Working Paper:

Using Semantic Similarity to Assess Adherence and Replicability of Intervention Delivery

Kylie L. Anglin , Vivian C. Wong, and Arielle Boguslav

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 paper proposes a set of natural language processing (NLP) techniques called semantic similarity as an innovative, low-cost, and scalable method of measuring implementation constructs. Semantic similarity methods are a semi-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 paper 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.

University of Virginia

Updated April 2021

EdPolicyWorks
University of Virginia
PO Box 400879
Charlottesville, VA 22904

Acknowledgements: The research reported here 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 paper. All errors are those of the authors.

Using Semantic Similarity to Assess Adherence and Replicability of Intervention Delivery

Kylie L. Anglin, Vivian C. Wong, and Arielle Boguslav

University of Virginia

April 2021

Abstract

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 paper proposes a set of natural language processing (NLP) techniques called semantic similarity as an innovative, low-cost, and scalable method of measuring implementation constructs. Semantic similarity methods are a semi-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 paper 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.

Introduction

Experimental and quasi-experimental evaluations in educational settings have dramatically increased over the last two decades. This development has improved our ability to determine "what works" in improving student outcomes. Yet, recent analyses have shown that the majority of the large-scale randomized evaluations of educational interventions do not identify significant or substantial effects (Baron, 2013; Schneider, 2018). In these cases, the next question often concerns fidelity: *was the intervention implemented as it was intended?* This question has led to increased calls for implementation research to be included in efficacy studies with strong causal claims, like randomized control trials (RCTs; see the Standards for Excellence in Education Research, Institute of Education Sciences, 2020a). When efficacy trials fail to include information on how an intervention was implemented, readers are forced to assume that the intervention was delivered uniformly as designed (Dobson & Cook, 1980). In practice, however, interventions often deviate from the original program design, particularly when they are delivered at-scale and outside of controlled laboratory environments. Implementation research not only provides needed information on deviations in program delivery that influence the conclusions drawn from studies with null results, but, when monitored in real-time, can also meaningfully improve program quality (Durlak & DuPre, 2008).

Unfortunately, implementation researchers face intense logistical, methodological, and budgetary constraints in their efforts to assess intervention fidelity. Reviews of education research show that implementation information is frequently missing from program evaluations (Dusenbury et al., 2003; O'Donnell, 2008). This may be due to a dearth of practical guidance about the best ways to assess intervention fidelity in field settings (Roberts, 2017), or because of the additional resources that are needed to collect implementation data. Traditional approaches to implementation research require the development and validation of reliable fidelity measures for each new intervention. Unfortunately, it can be difficult to produce new fidelity measures with appropriate measurement properties for each new evaluation context, especially in cases where the intervention is complex and includes multiple components. To further complicate the issue, there have been limited methodological developments in designing treatment fidelity measures that have acceptable psychometric properties (Gresham, 2017; Sanetti & Kratochwill, 2009), and even fewer developments in designing measures for assessing treatment consistency in replication and scale-up studies. Finally, the process for hiring, training, and employing observers is time-consuming and expensive and may be altogether infeasible when sessions occur at different times, in different settings, and with multiple research teams.

Given the above challenges, the education research community would benefit from low-cost and scalable methods for assessing treatment fidelity in field settings. In this paper, we propose the use of a natural language processing (NLP) technique, semantic similarity, to describe how closely transcripts from intervention sessions delivered in field settings adhere to a standardized treatment protocol, and how consistently the protocol is replicated across sessions. At its core, semantic similarity quantifies the distance between two texts – such as transcripts from intervention sessions – based on the likeness of semantic content. To produce a measure of *intervention adherence*, we propose a semantic similarity method that evaluates the similarity between intervention transcripts and a scripted intervention protocol. To produce a measure of *intervention replicability*, we propose a method that quantifies the similarity of intervention transcripts to one another. The replicability measure reflects the extent to which the intervention was delivered consistently across potential sources of variation (intervention sessions, participants, interventionists, sites, or studies). A key strength of the semantic similarity approach is that it can be adapted to a number of implementation contexts simply by changing the documents to which transcripts are being compared.

The method is best applied in cases where the intervention is highly structured and delivered through verbal interactions with participants, and where transcripts of the intervention sessions are available. These sorts of structured, interactive intervention sessions are common in some domains of education, including in reading and mathematics instruction for struggling learners, in special education, and in behavioral education. These interventions often employ a pedagogical approach called "direct instruction," which requires explicit and unambiguous delivery of instructions, explanations, and practice models to support student learning. Given that direct instruction is often difficult to implement in field settings – where small variations in delivery can result in erroneous interpretations from students – interventions that employ these techniques commonly provide highly structured guidance for implementers on both the wording and sequence of learning activities (Adams & Carnine, 2003; Stockard et al., 2018). We propose semantic similarity techniques as a method of measuring implementation for these sorts of interventions where an interventionist's choice of language is considered important to the intervention's theory of change.

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 paper serves as a primer on NLP methods for semantic similarity, generally, and demonstrates how they can be applied for exploring implementation in education settings. The paper also provides a proof of concept for using semantic similarity

measures to quantify intervention adherence and replicability in low-cost and scalable ways. To this end, we apply the approach to a series of RCTs in teacher education that evaluate the impact of TeachSim – a structured coaching protocol – on preservice teachers' pedagogical performance in simulated classroom environments (Cohen et al., 2020). In these RCTs, treatment group teachers received a five-minute coaching session where the coach was expected to follow a structured conversation protocol. Through this application, we demonstrate that semantic similarity methods may be applied in evaluation contexts where the intervention protocol is standardized but not invariant – that is, the interventionist delivers instruction or feedback in a highly structured, explicit, and sequential manner, while having the flexibility to select a skill that is most appropriate for the learner.

The remainder of this paper is structured as follows. First, we provide a short overview of implementation constructs, summarizing how these constructs are traditionally measured and providing examples of recent literature using NLP for their measurement. Second, we discuss semantic similarity measures and how they may be used to assess implementation constructs like intervention adherence and replicability of intervention delivery. Third, we detail the NLP techniques that researchers may use to calculate semantic similarity. Fourth, we demonstrate the method's potential in an evaluation of coaching protocol for preservice teachers, discussing approaches to interpreting semantic similarity scores and demonstrating the face validity of the method through a qualitative analysis of transcripts. Finally, we conclude with a discussion of emerging questions for using semantic similarity scores in evaluation contexts, as well as advantages and disadvantages of a semantic similarity approach to measuring implementation constructs.

Constructs and Measures in Implementation Research

Understanding intervention delivery is critical in evaluation settings for multiple reasons. First, if participants do not receive the full intervention, receive an unexpected intervention, or receive highly variable intervention components, researchers need to know this in order to appropriately interpret the results of a study (Fixsen et al., 2005; Rossi et al., 1985). Second, measures of implementation may be useful for explaining variations in outcomes within a single study (Schochet et al., 2014), and for explaining variations in effects across multiple studies (Steiner et al., 2019; V. Wong & Steiner, 2018). Finally, ongoing monitoring of intervention implementation in the field provides opportunities for researchers to ensure that the intervention is delivered according to the standardized protocol, or to provide additional supports to ensure appropriate implementation (Durlak & DuPre, 2008; Fixsen et al., 2005).

Implementation constructs may broadly be grouped into two categories: those for which there is some concept of ideal implementation and those for which the researcher is agnostic to how the intervention is implemented but nonetheless wishes to measure variation in delivery. The former category tackles questions of treatment fidelity, defined as the degree to which a treatment is implemented as intended by the designers (Dumas et al., 2001; Nelson et al., 2012; O'Donnell, 2008). In their review of prevention literature, Dane and Schneider identify five primary conceptualizations of fidelity: dosage, adherence to the program design, quality of program delivery, participant responsiveness, and program differentiation (1998). Of these, the most commonly measured construct by far is *dosage*: the amount of the intervention that is delivered. Other common constructs include *program adherence*: the extent to which implementers deliver the components of a treatment protocol; *program quality*: how well the treatment protocol is delivered; and *participant responsiveness*: the degree to which participants engage with the treatment. There is also increasing attention being paid to *program differentiation* defined as the difference between the program and business-as-usual or in RCTs, between the treatment and control conditions (Hulleman & Cordray, 2009; Institute of Education Sciences, 2020b). When researchers do not have a pre-specified concept of ideal implementation, they may wish to measure *program variation* or *adaptation* rather than fidelity. For example, researchers may be largely agnostic to implementation styles if they believe that interventions should be adapted to individual contexts (Dane & Schneider, 1998). Even when adaptation is viewed as a positive development, researchers still need to document and measure variation to better understand the nature of the intervention and interpret effects.

Measures of implementation vary depending on researchers' conceptualization of the construct and the resources available to them. On one end of the spectrum, researchers may use easy-to-collect data like attendance records, counts and lengths of intervention sessions, implementer self-reports, or logs from an online platform (Dane & Schneider, 1998; Dusenbury et al., 2003; Hulleman & Cordray, 2009). Other the other end of the spectrum, researchers may hire trained observers to rate each intervention session against a rubric or conduct qualitative interviews of interventionists and participants (Hulleman & Cordray, 2009; Marsh et al., 2017). In the following section, we propose semantic similarity as a method of producing measures of implementation that are less expensive than hiring trained observers or conducting interviews, but that can also provide information beyond measuring dosage from attendance records or counts of intervention sessions.

NLP Measures of Implementation Constructs

This paper sits within the burgeoning literature applying NLP techniques to education data (Reardon & Stuart, 2019) and more specifically within a smaller body of literature using NLP to understand program implementation. To our knowledge, researchers have so far applied three NLP techniques to measure implementation in education: dictionaries, topic modeling and text classification. In a dictionary-based approach to NLP, the researcher parses each document searching for instances of key terms. For example, in a study describing district responses to deregulation under the Texas District of Innovation statute, Anglin documented variation in regulatory exemptions by searching District of Innovation plans for regulatory statutes of a specified format (Anglin, 2019). This is an example of using NLP to document program variation (here, variation in regulatory exemptions claimed). Similarly, Sun, Liu, Zhu, and LeClair use topic modeling, a method of automatically extracting patterns of semantic meaning (topics) from text, to document variation in reform strategies found in school improvement plans (Sun et al., 2019). Topic modeling is an automated approach to the kind of coding that qualitative researchers undertake – grouping texts into categories of shared meaning – which makes it well suited to documenting and understanding treatment variation.

In cases where researchers have some a priori notion of better and worse implementation and wish to measure fidelity, automated classification can be a useful approach. In this approach, researchers label a subset of data by hand and train an algorithm to recognize the text features that correspond to those labels. For example, in a study of text-message based college counseling, Fesler trained a classifier to identify productive engagement between a college counselor and text-message recipient. This study could be characterized as an application of NLP to measuring the implementation construct of participant responsiveness. Similarly, Kelly, Olney, Donnelly, Nystrand, and D'Mello used an automated classifier to identify authentic questioning by teachers, an example of using NLP to measure quality of implementation (Kelly et al., 2018). Classification techniques take advantage of the highly scalable nature of NLP; once the algorithm has been trained, it may be applied to new treatment sessions at a negligible additional cost. However, the classification approach has substantial start-up expenses as researchers need to hand label documents. In the previous examples, Fesler hand-labeled 551 interactions while Kelly and team hand labeled 451 transcripts. This paper proposes a comparatively lower-cost solution using semantic similarity which does not require the hand labeling of documents.

Semantic Similarity as a Measure of Intervention Adherence and Replicability

Semantic similarity (also termed "document similarity" in information retrieval) is an umbrella term for a suite of NLP tools used to quantify the similarity of two or more texts. The intuition behind semantic similarity methods is simple – texts can be represented by their vocabulary and compared to one another by the relatively frequency with which they use a set of words or phrases. For now, we introduce the simplest approach to semantic similarity using word frequencies, though in the next section we discuss alternative representations of texts.

We'll begin by defining a few terms within the NLP context: a *document* is a single text of interest. A *corpus* is the full set of documents a researcher is interested in comparing. From the corpus, a researcher creates a *document-term matrix* where each row corresponds to a document $(i = 1, \ldots, N)$ and each column corresponds to a word in the corpus. Then, each document is represented by a vector $W_i = (W_{i1}, W_{i2}, \ldots W_{im})$, 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. The process of separating a document into a set of units (here, words) is referred in the NLP literature as *tokenization* while the process of creating a document-term matrix is referred to as *vectorization*.

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

$$sim(d_1, d_2) = \frac{\overrightarrow{V_1}(d_1) \cdot \overrightarrow{V_2}(d_2)}{|\overrightarrow{V_1}(d_1)||\overrightarrow{V_2}(d_2)|},$$

The numerator here is the dot product (sum of products) of the two document vectors: in other words, the sum of the product of the two documents' word frequencies in each column. For example, we would multiply the frequency of the first word in the first document by the frequency of the first word in the second document and add this to the product of the word frequencies for the second word, and so on. The denominator is the product of the magnitude of the two vectors (here, the number of words in each document). This normalizes 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.

Applications of Semantic Similarity in Education Settings

The canonical application of semantic similarity is in plagiarism detection; teachers and academics need some method of detecting when a writer has made extensive use of someone else's work, even when the plagiarizer has made inconsequential changes to the vocabulary or word order to avoid detection. Given two documents, a reader could likely identify whether the texts are suspiciously similar, but the problem quickly becomes overwhelming when a reader needs to compare one document to a large corpus of potential source documents which may have been plagiarized. For this reason, computer scientists have developed semantic similarity methods which can automatically detect plagiarism. Using these measures, teachers can identify which document from an arbitrarily large corpus is most similar to a student essay, and whether some essays are more likely than others to have been plagiarized.

The challenge of assessing intervention adherence and replicability in interactive standardized interventions is similar to the problem of plagiarism, albeit with fewer moral implications. As in the plagiarism case, researchers want to quantify the similarity of two or more documents (here, transcripts) and need some method of detecting derivative text (here, speech) *even if there are lexicographical differences in language that do not change the semantic meaning of the words.* This approach has most potential value for assessing intervention fidelity when the treatment protocol is delivered verbally and is highly structured or scripted.

In education and clinical settings, such protocols are common in interventions that include direct or explicit instruction. Direct instruction is characterized as a teacher-centered approach, often with structured or scripted demonstrations of step-by-step routines and practice opportunities and consistent feedback for efficient delivery of instructional content (Adams & Carnine, 2003). The approach has been shown to have potential value in supporting struggling readers and students with disabilities, and in behavior education. For example, two prominent reading curricula, Open Court and Success for All, include scripted lesson plans with sequenced learning activities for teachers to deliver reading instruction (Borman et al., 2007, 2008). Similarly, in *Response to Intervention* programs for struggling readers, teachers are asked to provide explicit, scripted instruction to small groups of students, with increasing intensity of intervention services as students' needs increase (Fuchs & Fuchs, 2006). In special education, students with autism spectrum disorder (ASD) are often taught "social scripts" for improving language skills and peer interactions in academic and social settings (Ganz et al., 2008; Goldstein, 2002; Stevenson et al., 2000). Finally, in comprehensive school-wide behavioral interventions such as Positive Behavioral Interventions and Supports (PBIS), teachers

learn highly-structured and consistent approaches for interrupting, correcting, and redirecting students' off-task, disruptive behaviors (Horner & Sugai, 2015). In each of these cases, teachers or interventionalists are provided a standardized protocol or script that they are expected to deliver; fidelity may be evaluated by examining how closely the language in the intervention sessions resembles or adheres to a scripted protocol.

Intervention Adherence

With semantic similarity, adherence scores can be determined by examining the cosine similarity of session transcripts and a benchmark script. That is, for each transcript, the researcher creates a document-term matrix from the full set of intervention transcripts and the benchmark script of interest. While the benchmark script itself depends on the nature of the intervention protocol, in general, it should include all components of the intervention protocol with scripted 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 A[1].

Then, for a given transcript of an intervention session, document d_i, and a benchmark script, s, script similarity is determined by the following:

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

where $sim(d_i, s)$ is the cosine similarity of the two documents ranging from 0 (no shared terms) to 1 (identical relative word frequencies). From there, the researcher can determine which intervention sessions are closest to the benchmark and which intervention sessions deviate more substantially by exploring the distribution of script similarity measures. The researcher can also calculate the average script similarity for a study, site or interventionist to compare the relative intervention adherence according to a benchmark script.

Intervention Replicability

Beyond intervention adherence, researchers also commonly want to understand how *consistently* an intervention is replicated across participants, interventionists, sites, or studies. Within a single study, consistency can be considered an important counterpart to adherence; Dumas and colleagues argue that interventions satisfy adherence requirements if and only if the intervention is

[1] In the TeachSIM application, intervention sessions are short – just five 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 sub-section to the corresponding section of the protocol.

delivered in such a way that is true to the theory of change *and* if it is delivered in a "comparable manner to all participants" (Dumas et al., 2001, p. 38). Moreover, when an intervention is replicated across sites or studies, conclusions from these replications often rely on the assumption that the treatment and control conditions are "identical in both studies, that is, there is no (unobserved) variation in the implementation of the treatment-control contrast across studies (Steiner et al., 2019, p. 283)". With semantic similarity, a researcher can measure the replicability (consistency) of intervention delivery by calculating the similarity of intervention transcripts to one-another.

To calculate the replicability score within a single study, the researcher first creates a document-term matrix for the full set of documents which will be compared. Then, the researcher calculates a pairwise similarity measure where each transcript in a study is compared to 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 one 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:

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

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

To calculate intervention replicability across two or more studies, a researcher creates a document-term matrix of every transcript across all studies of interest. Consider two studies, Study 1 and Study 2, where Study 1 has *n* documents and Study 2 has *m* documents. Then, the similarity of Study 1's document *j* to Study 2 is calculated by comparing document *j* to every document in Study 2:

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

and the average similarity of Study 1 and Study 2 is calculated:

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

Similar to the adherence measure described above, this method yields a replicability score that ranges between 0 and 1, where 1 indicates perfect replicability in transcripts and 0 indicates no semantic overlap across transcripts.

Thus far, this is the only measure we know of that quantifies replicability of intervention delivery based on the similarity of semantic content. Currently, most approaches to assessing intervention replicability involve comparing measures of treatment fidelity across participants, sites, or studies. However, our replicability score provides a direct quantitative measure of how consistently intervention sessions are delivered across those delivering the treatment, whether or not the intervention is delivered with high adherence. The measure may be especially useful in cases where intervention adherence is low, but the researcher still wants to know whether sessions were delivered consistently. Understanding both dimensions of intervention fidelity – adherence and replicability – provides researchers with important insights for understanding how the intervention was actually delivered, as well as for developing appropriate implementation supports.

Uses of Semantic Similarity Methods for Monitoring Field Evaluations

An advantage of semantic similarity approaches is that once intervention sessions have been transcribed, the adherence score can be calculated automatically. This means that in evaluation studies where transcript data are available and researchers are still in the field, the method may be used to identify sessions that stray from the benchmark and if needed, provide interventionists with additional training. Further, the measure is substantially less costly than hiring trained observers to rate each session according to an adherence rubric (the most common approach to assessing intervention fidelity).

Importantly, the method is not meant to replace all types of implementation research and is limited in its ability to evaluate other types of fidelity 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. Nor does the method evaluate the quality of the delivery, with the exception of semantic deviations from the benchmark. However, the method can save time and resources by flagging possible deviants from an intervention protocol in field settings for further human examination. In this way, script similarity for assessing intervention adherence may best be understood as an efficient and scalable, but narrow, measure of adherence to a standardized benchmark. In many field settings – where it is often impossible to observe and monitor intervention fidelity at all – even narrow but feasible measures of fidelity may provide researchers with an essential tool for supporting implementation.

NLP Techniques for Semantic Similarity

In this section, we provide an overview of NLP techniques beyond word frequencies which can be used to calculate semantic similarity between documents. Each of these techniques relates to decisions regarding the columns and values in a given document-term matrix. The following techniques are similar to the types of data processing that researchers do with non-text data: selecting which variables to include in a model, how to weight them, and whether some variables should be aggregated and transformed. In NLP contexts, the variables we are processing are the words in the corpus of documents. Note that the technical choices that a researcher makes in representing their corpus can have a substantial impact on their study's adherence and replicability scores. In practice, we recommend that researchers employ several techniques and test the robustness of their findings to technical decisions.

Prioritizing the Words that Matter

Document-term matrices quickly grow to very large dimensions as there is a column for every unique word in the document corpus. 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 may add very little meaning: words like *a*, *an*, *the*, and *to*. These words are commonly 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 pre-defined lists of stop words. 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 common weighting technique that researchers apply is frequency-inverse document-frequency (tf-idf) which assigns weights based on a word's relative frequency in the full corpus of documents (here, transcripts and scripts). Formally, tf-idf weights are determined by the following formula:

$$tfidf_{t,d} = tf_{t,d} * \log \frac{N}{df_t}.$$

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 any pre-processing, all words in a document-term matrix are treated as wholly distinct from one another. This is particularly 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 treating all derivatives of a word as a single entity using a method termed lemmatization. Lemmatization reduces each word to its root form – for example, *teach*, *teacher*, *teachers*, and *teaches* would all be represented by the root word, *teach*.

Even after lemmatizing, we still fail to capture the substitutability and similarity of words in a given context. To address this, researchers can incorporate Latent Semantic Analysis (LSA). LSA works under the assumption that the contexts in which a word does and does not appear is an appropriate method of determining the similarity of meanings of words to each other (Landauer et al., 1998). Similar to factor analysis, LSA is based on singular value decomposition and reduces the terms in a document-term matrix to a set of underlying factors which may be thought of as abstract 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; for example, in tests of synonym detection, Landauer and Dumais found that performance peaked with 300 dimensions when trained on a corpus of approximately 30,000 terms (Landauer & Dumais, 1997).

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 bigram 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 advanced programming skills, see Appendix B.

[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 (NLTK) 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.

The Impact of Pre-Processing 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 pre-processing 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 to 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, pre-processing 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 the cutoff after stop words have been removed. A single semantic similarity score on its own carries very little meaning. It is only through comparisons across different modeling approaches in our semantic similarity analysis that we gain insight. We recommend comparing several modeling approaches to interpret intervention adherence and replicability scores, which we will demonstrate below in the applied example below.

An Application to TeachSIM

In this section, we apply our proposed measures of intervention adherence and replicability to an experimental evaluation of the efficacy of a coaching protocol for improving teacher candidates' pedagogical skills. In the TeachSIM context, teacher candidates practice an instructional task for five minutes with student avatars in a mixed-reality simulated classroom environment. The task involved either "leading a text-based discussion" or "managing off-task student behaviors." Treated teachers then participate in a five-minute coaching conversation with a master educator designed to improve their pedagogical performance. During these sessions, coaches could choose one of four structured protocols depending on the targeted skill of the teacher candidate and the

instructional task. In coaching conversations focused on "leading a discussion," the four targeted skills for teachers included: probing for textual evidence, scaffolding student understanding, providing descriptive feedback, or probing for a warrant. In conversations focused on "managing off-task behaviors," 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. There were 14 coaches across the five studies, with four to five coaches per study and a turnover rate of approximately two coaches per study. In three studies, coaching conversations focused on improving teacher candidates' responses to off-task student behavior (Behavior Studies 1, 2, and 3). In the other two studies, coaching conversations focused on improving the quality of instructional feedback that teacher candidates provide to support students' understanding of a text (Feedback Studies 1 and 2). Feedback Study 1 was the first study conducted and was used as a pilot to inform the later development of the coaching protocol and student avatar script. 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. The coaching expert defined a structured coaching protocol that included five components where coaches: 1) ask the candidate to assess their own performance; 2) affirm an observed effective teaching practice; 3) identify 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 target 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 A shows an example script and how it aligns with the treatment protocol.

Note that these scripts demonstrate the structured nature of the intervention (as defined by the coaching protocol), but also that the intervention is not invariant. For example, the protocol allows for coaches to choose which teaching skill they'd like to target and which teaching practices

they would like to affirm. Further, the protocol is structured so that we would expect there to be patterns in language usage even in cases where the coach does not follow the protocol language verbatim. In fact, coaches were explicitly told that they did not need to use the protocol language verbatim.

Transcripts

Coaching sessions were video-taped and transcribed using either a professional transcription service or undergraduate research assistants. 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*)[3]. 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.[4]

Methods

Pre-processing. Before applying any of the NLP techniques discussed earlier in this paper, 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, impersonations, 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. The first matrix includes all of the terms in the corpus. In the second matrix, we excluded stop words from a popular pre-specified list (Python's Natural Language Toolkit – NLTK) and supplemented with a set of common pause fillers and vocal ticks like "uh" and "um". In the third matrix, we lemmatized the words, replacing all word derivatives with a single stem. In the fourth matrix, we weighted each term using tf-idf weighting.

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

Finally, in our fifth matrix, we performed LSA on a document-term matrix with stop word removal and tf-idf weighting and kept the 100 most common concepts.

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 to transcripts within the same study (for example, 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 to transcripts from other studies, we consider the score an *across-study replicability* measure.

TeachSIM Results

In this section, we demonstrate how semantic similarity methods may be used to provide descriptive measures of intervention adherence and replicability for single studies, and for results from multiple studies. We present measures of adherence and replication for each of the five studies and discuss how these results may be interpreted. Though we estimate five sets of adherence and replication scores using the approaches listed above, for most of the results, we limit our discussion to one, relatively simple, method of text processing for ease of interpretation: removing stop words and applying tf-idf weighting. However, once stop words have been removed, results are generally robust across methods indicating the usefulness of even simple NLP techniques. For more details on the results produced by each text processing method, as well as a narrative description of how scores change with each technique, we point readers to Appendix C.

Intervention Adherence

Table 1 provides an example of how adherence scores might be included in summary tables in an evaluation study report alongside other study 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 allows readers to conclude that they should pay close attention to Feedback Study 2 if they are interested in a relatively higher-adherence context.

Table 1 shows adherence scores from one set of analytic techniques, but, as discussed above, semantic similarity scores are sensitive to analytic decisions. Therefore, a table describing the

sensitivity of results to different specifications is useful. Table 2 shows the average adherence score for each study across each of the five pre-processing approaches. Within each pre-processing approach, we also rank the studies from lowest to highest adherence with shading; darker shading indicates higher adherence while the lowest-adherent study has no shading. Results in our case are largely robust; study ranks are relatively stable no matter the pre-processing techniques employed. Across techniques, Feedback Study 2 has the highest average adherence while transcripts from our pilot Feedback Study 1, have the lowest average adherence as indicated by four out of five techniques[5].

In Figures 1 and 2, we demonstrate how adherence scores can be used for monitoring treatment adherence. Figure 1 provides an example of how researchers might use visualization to identify abnormal transcripts that fall outside of the distribution of adherence scores. Here, we've 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 informing ongoing training. Here, we disaggregate Feedback Study 2, a study with a relatively wide distribution of adherence scores, by coach. The figure demonstrates 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*? More broadly, *how should semantic similarity scores be interpreted within a particular context*? Semantic similarity scores provide a quick and scalable way to code transcripts based on their similarity to a benchmark scripted protocol, but these scores are most useful with a "human in the loop." In particular, semantic similarity scores require some interpretation by experts with subject-matter knowledge of the intervention. To answer questions of interpretation within the TeachSIM context, we took two

[5] 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.

approaches which we recommend that researchers apply in their own contexts: an informal validation effort and qualitative analysis.

First, we asked a coaching expert who was blinded to script similarity results to pull three examples of ideal implementation of the behavior study protocol 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 transcripts with inadequate implementation of the protocol are well below the median script similarity 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 in order to qualitatively evaluate whether high-adherence scores are high enough and if low-adherence scores are below acceptable levels. 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. Qualitative analysis reveals that the lowest-adherence transcripts commonly include off-topic and unclear conversations and are often missing one or more of the 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 a strength, identify 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 about the simulator and does not clearly identify or explain a strength:

> *So, what's interesting about this is that even though it seems so odd, it actually helps teachers to build muscle memory. Yes. So, it's actually pretty effective but it does seem like … I know isn't it neat. 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 a strength (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.*

The transcript then goes on to identify an area of growth,

> *So, one of the things I'd like to focus our kind of our coaching conversation on is how we can have you offer more specific redirections for student behavior so you can keep doing what? Right. You can keep teaching.*

The coach continues by engaging the candidate in role play: "And I'll be Ethan this time. So, let's practice a specific response".

Another low-adherence transcript (0.15) demonstrates again how off-topic 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, does not clearly identify the candidate's strength, and fails to identify an area of growth:

> *What do you think?... I think that's the key when you're providing like behavioral redirections, and there are other you know like layers on this but like 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.*

These excerpts 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, when an implementer repeatedly uses an uncommon term while successfully delivering treatment components, the measure will nonetheless identify the transcript as low adherence. For

example, the following excerpt is from a transcript that received a very low script similarity score (0.13) despite delivering all components:

> *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. 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 are acceptable implementations of the treatment protocol. This indicates that script similarity may not be distinguishing between good and excellent implementation in the TeachSIM context.

This 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 script similarity measure, particularly for low script similarity scores. By sampling several transcripts for qualitative analysis, we gain an understanding for how to interpret different semantic similarity scores for a particular intervention. We are also able to identify limitations of the scores in this context – e.g. low adherence scores may represent cases where the coach uses very unusual language, but follows the intervention protocol. These validation exercises are relatively low-cost but have the potential to increase the value of semantic similarity measures in providing information on program implementation.

Intervention Replicability Across Studies

A key assumption for replication efforts with multiple studies is that intervention conditions are delivered consistently (Steiner et al., 2019; V. Wong et al., 2020). Using the replicability measure, we 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. 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 with the same pedagogical task and simulation scenario (behavior management or feedback). 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 was Feedback Study 1. This study also has the lowest replicability scores, implying that transcripts from this study do not stray from the script in the same ways. In order to gain insights into the amount of variation represented by replicability scores, we recommend that researchers take a similar approach to that which we took for adherence scores: sampling transcripts from either ends of the distribution for human interpretation.

Discussion

A semantic similarity approach to measuring intervention adherence and replicability brings many potential advantages. First, so long as a researcher has or is or is able to obtain transcriptions of treatment sessions, semantic similarity methods may be implemented at low-cost and are nearly infinitely scalable. Researchers only need transcriptions and moderate computer programming skills. We hope that the relatively low cost of measuring the similarity of treatment transcripts to a benchmark script will encourage researchers who would not otherwise include measures of fidelity to incorporate the measures presented here in their impact evaluations. Second, the automated nature of semantic similarity techniques means that 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. This provides a strong argument for including

semantic similarity measures of adherence alongside more complex, but potentially unreliable, approaches to fidelity measurement like observation rubrics. 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 treatment replication is a novel contribution for replication science. Transcript similarity directly addresses the question of treatment stability and consistency, measuring changes in intervention implementation that may not be captured using an adherence rubric.

Despite these advantages, semantic similarity measures are not a one-size-fits all solution. There are two primary considerations that researchers should evaluate before using a semantic similarity approach to assess intervention adherence or replication. 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 delivered through dialogue and employs standardized or at least highly structured language. 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 adherence measures for implementers, emphasizing that rote delivery is not required, and they should sample high-adherence transcripts for human review to ensure that implementers are appropriately responding to participants.

There are also cases where semantic similarity methods will simply be too blunt to satisfy the researcher's 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 benchmark script. Thus, script similarity may both underrepresent some components of the intervention and contain irrelevancies. If a researcher is simply interested in determining the relationship between fidelity 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.

The second consideration is resources. The greatest cost to using NLP techniques is the cost of obtaining transcriptions. Automated transcription services are available and generally low-cost, but often require human editing on the backend to increase accuracy. Thankfully, speech recognition

technology is continuing to improve and there is some evidence that even noisy transcriptions contain rich information on social interactions (Georgiou et al., 2011). Often, researchers have plans to transcribe intervention sessions regardless of their intent to apply NLP techniques (this was the case in the study of the impact of coaching presented in this paper). The cost of applying semantic similarity in these cases is quite low.

Ultimately, a researcher's decision on whether to incorporate semantic similarity measures of implementation constructs 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, the researcher is interested in discriminating between components of the theory of change, or the researcher has the resources, they should use traditional methods of assessing fidelity: observational rubrics and surveys. Or, if the researcher needs to be able to discriminate between components of the theory of change, but the study is too large to employ trained observers in every session, 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 paper 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 the 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. 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 pre-processing 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 non-adherence 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 paper 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 to 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 paper as a jumping off point and will adapt our proposed approach to their particular circumstances and research questions.

Bibliography

Adams, G., & Carnine, D. (2003). Direct Instruction. In *Handbook of Learning Disabilities* (pp. 403–416).

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

Baron, J. (2013). Randomized controlled trials commissioned by the Institute of Education Sciences since 2002: How many found positive versus weak or no effects. *Washington, DC: Coalition for Evidence-Based Policy*.

Borman, G. D., Dowling, N. M., & Schneck, C. (2008). A Multisite Cluster Randomized Field Trial of Open Court Reading. *Educational Evaluation and Policy Analysis*, *30*(4), 389–407. https://doi.org/10.3102/0162373708326283

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final Reading Outcomes of the National Randomized Field Trial of Success for All. *American Educational Research Journal*, *44*(3), 701–731. https://doi.org/10.3102/0002831207306743

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

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

Dobson, D., & Cook, T. J. (1980). Avoiding type III error in program evaluation. *Evaluation and Program Planning*, *3*(4), 269–276. https://doi.org/10.1016/0149-7189(80)90042-7

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.

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

Fixsen, D. L., Blase, K., Friedman, R., & Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature*. University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to Response to Intervention: What, Why, and How Valid Is It? *Reading Research Quarterly*, *41*(1), 93–99. JSTOR.

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* (Working Paper 23276; NBER Working Papers, pp. 1–54). https://www.nber.org/papers/w23276

Georgiou, P. G., Black, M. P., Lammert, A. C., Baucom, B. R., & Narayanan, S. S. (2011). "That's Aggravating, Very Aggravating": Is It Possible to Classify Behaviors in Couple Interactions Using Automatically Derived Lexical Features? In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective Computing and Intelligent Interaction* (Vol. 6974, pp. 87–96). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-24600-5_12

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

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*(1), 87–110. https://doi.org/10.1080/19345740802539325

Institute of Education Sciences. (2020a). *Standards for Excellence in Education Research*. https://ies.ed.gov/seer/index.asp

Institute of Education Sciences. (2020b). *Standards for Excellence in Education Research: Document treatment implementation and contrast*. https://ies.ed.gov/seer/implementation.asp

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. https://doi.org/10.3102/0013189x18785613

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1746–1751. 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.

Marsh, J. A., Bush-Mecenas, S., Strunk, K. O., Lincove, J. A., & Huguet, A. (2017). Evaluating Teachers in the Big Easy: How Organizational Context Shapes Policy Responses in New Orleans. *Educational Evaluation and Policy Analysis*, *39*(4), 539–570. https://doi.org/10.3102/0162373717698221

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

Řehůřek, R. (2011). *Gensim: Topic modelling for humans*. https://radimrehurek.com/gensim/auto_examples/index.html

Roberts, G. (2017). Implementation Fidelity and Educational Science: An Introduction. In G. Roberts, S. Vaughn, S. N. Beretvas, & V. Wong (Eds.), *Treatment Fidelity in Studies of Educational Intervention* (pp. 1–21).

Rossi, P. H., Freeman, H. E., & Sandefur, G. D. (1985). *Evaluation: A Systematic Approach*. SAGE Publications.

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

Schneider, M. (2018, December). *Message from IES Director: A More Systematic Approach to Replicating Research*. Institute of Education Sciences (IES), part of the U.S. Department of Education (ED). https://ies.ed.gov/director/remarks/12-17-2018.asp

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding Variation in Treatment Effects in Education Impact Evaluations: An Overview of Quantitative Methods* (Issue April, pp. 1–50). https://ies.ed.gov/ncee/pubs/20144017/pdf/20144017.pdf

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.

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*. *19–68*, 1–64.

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

Wong, V., & Steiner, P. (2018). Replication designs for causal inference. In *EdPolicyWorks Working Paper Series* (No. 62; EdPolicyWorks Working Paper Series, Issue 62). EdPolicyWorks. https://curry.virginia.edu/sites/default/files/uploads/epw/62_Replication_Designs.pdf

Table 1

Sample and Setting Characteristics by Study

	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 [.05]	0.26 [.06]	0.23 [.06]	0.16 [.06]	0.36 [.09]

Table 2

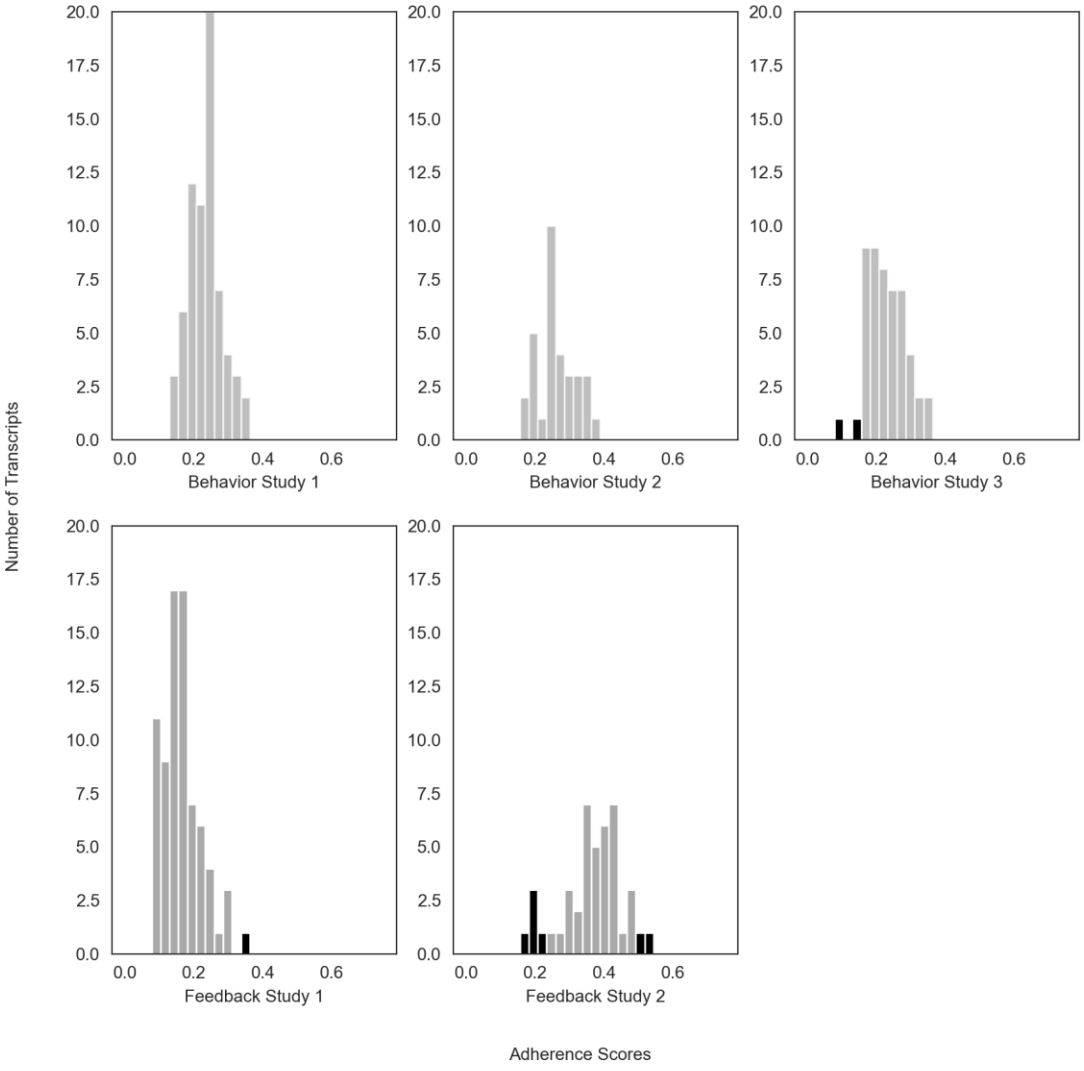
Study Adherence

	(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.

Figure 1

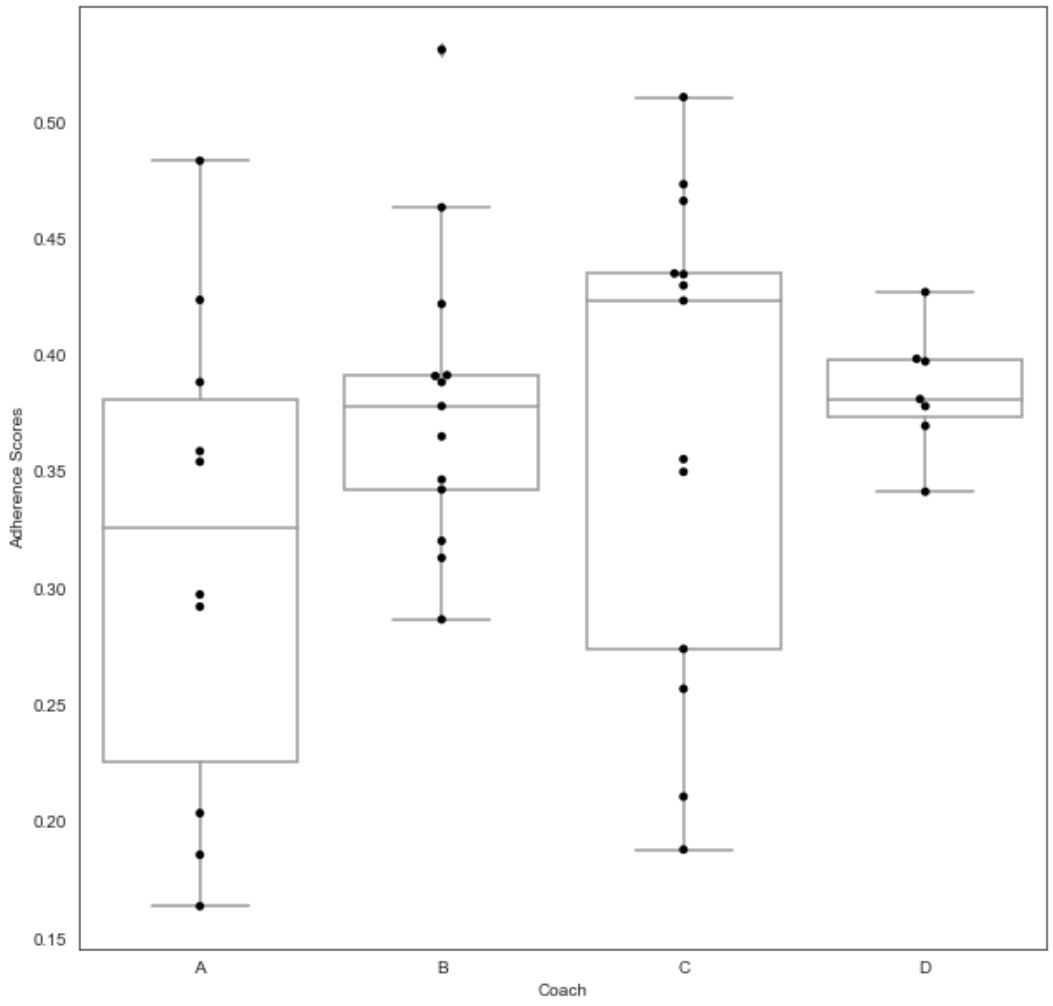
Distribution of Adherence Scores by Study, with Unusual Transcripts Highlighted



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.

Figure 2

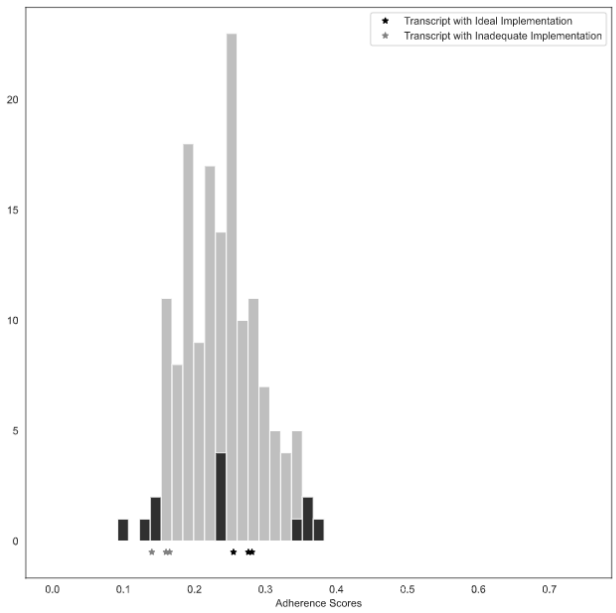
Distribution of Adherence Scores by Coaches within Feedback Study 2



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.

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.

Table 3

Replicability Matrix

	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 tf-idf 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).

Appendix A

Example Coaching Script for the Behavior Management Scenario

Below, we provide an example coaching script labelled 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 A1

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	To make your next simulation even stronger, I want you to focus on making your redirections more timely so that you

	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.	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 B

A Selective Overview of Advanced NLP Techniques

Each of the methods described in the main body of the paper are relatively straight-forward 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 1000). Thus, each document is represented as a high-dimensional matrix. Applying cosine similarity to multiple matrices is not straight-forward. To sidestep this problem, researchers often simply average the word embeddings for a document (reducing the word embeddings matrix to a vector; Řehůřek, 2011), 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 paper 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 (for example 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 (for example, 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 C

Semantic Similarity Statistics by Study and Pre-Processing 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, tf-idf weighting, lemmatization, and latent semantic analysis. The techniques are cumulative so that the final set of results uses all of the previous pre-processing methods. Each table demonstrates a consistent pattern. The highest similarity scores are produced without any text pre-processing. 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 B1

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.

Table B2

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.

Table B3

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.

Table B4

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.

Table B5

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.