An Application of Text Embeddings to Support Alignment of Educational Content Standards

Reese Butterfuss Harold Doran

Human Resources Research Organization

Paper Presented at Generative Artificial Intelligence for Measurement and Education Meeting

February 1, 2024

Abstract

Large language models are increasingly used in educational and psychological measurement activities. Their rapidly evolving sophistication and ability to detect language semantics make them viable tools to supplement subject matter experts and their reviews of large amounts of text statements, such as educational content standards. This paper presents an application of text embeddings to find relationships between different sets of educational content standards in a process termed content mapping. This content mapping process is routinely used by state education agencies and is often a requirement of the United States Department of Education peer review process. We discuss the educational measurement problem, propose a formal methodology, demonstrate an application of our proposed approach, and provide measures of its accuracy and potential to support real-world activities.

Keywords: content alignment, content mapping, natural language processing, semantic similarity, text embeddings

Introduction

Alignment is critical to educational and psychological measurement and is one component of the validity investigation process regarding an assessment system. It is a broad term and has implications for work taking many forms with variants including alignment of test items to standards (Webb, 1997), aligning two distinct sets of standards (Neidorf et al., 2016), or establishing relationships between performance level descriptors (PLDs) and test items (Schneider, Agrimson, & Veazey, 2022).

A common theme across these alignment approaches is that they involve subject matter experts (SMEs) in a time-consuming process in which they review large amounts of text data and/or test items in some form. In general, alignment activities require expert judges to review test content (e.g., items, standards, PLDs) and use their expert judgment to determine the degree to which there is similarity between the test content under review. Specifically, experts may evaluate a test question against a learning objective and assess whether it is measuring that standard, or experts may evaluate two sets of learning standards and find pairs of text statements that form matches between their intended learning objectives.

Content standards are formal definitions of what a student is expected to know and be able to do in a given subject area and grade level. These standards are often locally defined (e.g., by a state education agency (SEA)) and are used for instruction and assessment design in that jurisdiction. However, other sets of standards also exist and it is often useful to show how one set of standards formally relates to a different set of standards. This is generally necessary given that different sets of standards often express similar learning outcomes for students but differ in the language they use to express these outcomes. In some instances, demonstrating this relationship is required, such as addressing critical element 3.1 in the United States Department of Education (USDOE) standards and assessment peer review process (USDOE, 2018).

Establishing this relationship between different sets of standards is referred to as standards-to-standards *content mapping*, and the aim is to find if a standard in one set can be matched with a comparable learning standard in the second set. Some examples of content mapping include the Dynamic Learning Maps (DLMs) crosswalk to the Common Core State Standards (CCSS) (Dynamic Learning Maps Consortium, 2013), mapping the National Assessment of Educational Progress (NAEP) to the CCSS (Daro, Hughes, & Stancavage, 2015), mapping NAEP to state standards via the HumRRO approach (Vockley, 2009), mapping SAT to CCSS (Jackson, 2015), or mapping the World-Class Instructional Design and Assessment (WIDA) English Language Development Standards (ELD) to individual state standards (WIDA, 2020).

Content mapping is a difficult and time-consuming process that requires SMEs to evaluate many pairs of text statements and use their expert judgment to determine similarity. The largest challenge is that SMEs must often consider all possible pairs of text, which entails an overwhelming level of effort. For example, Figure (1) illustrates how large such tasks can be. In this example, there are 45 state standards and 38 NAEP standards. Content mapping in this scenario could involve SME evaluation of all $45 \times 38 = 1{,}710$ pairs of text statements to determine which standards are potentially aligned. In some cases, SMEs might only review organized subsets of standards presumably measuring similar content (e.g., limit within number sense dimension), but there can be large pairwise comparisons needed even within these subsets.

In practical educational settings, content mapping often involves teachers or other content experts working long hours in addition to their teaching activities to complete these mapping activities. Additionally, the sheer number of pairwise text statements reviewed represents an enormous amount of work, often resulting in cognitive overload and fatigue, both of which are root causes of errors. The challenge is seemingly obvious—is it possible to find ways to ease the task such that SMEs review smaller subsets of text pairs that contain the most probable matching pairs instead of reviewing overwhelming amounts of text? This question is what motivates our exploration using natural language processing (NLP).

Some prior work has explored the potential of NLP to content mapping. Zhou and Ostrow (2022) evaluated the accuracy with which transformer models trained from scratch can predict

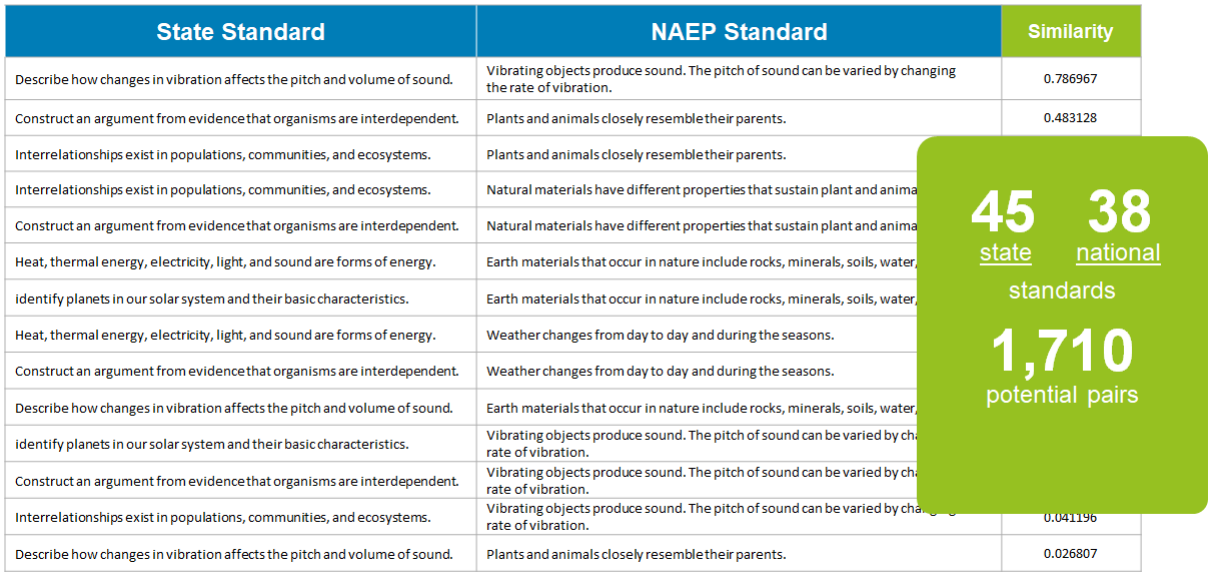


Figure 1: *Total Pairwise Comparisons*

which of 822 unique standards each item in a large corpus of 9,836 unique grade 3 – 8 English language arts items was aligned. A subset of their standards were CCSS standards. Their results showed that the predicted standard matched the true standard with 35% accuracy on the training set and 23% accuracy on the test set, but for the CCSS subset, the accuracy on the training set was 60% and accuracy on the test set was 29%. The authors attributed the steep drop in accuracy for the CCSS test set to overfitting due to small training sample sizes. To represent the meaning of the test items, their method leveraged word-level embeddings that were averaged over the words within each assessment item. An alternative approach is to derive sentence embeddings to represent the meaning of text, which better captures context and therefore deeper, more nuanced meaning of text input compared to word-level embeddings (Arora, Liang, & Ma, 2017; Reimers & Gurevych, 2019). Khan, Rosaler, Hamer, and Almeida (2021) developed Catalog, a content-tagging system that uses semantic similarity derived from a transformer-based semantic matching algorithm to match educational materials (e.g., reading passages) to 24 unique Next Generation Science Standards (NGSS). Specifically, Catalog combines each text element from the educational materials with each standard, and processes the combined text using a transformer model pre-trained for next token prediction to derive probability values for each text element from the educational materials conditioned on the NGSS standards. Then, the educational material is processed in isolation via the same method to derive unconditional probabilities, and finally, the unconditioned and conditioned probabilities are compared to produce a single score by which to rank-order NGSS standards for each text element in the educational standards. Results from Catalog suggest that it performed either similarly or superior to human judgment in accurately matching high school biology textbook passages to NGSS standards.

The existing recent NLP-based approaches in the extant literature show promise, but there are shortcomings and barriers to practical use that could be addressed in subsequent work. Specifically, the approach developed by Zhou and Ostrow (2022) requires sufficiently large corpora to train the transformer models, does not leverage sentence embeddings, and struggles with instances when

a single item aligns with multiple standards. The approach developed by Khan et al. (2021) indeed leverages sentence embeddings and does not require training data, but it is designed to facilitate the alignment of educational content with only NGSS standards. Moreover, the goals of these approaches focus more on fully *automating* the alignment process rather than on providing a means of increasing the efficiency and effectiveness of SMEs during the alignment process.

The latter point is critical and is a major objective for our work. Our view is that NLP-based applications for alignment should only serve as a supporting mechanism to create an organized pathway for SMEs to decide which standards are aligned. Human expert judgment is still at the core of our process and alignment decisions are not automated by a text classification approach.

Purpose and Organization

This paper presents a method that can assist SMEs by leveraging transformer models, text embeddings, and the cosine similarity index to improve the efficiency of the content mapping process. Specifically, we developed an intuitive method that can potentially apply to the alignment of any content standards and requires no text preprocessing or labeled training datasets. The practical goal of the method is not to automate content alignment outright but rather to facilitate the alignment task by decreasing the workload imposed on SMEs. To this end, we sought to constrain the search space SMEs must negotiate during alignment by limiting the number of pairs to only those that share high semantic overlap and are therefore more likely candidates for alignment.

Hence, our approach is *NLP-assisted* and not *NLP-driven*. That is, an NLP-driven method is one where the approach on its own forms a classification between matched pairs, whereas an NLP-assisted method is used to indicate that the language model only helps inform the SMEs about which pairs seem most likely to correspond, and then SMEs review the subset and make final judgments. This approach makes use of NLP techniques as a supporting mechanism to initially form probable pairs and reduce the total number of text pairs experts would need to review. The method itself does not establish alignment or make any decisions.

The remainder of the paper is organized into three sections. First, we describe the concept of text embeddings. Second, we formalize a methodology that uses these embeddings as a supporting mechanism to facilitate content mapping. Third, we demonstrate this method using real NAEP data along with its accuracy metrics to demonstrate its potential in real-world settings.

Sentence Embeddings as a Supporting Tool

A *sentence embedding* is a transformation of a text statement into a numeric representation of the text's meaning. These embeddings are obtained via large language models (LLMs) that convert text to embeddings and also capture the semantic meaning inherent in the text itself. Embeddings are not limited to single sentences but can more broadly convert text statements longer than a single sentence into a numeric representation and so for this reason we use the term *text embedding* to apply the concept of our work more broadly. This differs from other approaches that only convert individual words to embeddings, such as GLOVE and Word2Vec (Zhou & Ostrow, 2022), and other historical text transformation approaches that are designed only to evaluate whether text strings are similar enough to be considered the same piece of text, not whether they are capturing similar meaning such as a Levenshtein distance (Doran & Van Wamelen, 2010).

Importantly, the numeric transformations of each text statement now allow for empirical methods to establish relationships between large numbers of text statements. Generating embeddings for each standard transforms each standard into a numeric vector representation, and each has a location in a high-dimensional space. Because each standard has a location in this vector space, it is possible to assess the relation between any two standards. For example, assume we have two text statements that have each been transformed into a numeric embedding. We can then establish a statistical relationship between them with a common statistic used in machine learning for relating pairs of text embeddings called the cosine similarity index, which is akin to a Pearson correlation when the means of the vectors are centered on zero. It is a distance metric that effectively determines how similar two different pairs of text are in terms of their semantic similarity and has the same range and interpretation as the Pearson correlation.

When this work is performed in large batches (i.e., converting large amounts of text into embeddings), then perhaps we can also more efficiently compare large numbers of text pairs and find relationships between them. This could potentially preserve large amounts of human effort by forming subsets of the most probable matching pairs, thereby allowing SMEs to evaluate a smaller number of possible matching pairs instead of evaluating all possible pairs.

Outlining an NLP-Based Framework for Content Mapping

We can begin to imagine a new way that content mapping with large batches of text might unfold with the support of text embeddings. Assume there are two (or more) sets of content standards to be aligned. Informally, an outline and high-level sketch of a new, simplified process can be framed as:

1. Compute text embeddings between all pairs of text statements in both sets.
2. Compute cosine similarity between all pairs of embeddings.
3. Sort each statement with its corresponding pairs in descending order by cosine similarity.
4. Ask SMEs to find the matching pair in a subset of the ordering from (3) instead of searching over the entire set of text statements.

The general idea is that the cosine similarity index can be used to find which pairs of text have the greatest likelihood of being matched. Then, by rank ordering the text statements by their cosine similarity, SMEs will review a smaller subset of text statements where, hopefully, the true match will be somewhere in the top few text statements. Viewed from this perspective, our method orders content in a manner that resembles the Bookmarking standard setting method where SMEs review test items ordered by their response probabilities (RP) (Mitzel, Lewis, Patz, & Green, 2001). In our approach, we order content standard pairs by their cosine similarity in the same way that test items are ordered by RP values, and SMEs would review this ordered set to locate the matched pair. This is just a general sketch of the process, and the next section formalizes this implementation. Readers less interested in the technical approach can skip the formalities.

Formalizing the Method

Let the collection of text statements be represented as two distinct sets, $\mathcal{A} = \{1, 2, j, \ldots, J\}$ and $\mathcal{B} = \{1, 2, k, \ldots, K\}$. Here, the elements within each set, denoted $\mathcal{A}_j \in \mathcal{A}$ and $\mathcal{B}_k \in \mathcal{B}$, are individual text statements, long or short. The aim is to find similar pairs of text, $\mathcal{A}_j \simeq \mathcal{B}_k$, where this notation indicates these two elements are measuring a similar learning objective and would be considered a matched pair. Another way to state this is $\mathcal{A}_j \cap \mathcal{B}$ and is used to mean that text statement $\mathcal{A}_j$ is contained within $\mathcal{B}$.

In a typical alignment workshop, SMEs would take text statement $\mathcal{A}_j$ and compare it to all elements contained in $\mathcal{B}$. Making all of these pairwise comparisons is a large effort involving a total of $J \times K$ total comparisons when comparing all elements in $\mathcal{A}$ to all elements in $\mathcal{B}$. The formal objective is to reduce the dimension of $\mathcal{B}$ finding the subset such that $\mathcal{B}_p \subseteq \mathcal{B}$ so that SMEs have a smaller space to navigate making this task more reasonable. Then, SMEs search to determine $\mathcal{A}_j \cap \mathcal{B}_p$ where $\mathcal{B}_p$ is the subset containing the most probable match to $\mathcal{A}_j$. Note that we intentionally do not use a classification approach where $\mathcal{A}_j \simeq \mathcal{B}_k$ is identified for the SMEs. We believe doing so could lead to confirmation bias and would remove the SMEs expert judgment.

Let the jth embedding in $\mathcal{A}$ be denoted as $\boldsymbol{\alpha}_{j,\mathcal{A}} \in \mathbb{R}^N$ and the kth embedding in $\mathcal{B}$ be denoted as $\boldsymbol{\alpha}_{k,\mathcal{B}} \in \mathbb{R}^N$. Each embedding, $\boldsymbol{\alpha}$, is a real-valued vector of length N both spanning the same plane. Consequently, the angle between them is a metric describing the degree to which they are related. For this reason, we can compute for each embedding pair, $\mathcal{L}_{i(\mathcal{A},\mathcal{B})} = \psi(\boldsymbol{\alpha}_{j,\mathcal{A}}, \boldsymbol{\alpha}_{k,\mathcal{B}}) \; \forall \; j, k$ where $\psi(\cdot)$ is the cosine similarity metric.

Let $\boldsymbol{\mathcal{L}}_{j,k:K} = (\mathcal{L}_{1(\mathcal{A},\mathcal{B})}, \mathcal{L}_{2(\mathcal{A},\mathcal{B})}, \mathcal{L}_{p(\mathcal{A},\mathcal{B})}, \ldots, \mathcal{L}_{K(\mathcal{A},\mathcal{B})})$ be the vector of all cosine similarities between $\mathcal{A}_j$ and every other element in $\mathcal{B}$. This notation, $\boldsymbol{\mathcal{L}}_{j,k:K}$, is used to mean the cosine similarity between $\boldsymbol{\alpha}_{j,\mathcal{A}}$ and all elements in $\mathcal{B}$, hence it is always of length K and there is one of these vectors for each element in $\mathcal{A}$, or $(\boldsymbol{\mathcal{L}}_{1,k:K}, \boldsymbol{\mathcal{L}}_{2,k:K}, \ldots, \boldsymbol{\mathcal{L}}_{J,k:K})$.

Once these values are obtained, then for each $\mathcal{A}_j$ sort its corresponding cosine vector, $\boldsymbol{\mathcal{L}}_{j,k:K}$, in descending order. The first element of this sorted vector has the highest cosine similarity index, suggesting it has the highest degree of semantic similarity and the last element has the smallest cosine similarity. If the cosine similarity index is useful at capturing the similarity between content standards, then we might expect the first element, or approximately so, would be the aligned or matching standard pair. In this case, we could ask SMEs to review only a subset of the most probable matching text pairs, $\boldsymbol{\mathcal{L}}_{j,k:K} = (\mathcal{L}_{1(\mathcal{A},\mathcal{B})}, \mathcal{L}_{2(\mathcal{A},\mathcal{B})}, \ldots, \mathcal{L}_{p(\mathcal{A},\mathcal{B})})$, letting p be some positional element in the vector that is less than K. When, $p \ll K$ the search space is substantially reduced and human effort is minimized.

The determination of p is the fundamental task so that we can form the subset $\mathcal{B}_p$. If all standards in set $\mathcal{B}$ were randomly sorted, then the true match, $\mathcal{A}_j \simeq \mathcal{B}_k$, is uniformly distributed over the range of K and any choice $p \in \{1, 2, \ldots, K\}$ has an equal probability of containing the true match. Then, the uniform cumulative distribution function provides that the true match randomly appearing in the top p sorted cases would be $\Pr(\mathcal{A}_j \cap \mathcal{B}_p) = \frac{p}{K}$. Of course, we intentionally sort by the cosine similarity. However, this establishes the null distribution for how often the true match would be observed in the top p due to random chance alone and we can compare our observed results to this null probability to assess how much more efficient our proposed approach is relative to what is expected under random chance alone.

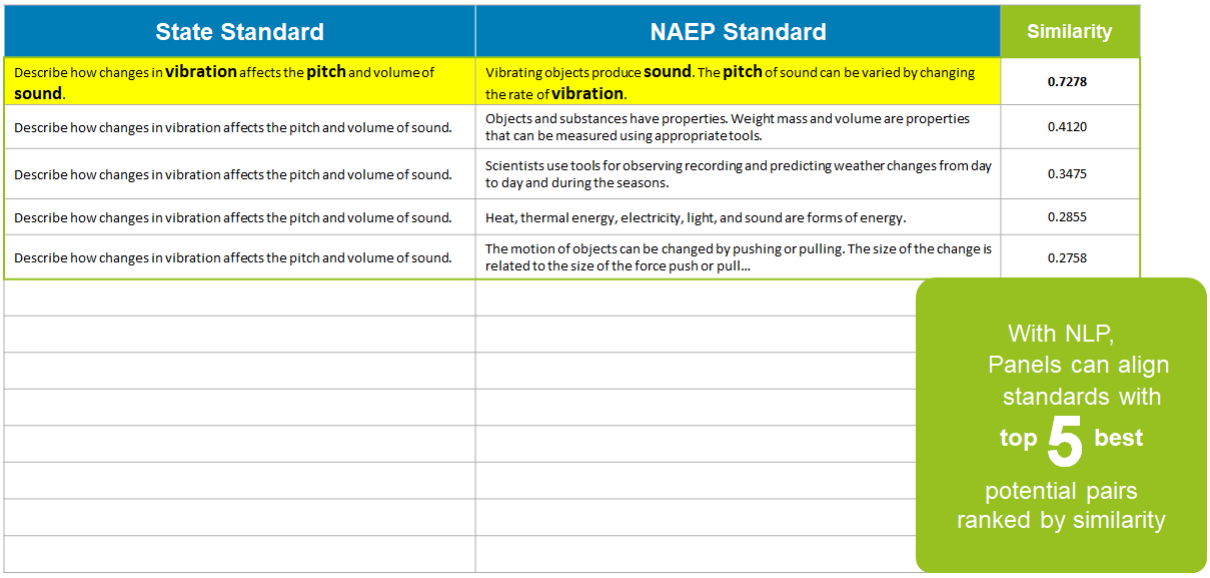


Figure 2: *Reduced Set of Comparisons*

Example

Figure (2) provides one example of this approach. In this example, the element $\mathcal{A}_1$ = "*Describe how changes in vibration affects the pitch and volume of sound.*" is compared to all text statements in the NAEP set, or $\mathcal{B}$. The column "Similarity" is the vector of cosine similarities between this standard and all other NAEP standards, or $\boldsymbol{\mathcal{L}}_{1,k:K}$ to correspond to the notation in the method section. The values here are sorted according to cosine similarity in descending order such that the pairs sharing the highest degree of similarity are presented first. We observe that $\mathcal{A}_1$ has a cosine similarity of .73 with the statement "*Vibrating objects produce sound. The pitch of sound can be varied by changing the rate of vibration.*" and a cosine similarity of .29 with the statement, "*Heat, thermal energy, electricity, light, and sound are forms of energy.*"

The idea is that the pairs with the largest cosine similarity are the most probable matches. But the question is how many of those pairs need to be in the subset for experts to review? Should we show SMEs text pairs with cosine similarities above a specific value or should humans review the top p ordered pairs of text statements? We further explore this in the analysis section using our labeled data.

Data

Our work makes use of a labeled data set where groups of three SMEs–one for each grade level 4, 8, and 12–reviewed individual standards from 30 states and aligned each state standard to the corresponding NAEP standard. The number of standards represented within each state varied across Grade 4 (22 states included $<$30 standards, eight included 30+), Grade 8 (11 states included $<$30 standards, 19 included 30+), and Grade 12 (13 states included $<$30 standards, 17 included 30+). In the original work, SMEs labeled the data as 0 = "not aligned" (n_4 = 60; n_8 = 103; n_{12} = 209), 1 = "partially aligned" (n_4 = 243; n_8 = 520; n_{12} = 739), or 2 = "fully aligned" (n_4 = 321; n_8

= 545; n_{12} = 97). In the original study, items that were judged to be either partially aligned or fully aligned were simply combined into an "aligned" category. Across grades, the standards covered the following content domains: Earth and Space Science (n = 1483 standards), Life Science (n = 1717 standards), and Physical Science (n = 2108 standards). Because we have labeled data, we treat the human-assigned match as the "true match." Table (1) provides an overview of the data used for this work.

Grade	No. of States	No. State Standard	No. NAEP Standard	No. Potential Pairs
4	30	768	33	25,344
8	30	1,832	43	78,776
12	30	2,247	49	110,103

Table 1: Labeled Data

Analysis

We began by computing text embeddings for all pairs of text in our data using four different popular open-source language models including all-mpnet-base-v2, all-distilroberta-v1, all-MiniLM-L6-v2, and multi-qa-MiniLM-L6-cos-v1, all of which are models in the Hugging Face ecosystem (Wolf et al., 2019). These models were selected because, at the time of this writing, those particular models were showing high predictive abilities and at the top of the Hugging Face "leaderboard". We did initially experiment with others, including Vicuna and T5 variants, but our work with those was computationally expensive, and our results did not suggest they offered any substantial benefit over other LLMs used. Once the embeddings were derived, we computed pairwise cosine similarities between all pairs as in $\boldsymbol{\mathcal{L}}_{j,k:K}\ \forall j$.

Importantly, we did not "fine-tune" any of these LLMs using our labeled data, we use them "out-of-the-box" and used the labels to determine how accurate our proposed method is to support the human reviews. Specifically, once the text statements were sorted in descending order using the cosine similarity, we found the proportion of state standards for which that the true match NAEP standard as determined by SMEs was among the top p ordered pairs.

To illustrate, suppose we had 10 text statements in $\mathcal{A}$ and 50 statements in $\mathcal{B}$. Assume that in eight cases, the true match is within the top five sorted pairs and the other two cases are within the top 10. We would say that 80% of the time, the true match is in the top five and 100% of the time it is within the top 10. The implication is that the text processing as described in the method section could be carried out before a content mapping workshop and SMEs could be asked to review only the top five text statements instead of reviewing all 50 statements for each pair. This would reduce the level of effort from reviewing as many as $5 \times 50 = 250$ pairs of text down to only $5 \times 5 = 25$ pairs of text.

Of course, in two of those pairs, the true match might fall outside the top five. Our proposed approach would provide experts with all text pairs ordered by their cosine similarity and ask them to review the top five to find a match. If that match is not within the top five, SMEs keep reviewing potential pairs as-needed.

The proposed approach may be help address additional complexities that can arise during content alignment. First, one cannot assume that each standard in one set indeed has a corresponding standard in the other set. There may be cases for which the SMEs determine that a given state standard is unique and therefore has no matching or parallel NAEP standard, for example. In our labeled data, 13.1% of state standards had no corresponding NAEP standard. Our method may provide a convenient way for SMEs to more quickly conclude that there is no equivalent standard because they would be reviewing the most likely rank-ordered candidates first. Second, one cannot assume that each standard in one set aligns with only one standard in the other set. In our labeled data, 7.7% of state standards were aligned with two or more NAEP standards. Again, our proposed method would likely prove helpful because the standards with which the target standard is aligned would be among the most similar pairs. Lastly, in a similar vein, different bodies of standards are often not written to the same level of specificity. For a simple example, a mathematics standard may be phrased "Student can perform basic arithmetic operations (adding, subtracting, multiplying, dividing) with whole numbers." The comparison set of standards may include four separate standards that capture the same idea, one each for adding, subtracting, multiplying, dividing. This asymmetry can occur in both directions between the two bodies of standards. Due to the nature of cosine similarity, each of the four separate standards in the comparison that capture the same idea as the single standard would be among the top-ranked pairs and would therefore permit SMEs to determine alignment quickly and easily.

Results

Table (2) provides descriptive statistics for cosine similarity values from each language model for each grade level, including overall (i.e., all possible pairs), only those standard pairs that SMEs judged as the "true matches," and only state standards that SMEs determined did not have a corresponding NAEP standard. Likewise, Figure (3 provides distribution density plots for the overall cosine similarity values and the true matches for each grade level and language model.

We describe the results in three subsections. First, for each grade level (i.e., grades 4, 8, and 12), we calculated an overall accuracy metric that captures the probability that the true match appears among the top p highest-cosine pairs, aggregating over the 33 states. Doing so directly informs the application of the method, as it provides information about the rank-order value of the true matches among all possible matches, which in turn indicates the minimum set size of high-cosine standard pairs that SMEs would need to consider. This information is critical to our goal of reducing the search space for human raters by reducing the number of viable pairs they must consider when making judgments about alignment.

Second, we calculated accuracy for SMEs judgments within four possible conditions of SME consistency and degree of alignment: (a) SMEs *unanimously* judged standards to be *fully* aligned, (b) SMEs were originally *split* in their judgments but reached a consensus that standards were *fully* aligned, (c) SMEs *unanimously* judged standards to be *partially* aligned, and (d) SMEs were originally *split* in their judgments but reached a consensus that standards were *partially* aligned. These results provide information about the extent to which cosine similarity is sensitive to SME judgment regarding whether items are partially vs. fully aligned.

Third, we calculated accuracy for separately for each state if that state included at least 30 standards in the original alignment study (n = 8 for grade 4, n = 16 for grade 8, n = 12 for grade

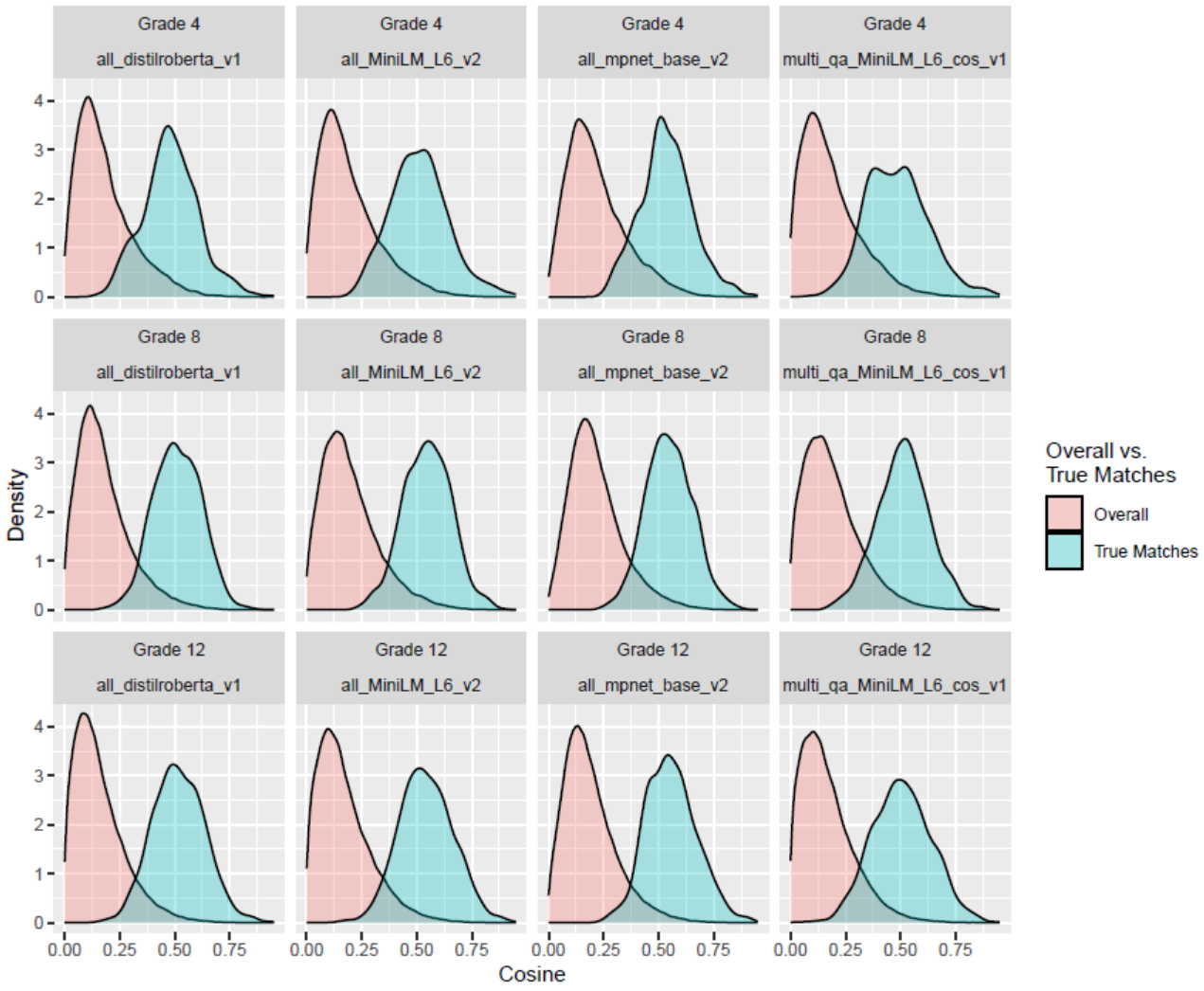


Figure 3: *Distribution Plots for each grade and language model*

Model	Grade 4	Grade 8	Grade 12
Overall - *M*(*SD*)			
all_MiniLM_L6_v2	.18(.14)	.19(.13)	.16(.13)
all_distilroberta_v1	.17(.13)	.17(.12)	.14(.12)
all_mpnet_base_v2	.21(.13)	.21(.12)	.18(.12)
multi_qa_MiniLM_L6_cos_v1	.16(.14)	.17(.13)	.14(.13)
True Matches Only - *M*(*SD*)			
all_MiniLM_L6_v2	.51(.13)	.55(.11)	.54(.12)
all_distilroberta_v1	.48(.12)	.51(.11)	.52(.12)
all_mpnet_base_v2	.54(.12)	.55(.11)	.56(.11)
multi_qa_MiniLM_L6_cos_v1	.49(.14)	.51(.10)	.51(.13)
Standards with No Matches - *M*(*SD*)			
all_MiniLM_L6_v2	.17(.13)	.18(.12)	.15(.13)
all_distilroberta_v1	.17(.12)	.16(.13)	.14(.12)
all_mpnet_base_v2	.21(.13)	.21(.13)	.18(.12)
multi_qa_MiniLM_L6_cos_v1	.15(.12)	.17(.10)	.14(.13)

Table 2: Descriptive Statistics

12) to examine the extent to which the accuracy results were consistent across states.

Overall Accuracy

For grade 4, there were 33 NAEP standards that were potential candidates for alignment with each of 770 state standards across the 33 states. Figure (4a) shows that the true match as designated by SMEs was the single highest-cosine pair 52%-56% of the time across the four language models. The true match appeared within the top two highest-cosine pairs between 71%-77% of the time. The true match appeared within the top five highest-cosine pairs between 89%-94% of the time. Thus, the NLP approach was approximately 94% accurate at capturing the true match within the top five highest-cosine pairs. The four language models performed similarly in terms of accuracy, although all-miniLM-L6-v2 and all-mpnet-base-b2 appeared to slightly outperform all-distilroberta-v1 by 3% and multi-qa-MiniLM-L6-cos-v1 by 5% accuracy. Additionally, the single worst case (i.e., the worst rank-order value of the true match) fell between the 15th and 21st ranked pairs across models. Referencing back to the methodology section, we can establish the probability that the true match would appear within the top five cases as 5/33 = .15, or a 15% chance. The observed probability of 94% is much larger suggesting this method yields results much better than expected by random chance alone.

For grade 8, there were 43 NAEP standards that were potential candidates for alignment with each of 1,833 state standards across the 33 states. Figure (4b) shows that the true match was the single highest-cosine pair 54%-58% of the time across the four language models. Similar to the grade 4 results, the NLP approach reached 93%-96% accuracy at capturing the true match within the top five highest-cosine pairs. Likewise, the four language models performed similarly in terms of accuracy, although multi-qa-MiniLM-L6-cos-v1 slightly underperformed compared to the other language models, showing 93% accuracy at capturing the true match within the top five

most similar pairs, whereas the other models showed 95-96% accuracy. The single worst case (i.e., the worst rank-order value of the true match) fell between the 16th and 23rd-ranked pairs across models. Here we expect a 5/43 = .12, or 12% chance that the true match would appear in the top five. Again, the observed rate of 95% is much higher than what is expected from random chance.

For grade 12, there were 49 NAEP standards that were potential candidates for alignment with each of 2,247 state standards across 33 states. Figure (4c) shows that the true match was the single highest-cosine pair 59%-64% of the time across the four language models. The NLP approach reached 93%-97% accuracy at capturing the true match within the top five highest-cosine pairs. As before, the four language models performed similarly in terms of accuracy, although multi-qa-MiniLM-L6-cos-v1 slightly underperformed. The single worst case (i.e., the worst rank-order value of the true match) fell between the 16th and 23rd ranked pairs across models. Here we expect a 5/49 = .10, or 10% chance that the true match would appear in the top five. Again, the observed rate of 93% is much higher than what is expected from random chance.

Relating these results to our original question, "how many pairs do SMEs need to see?" suggests setting $p \approx 5$ provides roughly a 95% probability that the true match is within the top five ordered pairs. Using Bookmarking as an analogy, this is similar to SMEs reviewing only the first five "pages" to locate their bookmark location.

Grade	Observed Result	Expected Under Random Chance
4	89%–94%	15%
8	93%–96%	12%
12	93%–97%	10%

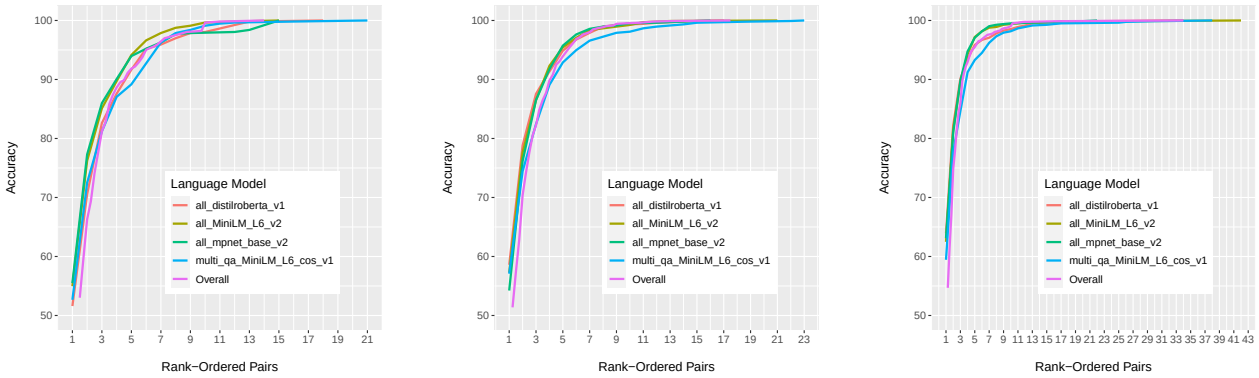
Table 3: Observed Percentage in Top 5

The results in Table (3) indicate a very large efficiency gain relative to the business as usual approach in content mapping. In these data, we observe there is roughly a 95% chance that the true match lives within the top five ordered statements whereas we would expect this to occur only about 10% to 15% of the time under random chance. Hence, to achieve this same 95% level of confidence with a traditional approach means SMEs would need to review 31, 41, and 47 text pairs in grades 4, 8, and 12, respectively instead of only the top five pairs.

Accuracy By SME Consistency and Degree of Alignment

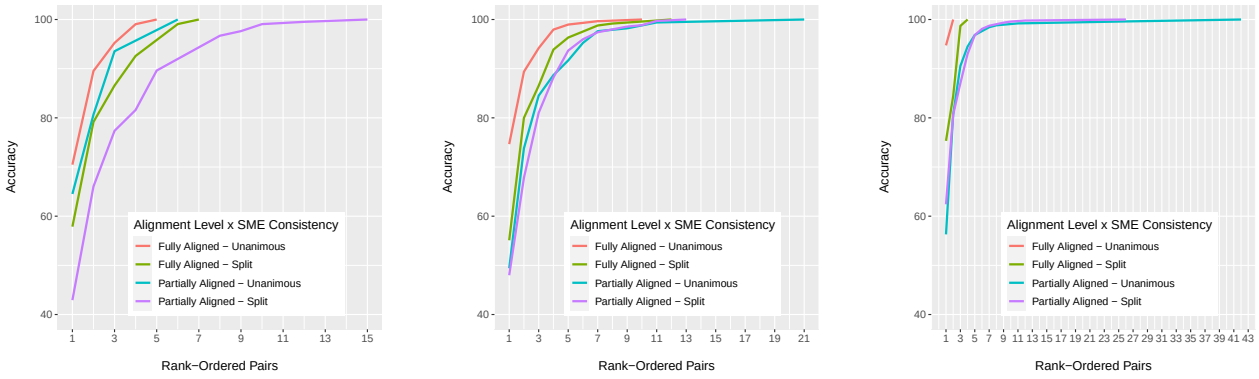
For grade 4, Figure (5a) shows that accuracy was highest for cases when SMEs unanimously judged that standards were fully aligned. This is a reassuring result showing that when SMEs tend to agree, the NLP-based approach also tends to produce better results. For such cases, the true match appeared within the top five pairs 100% of the time. Accuracy was lowest for cases when SMEs disagreed but reached a consensus that standards were only partially aligned. For such cases, the true match appeared within the top five pairs only 90% of the time. Accuracy for cases when SMEs were unanimous about partial alignment and for cases when SMEs were split about full alignment fell in between.

For grade 8, the results followed a similar pattern. Figure (5b) shows that accuracy was highest for cases when SMEs unanimously judged that standards were fully aligned, with the true match falling within the top five pairs 99% of the time. Accuracy was lowest for cases when SMEs



(a) *Grade 4* (b) *Grade 8* (c) *Grade 12*

Figure 4: Overall Accuracy Results



(a) *Grade 4* (b) *Grade 8* (c) *Grade 12*

Figure 5: Overall Accuracy Results

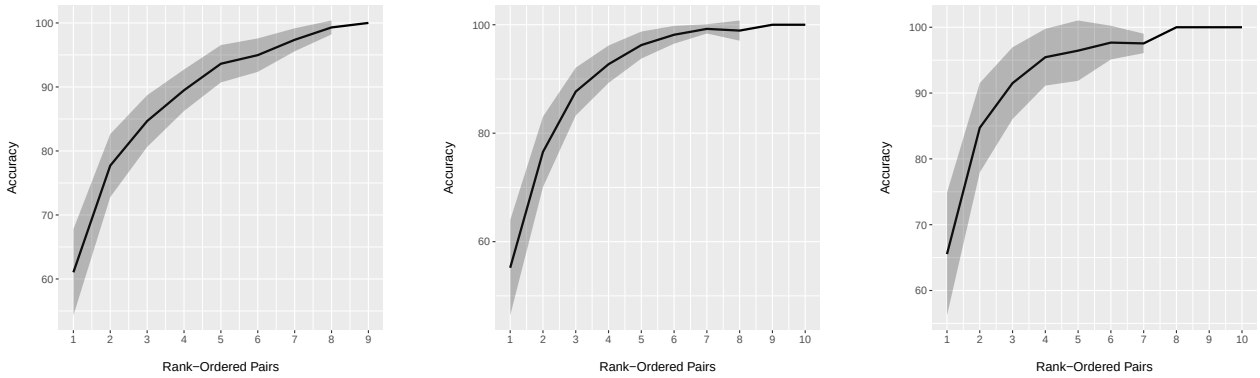
Note: Accuracy is based on rank_all_mpnet_base_v2.

disagreed but reached a consensus that standards were only partially aligned, with the true matches falling within the top five pairs 92% of the time).

Finally, for grade 12, Figure (5b) shows that accuracy was again highest for cases when SMEs unanimously judged that standards were fully aligned, with the true matches being the highest-cosine pair 95% of the time and falling within the top two pairs 100% of the time), whereas accuracy was lowest when SMEs judged standards only partially aligned, regardless of unanimity. However, for these cases, the true match still appeared among the top five pairs 96% of the time).

Accuracy Across States

For each plot in (6), the black line represents the mean accuracy for rank_all_mpnet_base_v2 across each state that consisted of at least 30 standards. We chose this language model just to illustrate consistency across states. The upper and lower bounds of the gray band represent the 95% confidence interval (CI). Because the 95% CIs were derived from aggregating data across states, they provide an estimate of the consistency of the accuracy results across states. For grade 4 (6a), the



(a) *Grade 4* (b) *Grade 8* (c) *Grade 12*

Figure 6: Accuracy Across States

Note: Accuracy is based on rank_all_mpnet_base_v2. The black line represents mean accuracy across states. The gray bands represent the 95% CI.

true matches fell within the top five highest-cosine pairs between approximately 92% and 97% of the time across states. For grade 8 (6b), the true matches fell within the top five pairs between approximately 95% and 99% of the time. Finally, for grade 12 (6c), the true matches fell within the top five pairs between approximately 92% and 100% of the time.

Discussion

Our goal was to develop an intuitive, generalized NLP-based approach to support content mapping activities and our proposed approach shows significant promise for real-world application. The approach we proposed offers a straightforward way of using text embeddings to improve the efficiency of SME efforts by substantially reducing the search space they must review to find matching pairs of text. Specifically, the approach we developed guides the alignment task by reducing the number of potential pairs that SMEs must consider to only those that are most likely to be aligned due to high semantic overlap. The results showed the subset of pairs provided to SMEs can generally be reduced to the top five pairs of standards ranked by cosine similarity, in which case our data showed there is a 95% probability that the true match exists within that subset.

The current results highlight a major benefit for practitioners, as the method may reduce the amount of time SMEs spend on these activities, also potentially reducing cost and improving accuracy when SMEs are less prone to cognitive overload and fatigue. We view two immediate applications for this work including content mapping activities needed to address the comparability requirements described in the USDOE peer review between sets of standards or second when SEAs need to establish linkages between the measurement objectives from off-the-shelf test forms (e.g., ACT, SAT) and local state standards. We also view the method as potentially supporting the embedded standard setting approach where items are reviewed and related to performance level descriptors (Lewis & Cook, 2020).

A second benefit of our approach is that the pre-trained language models all behave similarly out of the box for this type of activity. The correlations among the cosine similarity values pro-

duced by each of the four models ranged from r = .84 - r = .90). The four language models used in the current work were selected because of accessibility and popularity (Wolf et al., 2019), but many others could be used in future work. Indeed, future work will examine the extent to which different language models produce different results across content mapping data sets, but these initial results suggest that text embeddings may be somewhat invariant to the specific LLM used, and the complicated task of training a model to be used for this purpose may be unnecessary. Compared to related work in the existing literature, the current approach does not require labeled datasets, text preprocessing, or model tuning to be useful for facilitating alignment, at least in the current context.

A third benefit is our approach on its own does not form the classification. It is only an organizing framework to make a very large problem smaller and SMEs are still fundamentally at the center of the classification decisions. In this way, there is little risk of using this method for at least two reasons. First, if the SMEs do not find the "true match" within the top five, they can keep reviewing additional matches until they do find the match. Ordering the pairs only pre-organizes the data and does not limit SMEs in their work. Second, there is no risk of classification error by the approach itself. Any misclassification that occurs is based in the SMEs judgment, which may be probable, but that remains probable even without using our approach.

The degree to which the current results generalize to other types of content mapping activities (e.g., different standard domains, items-to-standards alignment) needs further review and exploration. Still, in the current study, the 95% probability of capturing the true match within five pairs generalized across grade levels and different states in our data. Additionally, future work will examine the embeddings of the standards to identify the linguistic signatures of the 5% of true-match standards that the approach does not accurately identify within the top five pairs. If there are systematic differences between the accurate and inaccurate cases in terms of the variables within the embeddings, then it would be possible to identify difficult-to-match standards beforehand and use fine-tuned models for classifying them with their respective true-match standards.

The current results also showed that the accuracy of the proposed approach was slightly higher in grades 8 and 12 compared to grade 4. One potential explanation for the increase in accuracy is that the mean cosine similarity values of the top NAEP standard for each grade 4 standard (M = .145, SD = .048) is lower than that for grades 8 (M = .193, SD = .049) and 12 (M = .188, SD = .056). In line with these cosine results, it may be the case that the language used in the standards for grades 8 and 12 necessarily uses greater specificity to describe the more advanced concepts and phenomena than do the standards for grade 4, which would then mean that the NAEP standards with which the grade 8 and grade 12 standards were aligned would be more likely to share that high-specificity language, resulting in overall higher mean cosine similarity values and therefore greater accuracy.

The usefulness of the current work should be interpreted with caution in light of its limitations. Namely, our current labeled dataset consisted of alignment of only science state standards with only NAEP standards. Subsequent work will examine the usefulness of the approach for standards belonging to different domains (e.g., mathematics, English language arts), as well as items-to-standards alignment. Doing so is critical to establish the generalizability and broad utility of the approach. Moreover, examining the approach in different contexts will help identify boundary conditions. Once boundary conditions on the usefulness of the current simple approach are identified, we will examine the extent to which various modifications and extensions can address those boundary conditions. For example, few-shot classification may be useful because it can fine-tune

the language models based on very small, labeled datasets (i.e., only a few labeled cases per category/standard) to be more sensitive to the linguistic features that indicate alignment in a given context and therefore more accurately predict corresponding standards.

Overall, the current method shows potential for assisting SMEs in conducting content alignment by greatly reducing the number of viable pairs they must consider. We opted for an approach that assisted humans and retained their decision-making role rather than attempting to automate the task, as human judgment is the gold standard for alignment, and stakeholders may be relatively hesitant to endorse a fully automated approach in practice. Our future work in this context must continue to refine the method and examine boundary conditions to realize broad adoption and in turn, promote a more efficient and economical alignment process.

References

Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations.*

Daro, P., Hughes, G. B., & Stancavage, F. (2015). *Study of the alignment of the 2015 NAEP mathematics items at grades 4 and 8 to the common core state standards (CCSS) for mathematics* (Tech. Rep.). Commissioned by the NAEP Validity Studies (NVS) Panel. Retrieved from https://www.air.org/sites/default/files/2021-06/Study-of-Alignment-NAEP-Mathematics-Items-common-core-Nov-2015.pdf

Doran, H. C., & Van Wamelen, P. B. (2010). Application of the Levenshtein distance metric for the construction of longitudinal data files. *Educational Measurement: Issues and Practice*, *29*(2), 13-23.

Dynamic Learning Maps Consortium. (2013). *Dynamic learning maps essential elements for mathematics* (Tech. Rep.). Lawrence, KS: University of Kansas. Retrieved from https://dynamiclearningmaps.org/sites/default/files/documents/Math_EEs/DLM_Essential_Elements_Math_%282013%29_v4.pdf

Jackson, A. (2015). There's a surprising explanation for why the SAT is changing its format. *Business Insider*. Retrieved from https://www.businessinsider.com/the-sat-is-getting-a-format-change-to-align-to-the-common-core-2015-6

Khan, S., Rosaler, J., Hamer, J., & Almeida, T. (2021). Catalog: An educational content tagging system. In *Edm.*

Lewis, D., & Cook, R. (2020). Embedded standard setting: Aligning standard-setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice*, *39*.

Mitzel, H., Lewis, D., Patz, R., & Green, D. (2001). The Bookmark procedure: Psychological perspectives. In (p. 249-281).

Neidorf, T., Stephens, M., Lasseter, A., Gattis, K., Arora, A., Wang, Y., ... Holmes, J. (2016). *Comparison between the next generation science standards (NGSS) and the national assessment of educational progress (NAEP) frameworks in science, technology and engineering literacy, and mathematics.* (Tech. Rep.). Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/nationsreportcard/science

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Schneider, M. C., Agrimson, J., & Veazey, M. (2022). The relationship between item developer alignment of items to range achievement-level descriptors and item difficulty: Implications for validating intended score interpretations. *Educational Measurement: Issues and Practice*, *41*(2), 12-24.

USDOE. (2018). *A state's guide to the U.S. Department of Education's assessment peer review process.* https://oese.ed.gov/offices/office-of-formula-grants/school-support-and-accountability/standards-and-assessments/.

Vockley, M. (2009). *Three approaches to aligning the National Assessment of Educational Progress with state assessments, other assessments, and standards* (Tech. Rep.). Council of Chief State School Officers. Retrieved from https://www.commoncorediva.com/wp-content/uploads/2014/12/alignment_and_the_states_2009.pdf

Webb, N. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. research monograph no. 6.

WIDA. (2020). *WIDA English language development standards framework, 2020 edition: Kindergarten–grade 12* (Tech. Rep.). Board of Regents of the University of Wisconsin System. Retrieved from https://wida.wisc.edu/sites/default/files/resource/WIDA-ELD-Standards-Framework-2020.pdf

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... others (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhou, Z., & Ostrow, K. S. (2022). Transformer-based automated content-standards alignment: A pilot study. In G. Meiselwitz et al. (Eds.), *HCI international 2022 - late breaking papers. interaction in new media, learning and games - 24th international conference on human-computer interaction, HCII 2022, virtual event, June 26 - July 1, 2022, proceedings* (Vol. 13517, pp. 525–542). Springer. Retrieved from https://doi.org/10.1007/978-3-031-22131-6_39