AERA Open
January-December 2021, Vol. 7, No. 1, pp. 1–18
DOI: 10.1177/23328584211028615
Article reuse guidelines: sagepub.com/journals-permissions

A Natural Language Processing Approach to Measuring Treatment Adherence and Consistency Using Semantic Similarity

Kylie L. Anglin
University of Connecticut
Vivian C. Wong
Arielle Boguslav
University of Virginia

Though there is widespread recognition of the importance of implementation research, evaluators often face intense logistical, budgetary, and methodological challenges in their efforts to assess intervention implementation in the field. This article proposes a set of natural language processing techniques called semantic similarity as an innovative and scalable method of measuring implementation constructs. Semantic similarity methods are an automated approach to quantifying the similarity between texts. By applying semantic similarity to transcripts of intervention sessions, researchers can use the method to determine whether an intervention was delivered with adherence to a structured protocol, and the extent to which an intervention was replicated with consistency across sessions, sites, and studies. This article provides an overview of semantic similarity methods, describes their application within the context of educational evaluations, and provides a proof of concept using an experimental study of the impact of a standardized teacher coaching intervention.

Keywords: *fidelity, implementation, document similarity, replication, text as data, latent semantic analysis, treatment adherence, NLP, natural language processing*

When educational interventions are delivered at scale and outside controlled environments, they are rarely enacted exactly as their designers expected. Evaluators need implementation research to understand what a program looks like in the field as opposed to "in theory or on the drawing board" (Durlak, 2015, p. 1124). In evaluation contexts, implementation research often examines treatment fidelity to a prespecified program theory, addressing the question: *Was the intervention implemented as planned?* (Century & Cassata, 2016; Dumas et al., 2001; Nelson et al., 2012; O'Donnell, 2008). Even in cases where adaptations to the intervention are desired, measures of implementation help researchers understand program variations, interpret effects, and generate new hypotheses about how the intervention may be improved (Century & Cassata, 2016; Durlak & DuPre, 2008).

Unfortunately, researchers face intense logistical, methodological, and budgetary constraints in their efforts to understand program delivery. Traditional approaches to measuring fidelity require the development and validation of reliable measures for each new intervention (Gresham, 2017; Sanetti & Kratochwill, 2009). The researcher then needs to hire, train, and employ observers to rate each intervention session, a process which is time-consuming, expensive, and, at times, infeasible. More commonly, researchers employ less resource-intensive approaches; they may sample a few sessions for in-depth analysis, use easy-to-collect data like attendance and administrative records, or examine responses to self-report surveys. In some cases, researchers fail to collect any implementation data at all (Dusenbury et al., 2003; O'Donnell, 2008).

In this article, we propose a method for measuring implementation delivery using scalable natural language processing (NLP) techniques that provide richer information than attendance counts alone, and more representative information than sampling a few intervention sessions for in-depth analysis. The proposed methods are most applicable in settings where the intervention is delivered through verbal interactions with participants, and the interventionalist is expected to follow a structured protocol containing suggested language. These sorts of semiscripted protocols are common in multiple domains of education, including in special education, behavioral education, and reading and mathematics instruction for struggling learners. For example, students with autism spectrum disorder are often taught "social scripts" for improving language skills and peer interactions (Ganz et al., 2008; Goldstein, 2002; Stevenson et al., 2000), and in behavioral interventions such as Positive Behavioral Interventions and Supports, teachers learn to use highly structured consistent

approaches for redirecting students' off-task behaviors (Horner & Sugai, 2015). Structured implementation protocols are also commonly applied in pedagogical contexts that require unambiguous instruction, where even small variations in definitions and examples can result in student confusion. For these skills, intervention curricula often provide teachers with "highly structured guidance in wording, sequencing, and review of material" (Stockard et al., 2018, p. 481).

In this article, we use a branch of NLP techniques called semantic similarity to determine the extent to which standardized intervention protocols are delivered consistently and with adherence in field settings. At its core, semantic similarity quantifies the similarity between two or more texts based on their linguistic characteristics. In evaluation contexts, we use semantic similarity approaches to quantify variations in intervention transcripts. Importantly, the method characterizes similarities in semantic content but is robust to arbitrary differences in language that do not change the meaning of the text.

We apply semantic similarity to create two new implementation measures that are of interest in evaluation settings. To measure *intervention adherence*, we examine the semantic similarity of intervention transcripts to a scripted (though flexible) intervention protocol. The measure quantifies "the degree to which specified procedures are implemented as planned" (Dane & Schneider, 1998, p. 23) in highly scripted intervention settings, capturing the extent to which an implementer may have deviated from the protocol by omitting key components of the intervention or by introducing new aspects to the intervention (such as discussing unexpected topics). Semantic similarity may also be used to describe how consistently an intervention was delivered by measuring variation in intervention transcripts. To produce a measure of *intervention replicability*, we calculate the semantic similarity of transcripts within and across potential sources of variation (e.g., participants, interventionists, sites, or studies). While the adherence measure will often be of most interest when a program theory prioritizes adherence to a structured protocol, the replicability measure may be used to characterize program variations in field settings even when adaptations to the intervention are desired or of research interest.

Though semantic similarity techniques have a long history in computer science and information retrieval (Manning et al., 2008; Salton & Buckley, 1988), these methods are new in their application to implementation research. This article serves as a primer on NLP methods for semantic similarity, generally, and demonstrates how they can be used to analyze implementation in education settings specifically. To this end, we apply the approach to a series of randomized control trial (RCT) evaluations in teacher education that examine the impact of TeachSIM—a 5-minute structured coaching protocol—on preservice teachers' pedagogical performance in simulated classroom environments (Cohen et al., 2020). Through this application, we show that semantic similarity measures of intervention adherence and replicability are a scalable and informative option for implementation research, particularly when resources are scarce.

The Use of Natural Language Processing in Implementation Research

This article sits within a burgeoning literature that applies NLP techniques to education data (Reardon & Stuart, 2019) and, more specifically, within a smaller body of literature that uses NLP to describe program implementation. NLP allows researchers to analyze large bodies of text data (whether written or verbal) to gain insights into educational processes. The methods can range from relatively simple dictionary-based approaches (e.g., searching texts for instances of key terms) to more state-of the art machine learning algorithms. For example, in a study describing district responses to deregulation under the Texas District of Innovation statute, Anglin (2019) used machine learning algorithms to identify relevant policy documents scraped from school district websites and then to document variations in regulatory exemptions claimed by school districts. In another study, Sun et al. (2019) used topic modeling, a method of automatically extracting patterns of semantic meaning (topics) from text, to document policy variations in reform strategies found in school improvement plans.

NLP may also be used to automate human ratings of intervention fidelity. In these approaches, researchers first hand-label a subset of documents and then use a machine learning classification approach to recognize text features that correspond to the hand-labels. In a study of text-message based college counseling, Fesler (2020) trained a classifier to identify productive engagement between college counselors and text-message recipients. Kelly et al. (2018) used an automated classifier to identify authentic questioning by teachers. These articles demonstrate that automated classification techniques can identify complex, substantively meaningful features of implementation. Furthermore, they take advantage of the highly scalable nature of NLP—once the algorithm has been trained, it may be applied to new treatment sessions at negligible additional cost. However, classification algorithms require substantial start-up costs, that is, they require that researchers develop a valid and reliable coding system for hand-labeling documents, as well as enough documents that have been correctly labeled for the classifier to produce accurate results. For example, Fesler hand-labeled 551 interactions while Kelly et al. hand-labeled 451 documents.

The NLP methods used in this article do not require the same start-up costs as classifier approaches. Semantic similarity only requires transcripts from intervention sessions and a scripted protocol that the researcher deems representative

of high-quality intervention delivery. However, semantic similarity does not have the flexibility that classifier approaches have in capturing potentially nuanced features of implementation. Instead, it provides a summary rating of how similar or different intervention sessions are from a standardized protocol (though with some degree of flexibility allowed). In this way, semantic similarity may be understood as a highly efficient and scalable—but narrow—measure of intervention fidelity.

An Introduction to NLP Techniques for Semantic Similarity

In this section, we provide an overview of NLP techniques researchers may use to calculate the semantic similarity of texts. We begin by defining a few terms within the NLP context: A *document* is a single text of interest, and a *corpus* is the full set of documents a researcher is interested in analyzing. To measure semantic similarity, the researcher first *tokenizes* the documents, separating the strings of text into a set of units (most commonly words), and *vectorizes* the corpus, representing the texts numerically. Then, the researcher calculates the similarity of the vectors using a distance metric like cosine similarity (discussed later in this section).

In order to vectorize a corpus, researchers often create a *document-term matrix* where each row corresponds to a document $\left(i = 1,\ldots,N\right)$ and each column corresponds to a word in the corpus. Then, each document is represented by a vector $W_i = \left(W_{i1}, W_{i2}, \ldots, W_{im}\right)$, where W_{im} counts the frequency of the *m*th word in the *i*th document. The values in the columns are the frequency with which a document uses each word.

Prioritizing the Words That Matter

When document-term matrices contain every word in the corpus, they quickly grow to very large dimensions. Yet many of these words are unlikely to be useful in discriminating between texts. In particular, there will be a number of words that are common in every document, but that add very little meaning: words like *a, an, the*, and *to*. These words are referred to as *stop words*, and a first step to better prioritize important terms in a document-term matrix is to remove these words. In practice, researchers do not need to create a list of stop terms on their own as many software packages maintain predefined lists. However, researchers may edit these lists to better suit their context.

In addition to removing stop words, researchers may choose to weight words in their document-term matrix so that the words that are mostly likely capable of discriminating between documents are given greater weight. The most commonly applied weighting technique is term frequency–inverse document-frequency (tf-idf), which assigns weights based on a word's relative frequency in the full corpus of documents. Formally, tf-idf weights are determined by the following formula:

$$tf\text{-}idf_{t,d} = tf_{t,d} * \log\frac{N}{df_t}.$$

The greatest weight is given to words that occur many times in a few documents. The least weight is given to words that occur only a few times in a document and to words that occur in many documents. This system of weighting will down-weight stop words (without the researcher defining which words are common across all documents) while weighting the words in an extended but uncommon topic of conversation heavily.

Incorporating Shared Meaning Between Words

Without additional preprocessing, all words in a document-term matrix are treated as wholly distinct from one another. This is problematic when considering word derivatives like *teach* and *teaches*; it would not be appropriate to consider these words as having no shared meaning. To this end, document-term matrices can be improved by reducing each word to its root form through *lemmatization*—for example, after lemmatizing *teach, teacher, teachers*, and *teaches* would all be represented by the root word, *teach*.

Even after lemmatizing, we still fail to capture the similarity of words with different roots. To address this, the researcher can incorporate latent semantic analysis (LSA; Deerwester et al., 1990; Landauer et al., 1998). LSA uses singular value decomposition, a general form of factor analysis, to reconstruct the document-term matrix so that the first column contains the most information (capturing the most variance from the original matrix), the second a little less, and so on. Each column in the new matrix may be loosely understood as an abstract concept (composed of words that tend to occur in similar contexts) and the reconstructed matrix can be used as a dimension reduction method where the researcher only uses the first *X* concepts. It is up to the researcher to determine the number of abstract concepts to include, but between 50 and 300 is a common rule of thumb depending on the size of the corpus.[1]

Finally, if a researcher wishes to retain some of a word's context, they may create the document-term matrix using bigrams (word pairs), trigrams (word triples), or any *n*-gram. A document-term matrix made of bigrams would create a new term for every word pair. For example, the phrase, *work on your behavior management*, would be represented as a set of four bigrams: *work on, on your, your behavior, behavior management*. All of the above techniques have the advantage of being easily applied using common statistical software packages.[2] For a short overview of more advanced NLP techniques which may require more additional programming skills, see Appendix A.

Calculating Semantic Similarity

After preprocessing the texts and vectorizing the corpus, a researcher can calculate the *cosine similarity* of any two documents, d_1 and d_2 using the following formula:

$$sim(d_1, d_2) = \frac{\vec{V}_1(d_1) \cdot \vec{V}_2(d_2)}{\left|\vec{V}_1(d_1)\right|\left|\vec{V}_2(d_2)\right|}.$$

The numerator here is the dot product of the two document vectors: in other words, the sum of the products of the two documents' values in each column. The denominator is the product of the magnitude of the two vectors. In a simple document-term matrix, this would normalize the measure by the length of the documents so that it is the *relative* word frequencies which matter, rather than simply the percent of words shared between the documents. Cosine similarity measures may also be understood as the cosine of the angle between two document vectors. If two documents have equivalent relative word frequencies, the angle between their vectors will be zero degrees and their cosine similarity will be one (as the cosine of zero is one). If two documents do not share any terms, then, they will be perpendicular to one another, and their cosine similarity will be zero.

Semantic Similarity Measures of Adherence and Replicability

With semantic similarity, adherence scores can be determined by examining the cosine similarity of intervention transcripts and a scripted treatment protocol. In general, the scripted protocol should include all core components of the intervention with suggested language for how each component should be delivered. We provide an example of such a script, labeled with components from a teacher coaching protocol, in Appendix B.[3] With this protocol, and the set of intervention transcripts, the researcher creates a document term-matrix. Then, for a given transcript of an intervention session, document d_i, and a scripted protocol, s, script similarity is determined by

$$Script\ Similarity_i = sim(d_i, s),$$

where $sim(d_i, s)$ is the cosine similarity of the two documents ranging from 0 to 1. From there, the researcher can determine which intervention sessions are most similar to the scripted protocol and which intervention sessions deviate more substantially. The researcher can also calculate the average script similarity for a study, site, or interventionist to compare relative intervention adherence.

Similarly, a researcher can measure the replicability (consistency) of intervention delivery by calculating the similarity of intervention transcripts to one another. The researcher calculates a pairwise similarity measure where each transcript in a study is compared with every other transcript in that study. The average similarity of document d_j to every transcript in a set of n transcripts including document d_j is calculated as

$$Similarity\ of\ d_j\ to\ the\ set = \frac{\sum_{i=1}^{n} sim(d_i, d_j) - 1}{n - 1}.$$

Here, we subtract 1 from the numerator and denominator so that the similarity of d_j to itself is not included. Then, the measure of intervention replicability is calculated using the following formula:

$$Within\text{-}Study\ Similarity = \frac{\sum_{i=1}^{n} Similarity\ of\ d_i\ to\ the\ set}{n},$$

where replicability is measured as the average similarity of each document to every other document in the set.

A researcher can also measure consistency across potential sources of variation, like implementers, sites, or studies, by calculating across-group similarity. Consider two groups of transcripts, Group 1 and Group 2, where Group 1 has n documents and Group 2 has m documents. Then, the similarity of Group 1's document j to Group 2 is calculated by comparing document j with every document in Group 2:

$$Similarity\ of\ d_j\ to\ Group\ 2 = \frac{\sum_{i=1}^{m} sim(d_{1j}, d_{2i})}{m},$$

and the average similarity of Group 1 and Group 2 is calculated as

$$Across\ Group\ Similarity = \frac{\sum_{i=1}^{n} Similarity\ of\ d_i\ to\ Group\ 2}{n}.$$

Similar to the adherence measure described above, this method yields a replicability score that ranges between 0 and 1, where 1 indicates perfect consistency and 0 indicates no semantic overlap across transcripts. The replicability measure can identify which implementers, sites, or studies are most similar to one another in terms of intervention delivery and may be especially useful in cases where intervention adherence is low, but the researcher wants to know whether sessions strayed from the protocol in similar ways. Understanding both dimensions of intervention fidelity—adherence and replicability—provides the researcher with important insights for understanding how the intervention was actually delivered, as well as for developing appropriate implementation supports.

The Impact of Preprocessing on Semantic Similarity Scores

The magnitude of semantic similarity measures depends not only on the similarity between two texts but also on the

size and characteristics of the vector space (the terms of comparison in the document-term matrix). Appendix C provides some intuition for how different preprocessing techniques alter semantic similarity scores within the TeachSIM context and demonstrates a few patterns. First, any two texts will almost certainly have a higher semantic similarity if cosine similarity is calculated on a document-term matrix with no pre-preprocessing compared with one where we have removed the stop words. This is because the texts have many stop words in common and by removing them, we are purposefully ignoring these similarities. Similarly, tf-idf weighting will, by definition, decrease the cosine similarity between texts as it gives greater weight to words that are uncommon. On the other hand, preprocessing techniques that attempt to address word similarities, like lemmatization and LSA, will *increase* the cosine similarity of documents. These techniques both reduce the size of the vector space and give documents credit for using similar words.

Because differing approaches to semantic similarity will result in measures on a different scale, we have to be careful in our interpretation of intervention adherence and replicability measures. We cannot, for example, set an a priori cut score of 0.50 to indicate low adherence; a transcript may be well above a 0.5 cutoff before stop words have been removed and well below after. A single semantic similarity score on its own carries very little meaning. It is only through comparisons across documents and sources of variation that we gain insight. We recommend comparing patterns of intervention adherence and replicability scores across several modeling approaches, which we will demonstrate in the applied example below.

An Application to TeachSIM

In this section, we apply our proposed measures of intervention adherence and replicability to the TeachSIM coaching protocol. In the TeachSIM context, teacher candidates practice an instructional task—either leading a text-based discussion or managing off-task student behaviors—for 5 minutes with student avatars in a mixed-reality simulated classroom environment. Treated teachers then participate in a 5-minute coaching conversation with a master educator designed to improve their pedagogical performance. During these sessions, coaches could choose one of four structured protocols, corresponding to four different targeted skills depending on the coach's assessment of the teacher candidate's strengths and weaknesses. In coaching conversations following simulations of text-based discussions, the four targeted skills for teachers included probing for textual evidence, scaffolding student understanding, providing descriptive feedback, or probing for a warrant. In conversations following behavior management simulations, the targeted skills included providing redirections that are timely, specific, succinct, or calm.

We analyze these coaching conversations across five conceptual RCT replications; Table 1 presents summary statistics for the RCTs. Three studies focused on behavior management (Behavior Studies 1, 2, and 3) while two focused on text-based discussions (Feedback Studies 1 and 2). Feedback Study 1 was the first study conducted and was used as a pilot to develop the coaching protocol. The goals of applying the semantic similarity measure in TeachSIM were to provide evaluation researchers with summary quantitative measures of the extent to which coaching protocol was delivered to treatment participants with adherence and consistency within and across studies, and to allow researchers to identify outlier sessions that may inform future training of coaches.

Coaching Protocol and Benchmark Scripts

Benchmark scripts were developed by a coaching expert with careful attention to the intervention's theory of change. After careful review of the existing coaching protocol and training documents, the expert identified five central components where coaches: (1) ask the candidate to assess their own performance; (2) affirm an observed effective teaching practice, explaining why the practice was effective; (3) identify and explain one of four skills for the candidate to target in the next session; (4) engage the candidate in role-play so that the candidate can practice their targeted skill; and (5) close the coaching session with positive reinforcement. Then, the coaching expert represented each of these components using idealized language, generating benchmark scripts. Because of variations in teachers' targeted skills and instructional tasks, the treatment protocol was represented by eight ideal scripts—one script for each targeted skill for the two instructional tasks. Appendix B shows an example script and how it aligns with the treatment protocol.

Transcripts

Coaching sessions were video-taped and transcribed using a professional transcription service. Table 1 presents the number of transcripts in each study. Sample sizes ranged from 45 to 76 coaching sessions per study. In the transcripts, each utterance was preceded by a speaker tag (where *Coach*: designates that coach speech follows and *TC*: designates that teacher candidate speech follows) and a time stamp (in the format *[hh:mm:ss]*). We cleaned plain text transcripts to exclude these speaker tags, time tags, and any formatting characters (for example newline, \n).[4] We also excluded teacher candidate dialogue to focus our analysis on coaches' implementation of the protocol rather than teacher candidate's reactions to the coach.[5] After cleaning the transcripts, the average length of coach text was 681 words.[6]

TABLE 1
Sample and Setting Characteristics by Study

Characteristics	Behavior Study 1	Behavior Study 2	Behavior Study 3	Feedback Study 1	Feedback Study 2
Sample characteristics of teacher candidates					
GPA	3.42	3.46	3.54	3.45	3.51
% Female	1.00	0.88	0.50	0.88	0.98
% Over the age of 21	0.18	0.16	0.08	0.42	0.19
% White	0.56	0.63	0.56	0.77	0.69
Location of high school attended					
% Rural	0.03	0.12	0.09	0.24	0.13
% Suburban	0.86	0.82	0.79	0.68	0.85
% Urban	0.11	0.06	0.13	0.07	0.02
Average SES of high school attended					
% Low SES	0.04	0.00	0.00	0.08	0.00
% Middle SES	0.59	0.61	0.57	0.61	0.68
% High SES	0.32	0.28	0.40	0.31	0.28
Majority race of high school attended					
% Primarily students of color	0.10	0.03	0.06	0.07	0.04
% Mixed	0.48	0.47	0.41	0.39	0.51
% Primarily White students	0.42	0.50	0.53	0.54	0.45
Pedagogical task in simulation	Behavior Management	Behavior Management	Behavior Management	Providing Feedback	Providing Feedback
Timing	Spring 2018	Spring 2019	Fall 2019	Fall 2017	Fall 2018
N (treatment transcriptions)	68	45	47	76	46
Mean and standard deviation (in brackets) of adherence scores from semantic similarity measure	0.23 [0.05]	0.26 [0.06]	0.23 [0.06]	0.16 [0.06]	0.36 [0.09]

Note. SES = socioeconomic status.

Method

Preprocessing. Before applying any of the NLP techniques discussed earlier in this article, we first created a context-specific dictionary where we replaced all student avatar names (Ethan, Ava, Dev, etc.) with the word *avatar*. We made a similar dictionary for off-task behaviors that the avatars might display (singing, humming, etc.), replacing them with the word *misbehavior*. This dictionary ensured that words which shared a similar meaning in our context were treated similarly in the analyses; for example, from an adherence perspective, it is unimportant whether a coach discusses one student avatar's behavior or another and so we do not discriminate between their names.

After replacing contextual synonyms, we created five document-term matrices using our full corpus of documents, including all transcripts and ideal scripts. In the first matrix, we included all of the terms in the corpus with no preprocessing. In the second matrix, we excluded stop words from a popular prespecified list (Python's Natural Language Toolkit—NLTK) supplemented with a set of common pause fillers and vocal ticks like "uh" and "um." In the third matrix, we additionally lemmatized the words, replacing all word derivatives with a single stem. In the fourth matrix, we incorporated tf-idf weighting, and finally, in our fifth matrix, we incorporated LSA.[7]

Analysis. After creating our document-term matrices, we calculated adherence scores for each transcript by measuring the cosine similarity between each transcript and the appropriate ideal script (matching the transcript's scenario and targeted skill). We then averaged the adherence scores of every transcript within each study to create summary adherence scores. We also calculated five replicability scores for each transcript by measuring the average similarity of every transcript in each study to transcripts from Behavior Study 1, Behavior Study 2, Behavior Study 3, Feedback Study 1, and Feedback Study 2. When a transcript was compared with transcripts within the same study (e.g., when we calculated the similarity of a Behavior Study 1 transcript to other Behavior Study 1 transcripts), we consider the score a *within-study replicability* measure. When a transcript was compared with transcripts from other studies, we consider the score an *across-study replicability* measure.

TABLE 2
Study Adherence

Study	(1)	(2)	(3)	(4)	(5)
Behavior Study 1	0.69	0.36	0.23	0.25	0.33
Behavior Study 2	0.74	0.39	0.26	0.28	0.38
Behavior Study 3	0.72	0.36	0.23	0.24	0.32
Feedback Study 1	0.63	0.3	0.16	0.18	0.25
Feedback Study 2	0.74	0.51	0.36	0.39	0.54
Remove stop words		X	X	X	X
Tf-idf weighting			X	X	X
Lemmatization				X	X
LSA					X

Note. Adherence scores were estimated by calculating the cosine similarity between each transcript and the appropriate benchmark script. Shading indicates a higher ranking by average adherence score for each study where a darker shading indicates higher adherence. Tf-idf = term frequency–inverse document-frequency; LSA = latent semantic analysis.

TeachSIM Results

Intervention Adherence. Table 2 shows the average adherence score for each study across each of the five preprocessing approaches. Given that semantic similarity scores are sensitive to analytic decisions, it is important that researchers observe whether patterns are consistent across techniques. In Table 2, we rank the studies from lowest to highest adherence within each preprocessing approach where darker shading indicates higher adherence. This table indicates the robustness of patterns; shading is relatively consistent across techniques.[8] Given this robustness, we limit our remaining discussion of results to one, relatively simple, method of text processing for ease of interpretation: removing stop words and applying tf-idf weighting. However, readers can view Appendix C for details on the results produced by each text processing method.

Table 1 provides an example of how adherence scores might be included in summary tables alongside other statistics like sample sizes and participant characteristics. Like the other information in Table 1, the adherence scores allow readers to quickly compare a key characteristic across studies. For example, Table 1 shows that Feedback Study 2 was the highest adherence study while Feedback Study 1, the pilot, was the lowest. We dig further into these results in Figures 1 and 2, demonstrating how adherence scores can be used for monitoring program delivery. Here, we have created a histogram of adherence scores for each transcript in every study. Where transcripts seem to stray from the distribution (highlighted in black), we recommend that researchers check to see if there are any transcription errors, implementer misunderstandings that need to be corrected, or conditions which result in particularly high adherence. Figure 2 provides an example of how researchers might use adherence scores to inform ongoing training. Here, we show that that there are two coaches with highly variable adherence scores. This suggests that these coaches, in particular Coach A, could benefit from additional training.

Interpreting Adherence Scores From Semantic Similarity Measures. A common question with semantic similarity scores is, "How close to the benchmark script is close enough?" An answer to this question requires some interpretation by experts with subject-matter knowledge; in other words, semantic similarity scores are most useful with a "human in the loop." To answer questions of interpretation within the TeachSIM context, we took two approaches which we recommend that researchers apply in their own contexts: an informal validation effort and a qualitative analysis.

First, we asked a coaching expert who was blinded to semantic similarity scores to pull three examples of ideal implementation and three examples of inadequate implementation of the behavior study protocol. We then observed where these transcripts lay on the distribution of script similarity scores. The scores of these transcripts are represented as stars on Figure 3. The figure shows that the three substandard transcripts are well below the median adherence score (0.24), indicating that the adherence scores are able to identify the transcripts which deviate too far from the protocol. The three transcripts identified as ideal implementations of the protocol are above the median, but not substantially so. This suggests that the measure is better able to identify low-fidelity transcripts than high-fidelity transcripts.

To gain an intuition for the meaning behind script similarity scores, we recommend that researchers sample transcripts from both ends of the distribution for qualitative analysis. In the TeachSIM behavior study context, we pulled the four transcripts with the highest adherence scores, the four transcripts with the lowest adherence scores, and the four transcripts closest to the median. Figure 3 highlights these transcripts in black on the histogram. We then asked a coaching expert to identify to what degree the coach addressed core components of the coaching protocol: identifying and explaining one of five potential strengths of the teacher candidate (acknowledging misbehavior, and/or providing redirections that are specific, succinct, timely, and calm), identifying and explaining one of four potential areas of growth, and engaging the candidate in role-play. Qualitative analysis reveals that the transcripts with the lowest semantic similarity adherence scores commonly include off-topic, unclear, and unfocused conversations and are often missing one or more of the core treatment components. In the four moderate and high adherence transcripts, on the other hand, implementation is, generally speaking, good enough; the coach never fails to identify and define a

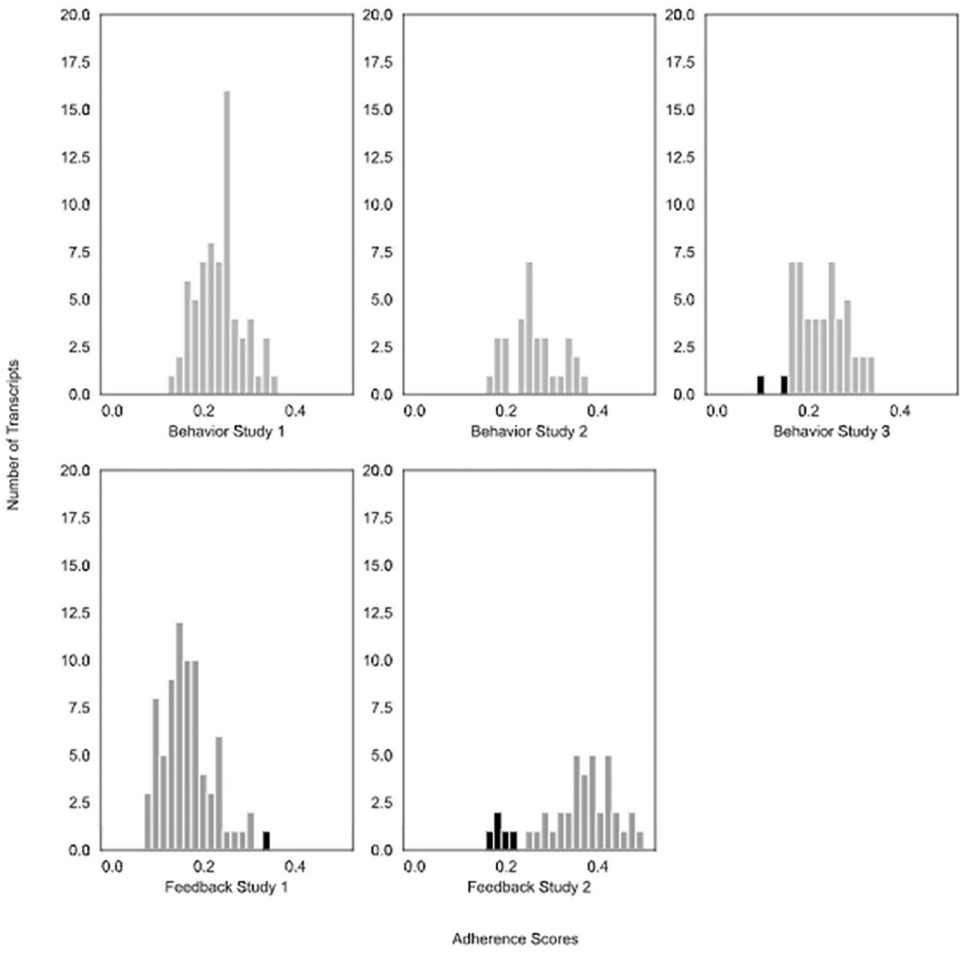

FIGURE 1. *Distribution of adherence scores by study, with unusual transcripts highlighted.*
Note. Adherence scores were estimated by calculating the cosine similarity between each transcript and the appropriate benchmark script using a document-term matrix with no stop words and tf-idf weighting. A higher score indicates higher adherence to the script. Potentially abnormal transcripts (based on visual examination) are highlighted in black. These are transcripts we have flagged for manual inspection.

strength, identify and define an area of growth, or engage the candidate in role-play.

As an example, the following excerpt is from the lowest adherent transcript (0.09). The coach begins in a short off-topic conversation[9] about the simulator and does not clearly explain how the candidate's strength (acknowledging misbehavior) benefits students:

> So, what's interesting about this is that even though [the simulator] seems so odd, it actually helps teachers to build muscle memory. Yes. So, it's actually pretty effective. . . . Okay, so I'm glad that you're interested by it. So, you definitely have some really good moves. So, you know, maybe thinking about teaching somewhere in your life like maybe professorship. So, one thing you did really well was noticing the kid who was starting to act out and we're going to just shape that a little bit, shape that a little bit to make it more precise. . . . So, as you went along what's really cool about you is that as went go along you got more proficient. And so, you're already sensing some of these things that we're going to talk about.

We can contrast this with a high adherence transcript which quickly and clearly identifies and explains the definition and importance of a strength (remaining calm; 0.37):

> So, how do you think that went in terms of your abilities to redirect Ethan or Dev's behavior. . . . One of the things that I saw that I really liked is that you keep your cool. That's the first piece that can really throw people off when they have a lot of redirections. Then what is a likely consequence of that? They'll feed off of your anxiety and then you'll have students that are likely to take advantage of that. So, the remedy for this and to build off of your sense of calm is to

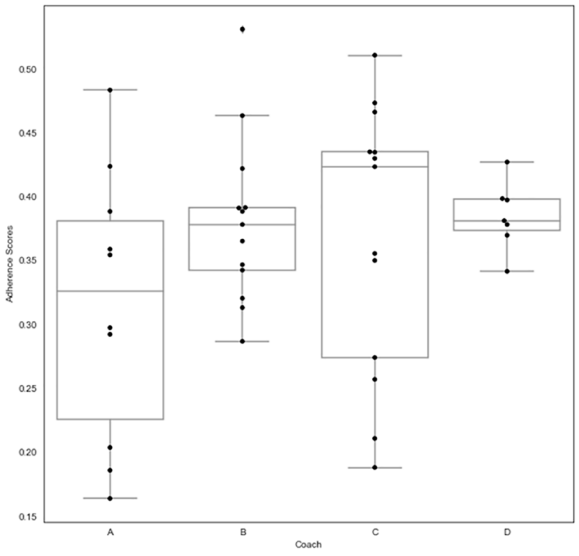

FIGURE 2. *Distribution of adherence scores by coaches within Feedback Study 2.*

Note. Adherence scores were estimated by calculating the cosine similarity between each transcript and the appropriate ideal script using a document-term matrix with no stop words and tf-idf weighting. A high score indicates higher adherence to the benchmark script. Boxes indicate the 50th percentile and interquartile range. Whiskers extend to all scores within 1.5 times the interquartile range. tf-idf = term frequency–inverse document-frequency.

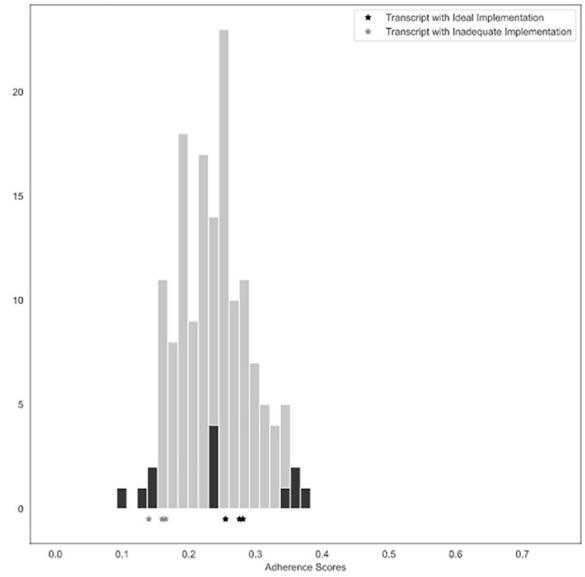

FIGURE 3. *Distribution of adherence scores in behavior studies, with transcripts analyzed by coaching expert highlighted and starred.*

Note. A coaching expert who was blinded to the adherence scores pulled three transcripts representing ideal implementation of the protocol and three transcripts representing inadequate implementation of the protocol. The scores from these transcripts are represented as stars on the above plot, where gray stars indicate inadequate implementation and black stars represent ideal implementation. We also pulled four transcripts with the lowest adherence scores, four transcripts with the highest adherence scores, and four transcripts which were closest to the median for qualitative analysis. These transcripts are represented by black bars in the histogram. Adherence scores were estimated by calculating the cosine similarity between each transcript and the appropriate benchmark script using a document-term matrix with no stop words and tf-idf weighting. A higher score indicates higher adherence to the script. tf-idf = term frequency–inverse document-frequency.

> address behaviors as soon as you notice them and be very specific with redirecting.

Another low-adherence transcript (0.15) demonstrates again how off-topic and unfocused conversations can crowd out other treatment components. The excerpt begins in an off-topic conversation about not being able to use detentions in the simulator and identifies several strengths and growth areas, rather than focusing on one of each, without any explanation of the definition or importance of these skills:

> I think [the] key when you're providing behavioral redirections—there are layers on this but at the base level, like simple, very specific. It is exactly what you want to have happen. . . . I was like, "No way detentions will work." Um, which you might not have in the in the classroom, so I understand that. I think you did a really nice job. And I think one of the things for behavioral redirections is being very specific. It's the same to being really calm and I think you did just a really nice job of it. So, like I this is going to come across as like me not having much to say, but it's just because you did a really nice job. You know there are things that you can say. Another thing you can think about the next time that Ethan is talking or Dev is talking or whoever is talking, you can use a lot of non-verbals too, if you feel comfortable with that. It still is very specific as to what you want. And you can be succinct. And it's also like pretty calm.

The above transcript exemplifies a key lesson learned from this exercise which could be used for program improvement: When coaches identify multiple strengths or multiple areas of growth, they often fail to clearly explain those skills. These excerpts further identify one strength of semantic similarity measures: They are well suited to identifying off-topic conversations and, to the extent that these off-topic conversations crowd out treatment components, the measure will appropriately flag these transcripts as low-adherence. However, where off-topic conversations do not crowd out treatment components, this feature may be considered a key limitation of the method; the similarity measure will flag off-topic conversations as deviating from the protocol, no matter whether such conversations are useful or harmful.

Another limitation of the measure occurs when an implementer repeatedly uses an uncommon term while successfully delivering treatment components. For example, the following excerpt is from a transcript that received a very low script similarity score (0.13) despite clearly identifying and defining a strength and an area of growth earlier in the transcript:

> So, I will be an off task student, and then you can provide me with some feedback. Yeah. "Ba ba da ba da ba da ba da ba da" It's okay.

TABLE 3
Replicability Matrix

Study	Behavior Study 1	Behavior Study 2	Behavior Study 3	Feedback Study 1	Feedback Study 2
Behavior Study 1	0.31	0.26	0.25	0.12	0.13
Behavior Study 2	0.26	0.31	0.27	0.12	0.14
Behavior Study 3	0.25	0.27	0.33	0.13	0.15
Feedback Study 1	0.12	0.12	0.13	0.25	0.21
Feedback Study 2	0.13	0.14	0.15	0.21	0.37

Note. The replicability index is calculated by calculating the pairwise similarity of each transcript in the study indicated in the first row to each transcript in the study indicated by the first column. Cosine similarity was calculated using a document-term matrix with no stop words and term frequency–inverse document-frequency weighting. Cells shaded in dark gray (on the diagonal) display the similarity of transcripts to other transcripts within the same study. Cells shaded in light gray display the similarity of transcripts to other studies within the same context (behavior management or feedback).

> You could just call me Ethan or Dev or Savannah or whoever. I'll respond to that. "Ba ba da ba da ba."

Here, given the semantic similarity score was estimated with tf-idf weighting on a corpus with no stop words, we suspect that the semantic similarity score is picking up on the repeated use of rare terms that were not included in the stop list: *ba* and *da*. However, these rare words were used by the coach to engage the candidate in role-play, an appropriate application of the coaching protocol.

Finally, we find that both the moderate- and high-adherence transcripts contain the key treatment components and that there are no substantial differences between these two groups of transcripts; all the analyzed transcripts with at least moderate adherence scores were acceptable implementations of the treatment protocol. This again indicates that script similarity may not be distinguishing between good and excellent implementation in the TeachSIM context.

This relatively low-cost qualitative exercise demonstrates the value of human expertise in interpreting semantic similarity scores. By asking an expert with content knowledge—who is blinded to script similarity scores—to identify low-fidelity and high-fidelity transcripts, we gain confidence in the validity of the adherence measure. By sampling several transcripts for qualitative analysis, we gain an understanding for how to interpret different semantic similarity scores for a particular intervention and gain insight for program improvement.

Intervention Replicability Across Studies. Using the replicability measure, we also assessed the extent to which the coaching protocol was implemented consistently within and across the five conceptual replication studies. Table 3 presents a replicability matrix showing the average similarity of transcripts in the row study to transcripts in the column study (similar to a correlation matrix). Intuition would tell us that transcripts should be most similar to other transcripts from the same study and least similar to transcripts from a different simulation context. Indeed, this is what we find. Looking at the Behavior Study 1 column, we see that Behavior Study 1 transcripts have the highest replicability to one another, followed by Behavior Study 2 and Behavior Study 3. Similarly, looking at the Feedback Study 1 column, we see that Feedback Study 2 is the best replication of Feedback Study 1.

The most striking feature of this table is the within-study replicability measure of Feedback Study 2; Feedback Study 2 transcripts are more similar to one another than are other transcripts, indicating a high degree of standardization (as well as adherence, as indicated by Figure 1). This follows from their adherence scores. Transcripts that are close to the benchmark script will be necessarily close to one another. Transcripts that are far from the benchmark script, on the other hand, may or may not cluster together. When replicability scores are used in conjunction with adherence scores, they are most useful for determining the similarity (or dissimilarity) of transcripts that stray from the script. In this case, our lowest adherence study Feedback Study 1, also has the lowest replicability scores, implying that transcripts from this study do not stray from the script in a consistent manner.

Discussion

A semantic similarity approach to measuring intervention adherence and replicability brings many potential advantages. First, so long as a researcher is able to obtain intervention transcripts, semantic similarity methods are nearly infinitely scalable. Researchers only need transcripts and moderate computer programming skills. We hope this will encourage researchers who would not otherwise include measures of fidelity (or who were previously sampling a few intervention sessions for fidelity assessment) to incorporate the measures presented here in their evaluations. Second, the automated nature of semantic similarity techniques means that, given the same transcript, semantic similarity measures

of intervention adherence and replicability will have perfect reliability; if the same method is applied to the same transcript, the same measure will result each time. Third, semantic similarity scores can be calculated in near real-time, potentially reducing the time between implementation and feedback. This allows researchers to use the measures presented here as informal diagnostics to quickly reveal when treatment sessions may be drifting from the protocol. Finally, we believe that our proposed measure of replicability is a novel contribution for replication science. Transcript similarity directly addresses the question of treatment stability and consistency (Steiner et al., 2019; Wong et al., 2020), measuring changes in intervention implementation that may not be captured using an adherence rubric or qualitative analysis.

Despite these advantages, semantic similarity measures are not a one-size-fits all solution. There are two primary considerations that researchers should evaluate. The first consideration is the construct validity. To provide an appropriate measure of adherence and replicability, semantic similarity methods rely on the assumption that the words used in a treatment session matter. For this reason, semantic similarity is most appropriate when the intervention is highly structured. However, even in these cases, researchers should carefully consider whether script similarity measures are inappropriately rewarding rote verbatim intervention delivery. If researchers do not want implementers to deliver a script verbatim, they should carefully frame the measure for implementers and sample high-adherence transcripts for human review to ensure that implementers are appropriately responding to participants.

There are also cases when semantic similarity methods will simply be too blunt to satisfy researchers' needs. Rubrics are capable of measuring multiple components of a theory of change while script similarity measures only a single construct—the similarity between a treatment transcript and a scripted protocol. If a researcher is simply interested in determining the relationship between adherence and the magnitude of a treatment effect, semantic similarity may be effectively incorporated into a model of heterogeneous treatment effects. On the other hand, if a researcher is interested in determining which components in a theory of change have the strongest relationship with effect sizes, semantic similarity is unlikely to be helpful. Furthermore, semantic similarity is limited in its ability to evaluate implementation constructs beyond adherence and replicability. For example, unlike trained observers, the method cannot make evaluative judgments about whether intervention sessions that stray from the benchmark script remain aligned with the intervention's theory and goals.

The second consideration is resources. The greatest cost-saving measure of semantic similarity is that it does not rely on human labor to rate intervention sessions. Based on our conversations with university researchers, we estimate that if we were to have used undergraduate research assistants to rate all 403 five-minute TeachSIM sessions using a fidelity rubric, this would have cost approximately \$15,100.[10] The cost of professionally transcribing intervention sessions for semantic similarity, on the other hand, was \$1,673.[11] If we additionally budgeted time for a coaching expert to provide qualitative analysis,[12] this total comes to \$2,198. Based on this back-of-the-envelope calculation, cost-savings in the TeachSIM case would be 85%. Furthermore, we expect cost-savings will increase as automated transcription services improve. However, TeachSIM is a relatively simple 1:1 intervention. Recording and transcribing interventions that occur in noisy classroom settings will be more difficult, particularly as students' voices may be muffled, labeling many speakers may be onerous, and compliance with requests to continuously wear a microphone may be low. These are all potential complications which should be considered before undertaking any NLP analysis of transcripts.

Ultimately, a researcher's decision on whether to use semantic similarity depends on their context, research questions, and resources. A semantic similarity approach is most appropriate when the treatment is highly structured, the researcher does not need to discriminate between components of the theory of change, and resources are scarce. If, on the other hand, a treatment is not highly standardized or the researcher has the resources, they should use traditional methods of assessing fidelity. Or, if the study is too large to employ trained observers in every session, but the researcher has the resources to label a large enough subset, a classification approach may be most appropriate.

Unresolved Issues and Areas of Future Research

The semantic similarity measures for assessing treatment adherence and replicability proposed in this article are still in a nascent stage of development. Though we believe that the TeachSIM example provides a useful proof of concept for the potential value of the method, questions remain for future research. First, semantic similarity could benefit from a formal validation study showing the relationship between semantic similarity measures, other implementation measures, and outcomes targeted by interventions. In practice, however, we suspect that semantic similarity measures will require additional validation in each new context. To this end, we recommend that researchers undertake an informal validation study similar to what we performed in TeachSIM—asking a content expert, who is blinded to the semantic similarity scores, to identify examples of high- and low-adherence transcripts and examining the extent to which their judgment matches the distribution of the scores. Furthermore, while in this study we only sample a limited number of transcripts for qualitative review, this approach can be extended with a

larger sample of transcripts or with additional reviewers. For example, researchers might ask both intervention experts and intervention implementers to review transcripts in order to triangulate interpretations. Second, because insights from semantic similarity scores come from observing and comparing the distributions of scores, there are open questions about sample size requirements for appropriate interpretation of scores. Hopefully, future research can provide guidance on the number of transcripts required akin to examining results from power analyses for determining appropriate sample sizes in studies. In the meantime, we suggest that researchers incorporate additional manual inspection in the early stages of a program before many transcripts have been analyzed. Once the distribution seems stable (i.e., when adding additional transcripts does not dramatically change the shape of the distribution) and researchers feel they have an intuition for the meaning behind similarity scores, they may then use the scores to monitor adherence with more confidence moving forward. Finally, a key concern in any NLP application is algorithmic bias. Depending on the preprocessing techniques applied, semantic similarity methods may penalize language that reflects gendered or cultural differences. This is an area which is ripe for research, but, ultimately, the extent to which such variations in language reflect true nonadherence or bias will depend on the intervention and theory of change. For this reason, we recommend that researchers incorporate qualitative review of transcripts and take steps to ensure that they understand how the measure is applied in their context in order to detect bias when it occurs.

Conclusion

This article demonstrates how NLP methods can help address many of the logistical, methodological, and budgetary challenges of implementation research. We propose semantic similarity methods as a low-cost, scalable method for assessing intervention adherence and replicability for highly structured interventions. In particular, we illustrate two measures: the similarity between transcripts and a benchmark script as a measure of adherence and the similarity between transcripts within and across studies as a measure of intervention replicability. An important advantage of the method is that it can be adapted to a variety of implementation constructs across a broad array of intervention types and contexts. For example, researchers may adapt semantic similarity methods to measuring treatment-control contrast by comparing language heard by the treatment group with the language heard by the control group. Alternatively, researchers may measure treatment variation across treatment modalities by comparing online to in-person conversations. To this end, we hope that researchers will view this article as a jumping off point and will adapt our proposed approach to their particular circumstances and research questions.

Appendix A

A Selective Overview of Advanced Natural Language Processing (NLP) Techniques

Each of the methods described in the main body of the article are relatively straightforward to apply using common statistical programming languages including Python, R, and Stata. However, they do not represent the current state of the art in NLP. In this appendix, we provide a short, selective overview of more advanced NLP methods which researchers may consider for incorporating the shared meaning between words and for considering a word's context within the document.

Incorporating Shared Meaning Between Words. Like LSA, word embeddings aim to capture the semantic meaning of words. They work with the underlying assumption that "a word is characterized by the company it keeps" (Firth, 1957). To this end, word embeddings are vectors which have been optimized so that words that appear in similar contexts are mapped close to one another in vector space (Mikolov et al., 2013). A reliable word embedding model will assign related words like student and child with vector that close to one another in vector space. These methods have proven to be highly effective at representing meaning. However, in practice, applying word embeddings to calculating the similarity between documents is difficult. Word embeddings represent each word with a vector (commonly with a length of 1,000). Thus, each document is represented as a high-dimensional matrix. Applying cosine similarity to multiple matrices is not straightforward. To sidestep this problem, researchers often simply average the word embeddings for a document (reducing the word embeddings matrix to a vector; Crossley et al., 2019), thereby losing much of the contextual information provided by the word embeddings.

Deep Learning Approaches for Considering Context

All of the techniques discussed in the main body of the article are considered "bag-of-words" models because they assume that documents can be represented as an unordered set of words. Though this assumption may seem unrealistic, bag-of-words models have been shown to be effective in a variety of contexts, including information retrieval (retrieving the most relevant document given some search query; Manning et al., 2008), inferring the author of a document (Gentzkow et al., 2017), and inferring an author's psychological state (Tausczik & Pennebaker, 2010). Nonetheless, there are several new approaches to representing documents which take into account word order and document organization. For example, one particularly effective approach to preserving word order is to use convolutional neural networks

(CNNs). CNNs were designed for visual classification tasks (e.g., classifying a photo as a photo of a dog, or not) and work by filtering data into a series of increasingly complex patterns. Because they preserve special relationships (e.g., a pixel or word's location within a photo or document), there is built-in support for considering a word's context (Kim, 2014; LeCun & Bengio, 1995). However, CNNs were designed for classification tasks and are less commonly applied to semantic similarity. In practice, this means that researchers would need to adapt available programs and that they will find substantially fewer references for their task.

Appendix B

Example Coaching Script for the Behavior Management Scenario

We provide an example coaching script labeled with the five components of the treatment protocol: opening, positive feedback, constructive feedback, practice, and closure. The script represents an ideal version of each of these for the behavior scenario where the targeted skill is providing timely redirections. The behavior scenario has four of these scripts, one for each potential targeted skill. The feedback scenario has four additional scripts as well.

TABLE B1
Example Script Aligned With Fidelity Components

Component	Description	Script
Opening	The coach asks for the teacher candidate's (TC's) thoughts about how the first simulation went.	How are you feeling about that first simulation?
Positive feedback	The coach provides positive feedback on one specific element of the TC's first simulation. The coach elaborates on their positive feedback by describing why the component(s) they praised is/are important.	I was excited watching you because I saw you make a face when Ethan started humming. That is so important because it shows me that you already have the lens to recognize misbehavior as soon as it begins. You noticed every time a student misbehaved.
Growth area	The coach names a specific area for growth, gives a definition for this growth area and elaborates on what this growth area means and why it is important. The coach connects the discussion to a specific example from the TC's first simulation and asks the TC to identify a better response to the student. The coach reinforces the importance of the growth area by asking a question(s) that supports the TC in reflecting on the difference between a response that incorporates the area of growth and a response that does not.	To make your next simulation even stronger, I want you to focus on making your redirections more timely so that you can address the misbehavior right away. This prevents the misbehaviors from distracting other students and taking away from class time. For example, I noticed in your last simulation that you were hesitant to correct Ethan. Next time Ethan hums I want you to immediately redirect the behavior. For example, you could say: Ethan, voice off, hands together. Let's look at another example. When Ethan misbehaves how could you respond immediately to redirect the behavior? Exactly, that's great. You could also say please stop humming. What would a response that's not timely look like? Why is the first response better than ignoring the behavior?
Practice	The coach indicates that they want the TC to practice implementing their feedback by engaging in a role-play. The coach provides positive reinforcement for at least one specific thing that the TC did well during the role-play.	Now I want you to actually practice redirecting a student. I will pretend to be an off-task student. I want you to redirect my behavior immediately. Why don't you start by pretending to teach the lesson? [Humming] That was great. You addressed my behavior right away.
Closure	The coach closes the conversation with a reminder of what the TC should focus on for the next simulation. The coach closes the conversation in a way that provides positive encouragement to the TC.	For the next session, you could try to keep a few redirections in mind for some common misbehaviors like talking or making noises. That will help address the behavior right away, before it can distract other students, without you having to spend time thinking about what to say first. I'm so excited to see you redirect student behavior immediately in the next session!

Appendix C

Semantic Similarity Statistics by Study and Preprocessing Technique

In this appendix, we display descriptive statistics resulting from semantic similarity measures for each study using five different text-preprocessing techniques: no pre-preprocessing, stop word removal, term frequency–inverse document-frequency (tf-idf) weighting, lemmatization, and latent semantic analysis. The techniques are cumulative so that the final set of results uses all of the previous preprocessing methods. Each table demonstrates a consistent pattern. The highest similarity scores are produced without any text preprocessing. Removing stop words dramatically reduces similarity scores. This is expected as we are removing the most common terms from the documents. tf-idf further reduces similarity scores; tf-idf weighting gives a greater weight to less common terms. Lemmatization, on the other hand increases similarity scores as it increases the number of shared terms in two documents. Finally, latent semantic analysis again increases similarity scores, but this behavior is not as predictable as the previous techniques.

TABLE C1
Behavior Study 1 Semantic Similarity Statistics

Script similarity					
Mean	0.69	0.36	0.23	0.25	0.33
SD	[0.05]	[0.06]	[0.05]	[0.05]	[0.08]
Range	(0.55, 0.82)	(0.25, 0.52)	(0.13, 0.36)	(0.15, 0.38)	(0.19, 0.53)
Within-study similarity					
Mean	0.83	0.55	0.3	0.31	0.42
SD	[0.02]	[0.04]	[0.04]	[0.04]	[0.06]
Range	(0.76, 0.87)	(0.41, 0.63)	(0.2, 0.38)	(0.21, 0.39)	(0.26, 0.52)
Remove stop words		X	X	X	X
tf-idf			X	X	X
Lemmatization				X	X
LSA					X

Note. Script similarity scores (measuring intervention adherence) were estimated by calculating the average cosine similarity between each transcript and the appropriate benchmark script. A higher score indicates higher adherence to the script. Within-study similarity scores (measuring replicability) were estimated by calculating the average pairwise cosine similarity of each transcript within Behavior Study 1 to every other Behavior Study 1 transcript. tf-idf = term frequency–inverse document-frequency; LSA = latent semantic analysis.

TABLE C2
Behavior Study 2 Semantic Similarity Statistics

Script similarity					
Mean	0.74	0.39	0.26	0.28	0.38
SD	[0.05]	[0.08]	[0.06]	[0.06]	[0.08]
Range	(0.56, 0.82)	(0.24, 0.52)	(0.16, 0.37)	(0.17, 0.4)	(0.25, 0.56)
Within-study similarity					
Mean	0.84	0.52	0.3	0.31	0.42
SD	[0.02]	[0.04]	[0.03]	[0.03]	[0.05]
Range	(0.77, 0.87)	(0.43, 0.59)	(0.23, 0.36)	(0.25, 0.38)	(0.32, 0.52)
Remove stop words		X	X	X	X
tf-idf			X	X	X
Lemmatization				X	X
LSA					X

Note. Script similarity scores (measuring intervention adherence) were estimated by calculating the average cosine similarity between each transcript and the appropriate benchmark script. A higher score indicates higher adherence to the script. Within-study similarity scores (measuring replicability) were estimated by calculating the average pairwise cosine similarity of each transcript within Behavior Study 2 to every other Behavior Study 2 transcript. tf-idf = term frequency–inverse document-frequency; LSA = latent semantic analysis.

TABLE C3

Behavior Study 3 Semantic Similarity Statistics

Script similarity					
Mean	0.72	0.36	0.23	0.24	0.32
SD	[0.05]	[0.07]	[0.06]	[0.06]	[0.08]
Range	(0.59, 0.8)	(0.15, 0.53)	(0.09, 0.34)	(0.1, 0.36)	(0.15, 0.49)
Within-study similarity					
Mean	0.84	0.58	0.32	0.33	0.45
SD	[0.02]	[0.04]	[0.04]	[0.04]	[0.06]
Range	(0.79, 0.87)	(0.45, 0.65)	(0.24, 0.41)	(0.25, 0.42)	(0.32, 0.56)
Remove stop words		X	X	X	X
tf-idf			X	X	X
Lemmatization				X	X
LSA					X

Note. Script similarity scores (measuring intervention adherence) were estimated by calculating the average cosine similarity between each transcript and the appropriate benchmark script. A higher score indicates higher adherence to the script. Within-study similarity scores (measuring replicability) were estimated by calculating the average pairwise cosine similarity of each transcript within Behavior Study 3 to every other Behavior Study 3 transcript. tf-idf = term frequency–inverse document-frequency; LSA = latent semantic analysis.

TABLE C4

Feedback Study 1 Semantic Similarity Statistics

Script similarity					
Mean	0.63	0.3	0.16	0.18	0.25
SD	[0.05]	[0.07]	[0.06]	[0.06]	[0.09]
Range	(0.47, 0.75)	(0.14, 0.53)	(0.08, 0.34)	(0.08, 0.39)	(0.1, 0.54)
Within-study similarity					
Mean	0.79	0.46	0.23	0.25	0.33
SD	[0.02]	[0.04]	[0.03]	[0.03]	[0.05]
Range	(0.73, 0.83)	(0.34, 0.57)	(0.18, 0.31)	(0.19, 0.32)	(0.25, 0.44)
Remove stop words		X	X	X	X
tf-idf			X	X	X
Lemmatization				X	X
LSA					X

Note. Script similarity scores (measuring intervention adherence) were estimated by calculating the average cosine similarity between each transcript and the appropriate benchmark script. A higher score indicates higher adherence to the script. Within-study similarity scores (measuring replicability) were estimated by calculating the average pairwise cosine similarity of each transcript within Feedback Study 1 to every other Feedback Study 1 transcript. tf-idf = term frequency–inverse document-frequency; LSA = latent semantic analysis.

TABLE C5

Feedback Study 2 Semantic Similarity Statistics

Script similarity					
Mean	0.74	0.51	0.36	0.39	0.54
SD	[0.04]	[0.08]	[0.09]	[0.09]	[0.13]
Range	(0.62, 0.79)	(0.36, 0.68)	(0.16, 0.53)	(0.2, 0.56)	(0.24, 0.74)
Within-study similarity					
Mean	0.84	0.55	0.35	0.37	0.5
SD	[0.02]	[0.04]	[0.04]	[0.04]	[0.05]
Range	(0.78, 0.88)	(0.48, 0.61)	(0.27, 0.42)	(0.29, 0.44)	(0.38, 0.58)
Remove stop words		X	X	X	X
tf-idf			X	X	X
Lemmatization				X	X
LSA					X

Note. Script similarity scores (measuring intervention adherence) were estimated by calculating the average cosine similarity between each transcript and the appropriate benchmark script. A higher score indicates higher adherence to the script. Within-study similarity scores (measuring replicability) were estimated by calculating the average pairwise cosine similarity of each transcript within Feedback Study 2 to every other Feedback Study 2 transcript. tf-idf = term frequency–inverse document-frequency; LSA = latent semantic analysis.

Appendix D

Robustness of Adherence Scores to Number of LSA Dimensions

	50 LSA components	100 LSA components	200 LSA components
Behavior Study 1	0.33	0.4	0.27
Behavior Study 2	0.38	0.46	0.3
Behavior Study 3	0.32	0.38	0.27
Feedback Study 1	0.25	0.3	0.2
Feedback Study 2	0.54	0.61	0.45
Remove stop words	X	X	X
tf-idf Weighting	X	X	X
Stemming	X	X	X
LSA	X	X	X

Note. Adherence scores were estimated by calculating the cosine similarity between each transcript and the appropriate benchmark script. Shading indicates a higher ranking by average adherence score for each study where a darker shading indicates higher adherence. tf-idf = term frequency–inverse document-frequency; LSA = latent semantic analysis.

Acknowledgments

The research reported in this article was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B140026 and Grant #R305D190043 to the Rectors and Visitors of the University of Virginia as well as the National Academy of Education and the National Academy of Education/Spencer Dissertation Fellowship Program. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors wish to thank Julie Cohen, Brian Wright, members of the University of Virginia School of Data Science capstone team, and members of the University of Virginia TeachSIM team for their feedback on earlier versions of this article. All errors are those of the authors.

Notes

1. For example, in tests of synonym detection, Landauer and Dumais (1997) found that performance peaked with 300 dimensions when trained on a corpus of approximately 30,000 terms.

2. Semantic similarity methods may be implemented using a number of programming languages. In the TeachSIM application, we used Python's spaCy module for tokenization and lemmatization and sklearn for vectorization. Python's Natural Language Tool Kit also offers a number of helpful text analysis functions. If researchers are unfamiliar with Python, R offers many reasonable alternatives (including the quanteda, Text2Vec, and spacyr packages) and Stata offers a package (lsamantic), which calculates text similarity using LSA.

3. In the TeachSIM application, intervention sessions are short—just 5 minutes. We hypothesize that semantic similarity measures would become noisier with longer intervention sessions. In these cases, researchers could consider building in distinct break points in transcripts which correspond to sections of a treatment protocol. Then, researchers could calculate the semantic similarity of the transcript subsection to the corresponding section of the protocol.

4. So long as time tags and speaker tags are denoted consistently, these can be automatically removed using regular expressions. We can also use the speaker tags to remove the text of speakers which are not relevant to the research question.

5. If we were instead interested in using semantic similarity methods to explore a construct like participant responsiveness, we might have instead chosen to exclude coach text and focus our analysis on participant speech.

6. A total of 504 words on average in Feedback Study 1 (SD = 136), 709 words in Feedback Study 2 (SD = 146), 740 in Behavior Study 1 (SD = 132), 768 in Behavior Study 2 (SD = 102), and 769 in Behavior Study 3 (SD = 125).

7. Results are robust to the inclusion of 50, 100, and 200 dimensions. See Appendix D.

8. We are not particularly concerned that our results are not robust to the inclusion of stop words. Differences in rankings for this naïve approach are not particularly informative. For example, one of the reasons Behavior Study 2 is more similar to its ideal script than Feedback Study 1 is that Behavior Study 2 and the ideal behavior script have the same most common words: to, you, and that. On the other hand, the most common word in Feedback Study 1 transcripts is "the" while the most common word in the feedback script is "to." These differences are unlikely to be meaningful.

9. Here, we define an off-topic conversation as any conversation that does not correspond to one of the five components of the coaching protocol. Given the 5-minute time limit for the coaching conversation, these off-topic conversations likely result in lower adherence to the coaching protocol to the extent that they take up time that would otherwise be spent implementing the components of the protocol in a higher quality or more thorough way. However, in cases where this does not occur, such off-topic conversations may beneficially support rapport between the coach and teacher candidate without not resulting in lower adherence. As we note above, the document similarity method cannot make these kinds of evaluative judgments to determine whether an off-topic conversation was beneficial or disruptive, which is why we suggest that researchers employ limited qualitative analysis to inform the interpretation of adherence scores and/or next steps to support ongoing adherence.

10. This is estimated at the University's undergraduate hourly rate, includes 5 hours of training per coder as well as weekly norming meetings, and allows for 15% of transcripts to be double coded.

11. In the TeachSIM context, obtaining professional transcriptions cost $0.83 a minute.

12. We have budgeted 15 hours at an hourly rate of $35.

References

Anglin, K. L. (2019). Gather-narrow-extract: A framework for studying local policy variation using web-scraping and natural language processing. *Journal of Research on Educational Effectiveness*, *12*(4), 685–706. https://doi.org/10.1080/19345747.2019.1654576

Century, J., & Cassata, A. (2016). Implementation research: Finding common ground on what, how, why, where, and who. *Review of Research in Education*, *40*(1), 169–215. https://doi.org/10.3102/0091732x16665332

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis*, *42*(2), 208–231. https://doi.org/10.3102/0162373720906217

Crossley, S. A., Kyle, K. & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*(1), 14–27. https://doi.org/10.3758/s13428-018-1142-4

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*(1), 23–45. https://doi.org/10.1016/S0272-7358(97)00043-3

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

Dumas, J. E., Lynch, A. M., Laughlin, J. E., Smith, E. P., & Prinz, R. J. (2001). Promoting intervention fidelity. *American Journal of Preventative Medicine*, *20*(3), 38–47. https://doi.org/10.1016/S0749-3797(00)00272-5

Durlak, J. A. (2015). Studying program implementation is not easy but it is essential. *Prevention Science*, *16*(8), 1123–1127. https://doi.org/10.1007/s11121-015-0606-3

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, *18*(2), 237–256. https://doi.org/10.1093/her/18.2.237

Fesler, L. (2020) *Opening the black box of remote college counseling using text-as-data* (CEPA Working Paper). Stanford Center for Education Policy Analysis. http://cepa.stanford.edu/wp20-03

Firth, J. R. (2018). A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis*, *1957*, 1–32.

Ganz, J. B., Kaylor, M., Bourgeois, B., & Hadden, K. (2008). The impact of social scripts and visual cues on verbal communication in three children with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities*, *23*(2), 79–94. https://doi.org/10.1177/1088357607311447

Gentzkow, M., Kelly, B.T., & Taddy, M. (2017). *Text as data* (NBER Working Papers). Cambridge, MA. https://www.nber.org/papers/w23276

Goldstein, H. (2002). Communication intervention for children with autism: A review of treatment efficacy. *Journal of Autism and Developmental Disorders*, *32*(5), 373–396. https://doi.org/10.1023/a:1020589821992

Gresham, F. M. (2017). Features of fidelity in schools and classrooms: Constructs and measurement. In G. Roberts, S. Vaughn, S. N. Beretvas, & V. Wong (Eds.), *Treatment fidelity in studies of educational intervention* (pp. 22–38). Routledge.

Horner, R. H., & Sugai, G. (2015). School-wide PBIS: An example of applied behavior analysis implemented at a scale of social importance. *Behavior Analysis in Practice*, *8*(1), 80–85. https://doi.org/10.1007/s40617-015-0045-4

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*(7), 451–464.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1746–51. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1181

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295x.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, *3361*(10), 1995.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. ArXiv Preprint. arXiv. http://ronan.collobert.com/senna/

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *Journal of Behavioral Health Services and Research*, *39*(4), 374–396. https://doi.org/10.1007/s11414-012-9295-x

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, *78*(1), 33–84. https://doi.org/10.3102/0034654307313793

Reardon, S. F., & Stuart, E. A. (2019). Education research in a new data environment: Special issue introduction. *Journal of Research on Educational Effectiveness*, *12*(4), 567–569. https://doi.org/10.1080/19345747.2019.1685339

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*,

24(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Sanetti, L. M. H., & Kratochwill, T. R. (2009). Treatment integrity assessment in the schools: An evaluation of the treatment integrity planning protocol. *School Psychology Quarterly*, *24*(1), 24–35. https://doi.org/10.1037/a0015431

Steiner, P. M., Wong, V. C., & Anglin, K. L. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift Für Psychologie/Journal of Psychology*, *227*(4), 280–292. https://doi.org/10.1027/2151-2604/a000385

Stevenson, C. L., Krantz, P. J., & McClannahan, L. E. (2000). Social interaction skills for children with autism: A script-fading procedure for nonreaders. *Behavioral Interventions: Theory & Practice in Residential & Community-Based Clinical Programs*, *15*(1), 1–20. https://doi.org/10.1002/(SICI)1099-078X(200001/03)15:1%3C1::AID-BIN41%3E3.0.CO;2-V

Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, *88*(4), 479–507. https://doi.org/10.3102/0034654317751919

Sun, M., Liu, J., Zhu, J., & LeClair, Z. (2019). Using a text-as-data approach to understand reform processes: A deep exploration of school improvement strategies. *Educational Evaluation and Policy Analysis*, *41*(4), 510–536. https://doi.org/10.3102/0162373719869318

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Wong, V., Anglin, K., & Steiner, P. (2020). *Design-based approaches to systematic replication studies* (No. 74; EdPolicyWorks Working Paper Series). https://education.virginia.edu/sites/default/files/uploads/epw/74_Design-Based_Approaches_to_Systematic_Conceptual_Replication_Studies.pdf

Authors

KYLIE L. ANGLIN is an assistant professor in research methods, measurement, and evaluation in the Neag School of Education at the University of Connecticut. Her research develops methods for analyzing program implementation in field settings using data science techniques, as well as methods for improving the causal validity and replicability of impact estimates.

VIVIAN C. WONG is an associate professor in research, statistics, and evaluation in the School of Education and Human Development at the University of Virginia. Her research focuses on evaluating interventions in early childhood and K–12 systems and in improving the design, implementation, and analysis of replications and randomized and quasi-experiments.

ARIELLE BOGUSLAV is a PhD student in education policy in the School of Education and Human Development at the University of Virginia. Her research interests primarily relate to teacher professional development and the application of behavioral insights to policy implementation.