the standards. The scores for those alignments were removed from the dataset.

4.1 Inter-rater Reliability Results

Figure 4 shows the various inter-rater reliabilities for each of the 10 relevance measures. Saracevic [14] reports that relevance judgment overlaps in various studies range between 30% and 80% depending on the experimental parameters. Previously reported results are at the bottom of this range and the binary version of our results is near the top.

Preliminary analysis of the data in Figure 4 confirms the difficulties associated with a reified, abstract notion of *relevance* reported earlier in [2] and [1]. Most prominent is the variability across the relevance dimensions: very high numbers for aspects such as *Appeal*, *Concepts, Background* and *Examples*, and low numbers for *Attachments* and *References*. The second prominent finding is the expected poor performance of the *Overall relevance* inter-rater reliability relative to some of the more precise aspects.

The binary inter-rater reliabilities we observed (IRR-1, 64-95%) were higher than we expected. While '*Overall relevance'* is weak compared to most of the others, it is still quite high compared to the results reported in [2] and [1]. Although additional analysis is

Table 1. Relevance aspects, Saracevic's relevance clues [14], and their operationalizations.

Relevance clue	Relevance aspect	Statement
Affective Match	R-1 Appeal	The document contains materials that are motivational or stimulating (interesting, appealing, or engaging) for students.
Content	R-2 Concepts	The document includes concepts, keywords, terms, and definitions from the standard.
Content	R-3 Background	The document provides interesting and important background material related to the standard.
Object	R-4 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	R-5 Non-textuals	I can use a non-textual component(s) *e.g.*, figures, tables, images, videos or graphics, *etc*.
Situational Match	R-6 Examples	I can use the real-world examples provided in the document in class.
Situational Match	R-7 Hands-on	I can use one or more of the hands-on, active engineering activities.
Situational Match	R-8 Attachments	I can use some of the attachments; *e.g.*, score sheets, rubrics, test questions, *etc.*
Situational Match	R-9 References	I can use references or Internet links to relevant materials elsewhere.
	R-10 Overall relevance	Overall, I consider this document relevant for this teaching assignment.
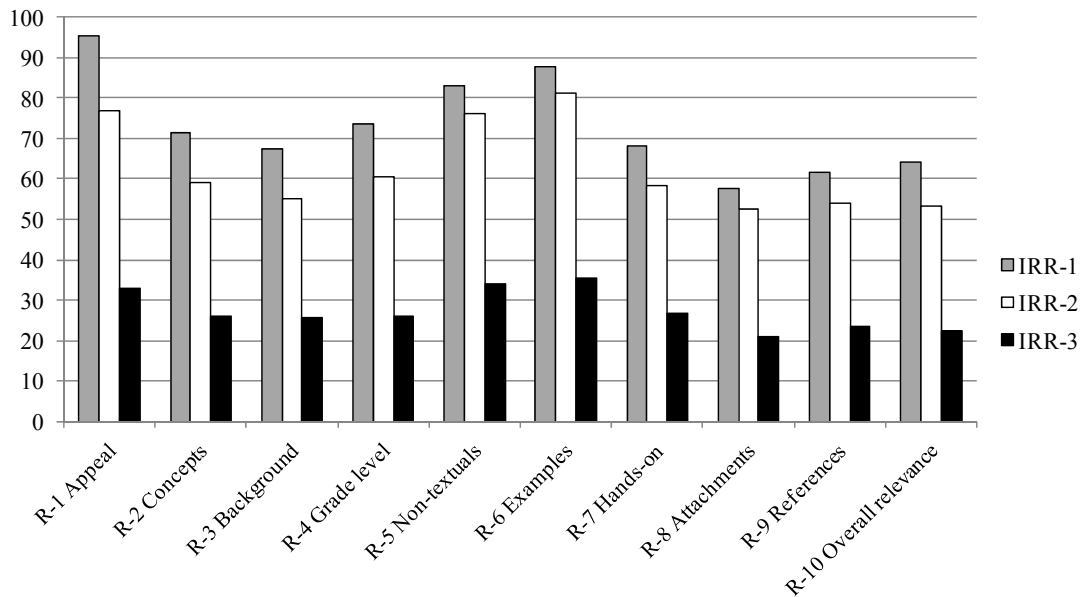
Figure 4. Inter-rater reliability for the 10 relevance measures.

needed to fully explain this pattern, we surmise that this is caused by several factors:

- The relevance question was asked in the context of an actual teaching task.

- The overall relevance question was only part of the evaluation done. Asking more detailed questions first may have affected responses to the overall relevance question.

- We used a six-point rather than a binary scale. It may be that this approach normalized the 'cutoff' point for relevance. If a binary rating was used, some judges may have decided, for example, that "somewhat agree" would not be good enough to equate with a good alignment in a standard/document pair.

- We asked users to score the relevance of a given set of document/standard pairs rather than to derive a set of alignments from a list of documents and standards.

  - Similarly, the inter-rater reliability scores on some of the relevance aspects are quite high. Although some of this might be the effect of all subjects being familiar and having affinity with the *TeachEngineering* system, we expect that the real reason for these high scores is the disambiguation of the relevance concepts; once it becomes clear which aspect of relevance we have in mind, the inter-rater reliabilities go up.

4.2 What About 'Overall Relevance'?

If these results indicate that overall relevance is neither a meaningful nor a very useful construct in measuring inter-rater reliability in curriculum-standard alignments, one can still ask if it cannot be considered as a complex, latent variable; *i.e.*, having a variable score based on a weighted combination of some of the more detailed relevance aspects.

Table 2 shows the correlations between all 10 relevance measures. Notice the absence of strong correlations between almost all of the specific relevance aspects (R-1–R-9), indicating that each measures an independent aspect of relevance.

Table 2. Correlations (r) between relevance aspects. See Table 1 for item descriptions.

	R-1	R-2	R-3	R-4	R-5	R-6	R-7	R-8	R-9
R-2	.16								
R-3	.36	.63							
R-4	.27	.20	.31						
R-5	.09	.19	.20	.18					
R-6	.21	.26	.27	.30	.42				
R-7	.34	.31	.40	.59	.24	.40			
R-8	-.02	.17	.21	.36	.35	.42	.39		
R-9	.06	.23	.20	.30	.25	.24	.45	.30	
R-10	.31	.76	.66	.50	.28	.35	.54	.35	.34

The statistical independence of the specific relevance aspects (R-1–R-9) implies that each of them can safely be used—no multicollinearity—as an independent variable in a multiple linear regression with overall relevance (R-10) as the dependent variable. Table 3 shows the results of the most parsimonious, best fit model: four independent variables (R-2, R-3, R-4 and R-7), all statistically significant at the p=.05 level.

Table 3. Multiple regression results.

	β	Std. Error	t-value	p
Intercept	-.272	.259	-1.048	.298
R-2 Concepts	.567	.083	6.794	< .01
R-3 Background	.173	.084	2.047	.044
R-4 Grade level	.322	.090	3.576	<.01
R-7 Hands-on	.194	.093	2.082	.041

The results show a good fit (R^2 = .75) with all variables statistically significant and all signs (β's) in the expected direction. From these results we conclude the following:

- For a sufficiently precise K-12 science or mathematics teaching task, 'overall relevance' of curriculum given a standard can be meaningfully and consistently constructed from the *Concepts*, *Background*, *Grade level* and *Hands-on* dimensions.

- Some of the other dimensions that yielded high inter-rater reliability; *e.g.*, *Appeal*, *Non-textuals* and *Examples,* do not contribute to overall relevance.

- Despite them not factoring into overall relevance, subjects strongly agreed about whether they could use non-textual components and examples from the curricular documents to teach the standard (μ(Non-textuals) = 2.1; μ(Examples) = 1.9).

5. REPERCUSSIONS

We propose that the above results have implications for the formation of a gold standard of curriculum-standard alignments as well as for the further development of alignment mechanisms.

5.1 Repercussions for Gold Standard Formation

If the above results can be replicated in larger groups of subjects, and a higher variation of curriculum-standard alignments, a gold standard for these alignments against which alignment mechanisms can be tested may be quite attainable. If indeed the poor inter-rater reliabilities reported in the literature are to a large extent an artifact of the operationalization of the context of relevance, quality or goodness-of-fit, more precise operationalizations that closely connect to the actual task for which the alignments are needed may very well strongly increase inter-rater reliabilities.

Furthermore, under such task-specific conditions, even the vague and abstract notion of 'relevance' may have a meaningful, although not necessarily very useful, interpretation as a complex, latent variable whose value can be consistently computed from its more precise dimensional scores.

5.2 Repercussions for Curriculum-Standard Alignment Mechanisms

If the above results indeed hold, they have important repercussions for the development of and overall approach toward curriculum-standard alignment. For instance, the data in this experiment indicate that very high levels of inter-rater reliability can be reached on different, uncorrelated relevance dimensions. This would mean that the documents in a digital educational library may align with a certain standard from one perspective, but not from another. Yet each of these perspectives is a highly agreed upon aspect of relevance for teaching science and mathematics on the K-12 level! This suggests a flexible approach to alignment. An approach, for instance, through which certain aspects and characteristics of a document or learning object can be included in certain cases, but not in others. This then creates the challenge of developing curriculum-standard alignment frameworks which are multivalued in nature; *i.e.*, they can flexibly involve 'evidence' of alignment from multiple sources and entirely different data types. Whereas the methods expressed by tools such as SAT and CAT rely largely on linguistic-semantic information, flexible tools should be able to combine this information with other information: How many examples does the document contain? Is the document linked to other documents which themselves have examples or hands-on activities? If a document is referred to by a really attractive and aligned document, does that mean the referred document also aligns?

6. CONCLUSION

A significant amount of good work has been done to facilitate the building of educational digital libraries capable of connecting educational standards to individual learning objects. The need for and value of this work increases as the supply of readily available digital learning objects increases. Collections like TeachEngineering hope to provide their users with a simple way to accurately locate useful resources for assigned standards, but development of improved tools depends on our ability to identify good matches in our collection for existing standards. We believe this study usefully explores how researchers can go about obtaining more consistent (and hopefully more useful) alignments to train, compare, and guide standards-based search facilities.

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

[3] Diekema, A.R., Chen, J. (2005) Experimenting with the Automatic Assignment of Educational Standards to Digital Library Content. In: Proceedings of the 5th ASM/IEEE Joint Conference on Digital Libraries. Denver, Colorado, June. The Association of Computing Machinery. New York. NY.

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

[5] Greenwood, J.D. (1982) On the Relation Between Laboratory Experiments and Social Behavior: Causal Explanation and Generalization. Journal for the Theory of Social Behavior. 12. 225-250.

[6] Lagoze, C., Payette, S., Shin, E., Wilper, C. (2005) Fedora: An Architecture for Complex Objects and their Relationships. Journal of Digital Libraries, Special Issue on Complex Objects. 6. 124-138.

[7] Lagoze, C., Kraft, D.B., Payette, S., Jesuroga, S. (2005) What Is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL. D-Lib Magazine. 11.

[8] Lazarsfeld, P.F. (1954) A Conceptual Introduction to Latent Structure Analysis. In: Lazarsfeld, P.F. (Ed.) Mathematical Thinking in the Social Sciences. Free Press. New York. NY.

[9] Levins, R., Lewontin, R.C. (1980) Dialectics and Reductionalism in Ecology. Synthese. 43. 47-78.

[10] Marshall, B., Reitsma, R., Cyr, M., Standards or Semantics for Curriculum Search? Proceedings of the 7th Joint ACM/IEEE Conference on Digital Libraries (JCDL 2007), June 17-22, 2007, Vancouver, British Columbia, Canada.

[11] Noble, W.G. (1981) Gibsonian Theory and the Pragmatist Perspective. Journal for the Theory of Social Behavior. 11. 65-85.

[12] Saracevic, T. (1975) Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Journal of the American Society of Information Science. 26(6). 321-343

[13] Saracevic, T. (2007) Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II: Nature and Manifestations of Relevance. Journal of the American Society of Information Science and Technology. 58. 1915-1933.

[14] 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.

[15] Van der Smagt (1985) Definieren en Relateren in Sociaal Wetenschappelijk Onderzoek. (Definitions and Relations in Social Science Research). Dissertation. University of Nijmegen, The Netherlands.