ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 20-312

Using Semantic Similarity to Assess Adherence and Replicability of Intervention Delivery

Kylie L. Anglin
University of Virginia

Vivian C. Wong
University of Virginia

Researchers are rarely satisfied to learn only whether an intervention works, they also want to understand why and under what circumstances interventions produce their intended effects. These questions have led to increasing calls for implementation research to be included in high quality studies with strong causal claims. Of critical importance is determining whether an intervention can be delivered with adherence to a standardized protocol, and the extent to which an intervention protocol can be replicated across sessions, sites, and studies. When an intervention protocol is highly standardized and delivered through verbal interactions with participants, a set of natural language processing (NLP) techniques termed semantic similarity can be used to provide quantitative summary measures of how closely intervention sessions adhere to a standardized protocol, as well as how consistently the protocol is replicated across sessions. Given the intense methodological, budgetary and logistical challenges for conducting implementation research, semantic similarity approaches have the benefit of being low-cost, scalable, and context agnostic for use. In this paper, we demonstrate how semantic similarity approaches may be utilized in an experimental evaluation of a coaching protocol on teacher pedagogical skills in a simulated classroom environment. We discuss strengths and limitations of the approach, and the most appropriate contexts for applying this method.

VERSION: October 2020

Suggested citation: Anglin, Kylie L., and Vivian C. Wong. (2020). Using Semantic Similarity to Assess Adherence and Replicability of Intervention Delivery. (EdWorkingPaper: 20-312). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/n5qj-7310

Using Semantic Similarity to Assess Adherence and Replicability of Intervention Delivery

Kylie L. Anglin and Vivian C. Wong

University of Virginia

September 2020

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. 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, Arielle Boguslov, Brian Wright, and members of the TeachSIM team at the University of Virginia for their feedback on earlier versions of this paper. All errors are those of the authors.

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. However, for effective interventions to be understood, researchers don't simply want to know *whether* an intervention works, they also want to understand *why* and *under what circumstances* an intervention makes a difference in students' lives. These questions have 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, 2020). 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; an implementation study provides needed information on how the program was *actually* delivered to participants.

In this paper, we introduce a natural language processing (NLP) technique called semantic similarity to describe how closely intervention sessions adhere to a standardized treatment protocol, and how consistently the protocol is replicated across sessions. At its core, semantic similarity methods quantify the distance between items based on the likeness of semantic content. In evaluation contexts, the method is best applied in cases where an intervention is delivered verbally and transcrips of the intervention sessions are available, and the choice of words are believed to be important for the interventions' delivery. To produce a measure of *intervention adherance*, researchers use semantic similarity methods to evaluate the similarity between session transcripts and a benchmark script of the intervention. To produce a measure of *intervention replicability*, researchers quantify the similarity of intervention transcripts with one another. A key strength of the approach is that it can be adapted to a number of implementation contexts simply by changing the documents to which transcripts are being compared.

Currently, implementation researchers face intense logistical, methodological, and budgetary constraints in efforts to assess intervention fidelity. Reviews of education research show that implementation fidelity research 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. For example, when Munter et al. (2014) asked their intervention developers to validate their fidelity coding scheme, they found that, even among experts, variation in fidelity coding was too great to be useful. To further complicate the issue, there has 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 of 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, with multiple research teams.

Given the above challenges, the education research community would benefit from low-cost and scalable methods for assessing intervention adherence and replicability in field settings. In this paper, we demonstrate how semantic similarity methods may be used to create new automated measures of fidelity. The remainder of this paper is structured as follows. First, we provide a short overview of traditional implementation approaches for assessing intervention fidelity. Second, we discuss how semantic similarity measures may be used to assess implementation constructs like intervention adherence and replicability. Third, we detail how researchers calculate and interpret semantic similarity measures of implementation. Fourth, we illustrate the method using data from an experimental evaluation of a coaching protocol on teacher performance in a simulated classroom environment. Finally, we conclude with a discussion of the 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, intervention implementation may be useful for explaining variations in outcomes in a single study (Schochet et al., 2014), and for explaining variations in effects across multiple studies (Steiner et al., 2019; Wong & Steiner, 2018). Finally, monitoring intervention implementation while practitioners

are in the field provide 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).

The most common form of implementation research 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). Dosage is defined as the amount of the intervention that is delivered and is commonly measured by counting each intervention session or by timing the length of the session (Dane & Schneider, 1998; Dusenbury et al., 2003). Adherence is defined as the extent to which implementation conforms to theoretical guidelines and is commonly measured using checklists or rubrics (Dumas et al., 2001; Dusenbury et al., 2003). Quality of delivery refers to way the program components are implemented. For example, researchers may measure teaching quality (Sarama et al., 2008) or implementer communication skills (Dumas et al., 2001) as part of their implementation study. Participant responsiveness is defined as the degree to which participants engage with the intervention. This may be measured using attendance or with more nuance using rubrics (Hulleman & Cordray, 2009). Finally, program differentiation (also referred to as strength of contrast) is defined as the difference between the program and business-as-usual, or in RCTs, between the treatment and control conditions. Program differentiation commonly relies on one of the other conceptualization of fidelity, contrasting the measure between the treatment and comparison group (Cordray & Pion, 2007; Hulleman & Cordray, 2009).

In each of the previous examples, the researchers had a pre-specified concept of ideal implementation, but there are also cases where a researcher has not defined quality but simply wishes to better understand the various ways an intervention may be implemented. This often occurs when the researcher has not designed the intervention but is nonetheless responsible for its evaluation, as is common with government programs. Even if the researchers have designed the intervention, they may be largely agnostic to implementation styles if they believe that interventions should be adapted to individual contexts (Dane & Schneider, 1998). If adaptation is viewed as a positive development, then researchers want to document and measure variation to better understand the nature of the intervention. In these cases, researchers commonly describe differences

in treatments across sites qualitatively (see Spillane, 1998) or they may more formally survey or interview the implementers (see Angrist et al., 2013; Marsh et al., 2017).

Semantic Similarity as a Measure of Intervention Adherence and Replication

Semantic similarity methods provide quantitative measures of *intervention adherence* to a standardized protocol, and of *intervention replicability* across different sessions, settings, and studies. While traditional forms of implementation research require the development of new, psychometrically validated measures for assessing implementation delivery in each unique context, semantic similarity measures may be used to measure intervention adherence and replicability across many contexts. However, it is important to note that these measures have specific and limited definitions as implementation constructs.For example, the adherence measure evaluates fidelity to a specified script but makes no attempt to differentiate whether an intervention was implemented with high or low quality. Nor does it measure participant responsiveness or program differentiation. Further, replicability, is one method of measuring adaptation but does not have a direct relationship with the fidelity constructs described above. Again, a study may have high or low replicability while having high or low quality intervention sessions.

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 – text can be represented by their vocabulary and compared to one another by the relatively frequency with which they use a set of words. The method may also incorporate a number of natural language techniques to better extract meaning from documents (for example, handling synonyms and context). These techniques, and an in-depth explanation of how semantic similarity is calculated, will be discussed in the next section. For now, we will simply define a few terms: a *document* is a single text of interest and a *corpus* is the full set of documents a researcher is in interested in comparing. From the corpus, a researcher creates a *document-term matrix* where each row corresponds to a document and each column corresponds to a word in the corpus. The values in the columns are the frequency with which a document uses each word. Semantic similarity is then calculated using a distance measure like cosine similarity (the cosine of the angle between document vectors). Cosine similarity ranges from zero to one, where two documents with a similarity of one share the exact same comparative words frequencies.

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 language-based standardized interventions is similar to the problem of plagiarism. Here, the documents of interest are transcriptions of intervention sessions. 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 differences in language that do not change the meaning of the text. In this section, we discuss a number of interesting comparisons a researcher can measure to gain a greater understanding of an intervention's implementation.

Intervention Adherence

Intervention adherence is commonly defined as "the degree to which specified procedures are implemented as planned or theorized" (Dane & Schneider, 1998, p. 23). In highly standardized, scripted interventions, adherence can be conceptualized as the degree to which an implementer's language follows a well-specified script of the intervention protocol. Many education interventions are delivered through standardized, scripted protocols, including interventions for struggling readers (Vadasy et al., 2006), coaching protocols for teacher candidates (Cohen et al., 2020), and behavioral therapies for students with autism spectrum disorder (Stevenson et al., 2000).

In evaluation contexts, adherence scores is 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. Then, for a given document, d_i, and a benchmark script s, script similarity is determined by the following:

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

From there, the researcher can determine which intervention sessions are closest to the benchmark and which intervention sessions deviate farthest by exploring the distribution of script similarity measures.

The researcher can also calculate the average script similarity for a study, site or interventionalist. In this measure, script similarity scores are summed across all transcript sessions and divided by the total number of transcripts (*n*), as given by:

$$Average\ Script\ Similarity = \frac{\sum_{i=1}^{n} sim(d_i, s)}{n}.$$

Here, average script similarity is interpreted as the average intervention adherence. In a single study, the measure may be used for reporting overall intervention adherence; in cases with multiple studies, sites, or interventionalists, average scores may be used to compare the relative intervention adherence according to a benchmark script.

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 resarchers are still in the field, the method may be used to identify sessions that stray from the benchmark and if needed, provide interventionalists with additional training. Another advantage is that once transcripts are obtained, there are no additional costs for training and employing observers to rate each isession according to an adherence rubric (the most common approach to assessing intervention fidelity).

However, the method is not meant to replace all types of implementation research. 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 program theory and goals. Nor does the method evaluate the quality of the intervention session itself, with the exception of semantic deviations from the benchmark. In this way, script similarity for assessing intervention adherence may best be understood as a quick and scalable – but narrow – measure of fidelity. In many field settings, however, even narrow but feasible measures of fidelity can provide researchers with an important tool for tracking and understanding how the intervention is actually delivered.

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 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)". In traditional implementation research, these assumptions are probed qualitatively, but semantic similarity provides a quantitative measure of consistency across studies.

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 document in a study is compared to every other document in that study. The average similarity of document d_j to every document in a set of *n* documents including document d_j is calculated as:

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

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 treatment 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 documents is calculated:

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

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 (with the same word frequencies and content) 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; it is also agnostic to the evaluation context in which it is applied. The measure may be especially useful in cases where intervention adherence is low (as measured by script similarity), but the researcher may still want to know whether sessions were delivered consistently, or if sessions were also highly deviant from each other as well. 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 intervention supports.

Other Comparisons of Interest

Beyond the comparisons presented here, we expect that researchers can derive their own similarity measures that are most well suited to address their questions of interest. For example, researchers may be interested in monitoring variations within and between interventionists so that they can intervene if an interventionist veers too far from the team. Or, in evaluation contexts where the researcher wants to measure treatment differentiation, semantic similarity methods may be used to quantify how similar and different conversations are between study conditions. We hope this paper serves as a jumping off point for researchers to think about designing their own measures of similarity for implementation research.

Calculation and Interpretation

In this section, we describe in greater detail how to calculate semantic similarity and provide an overview of techniques for representing text numerically. 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. Here, we first focus on the simplest method of representation, word frequencies, and later introduce more complex methods of representation which are often better able to discern differences in meaning.

Calculating Semantic Similarity with Word Frequencies

The first step in calculating semantic similarity is to create a *document-term matrix* where each row corresponds to a document $(i = 1, \ldots, N)$ and each column corresponds to a term in the set of

documents. 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. After a researcher has created their document-term matrix, the standard method of quantifying the similarity of two documents is to calculate the cosine similarity of their vector representations (Manning et al., 2008).

The cosine similarity of any two documents, d_1 and d_2 may be calculated using the following formula:

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

where the numerator is the dot product of the two document vectors and the denominator is the product of the magnitude of the two vectors. 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. Because cosine similarity is a measure of the angle between vectors; the magnitude of the vectors is irrelevant. For this reason, the length of the documents is also irrelevant; it is the *relative* word frequencies that provide meaning.

Using a simple document-term matrix to calculate semantic similarity assumes that all words in the corpus are wholly distinct with no relationship to one another and that all words matter equally in distinguishing between texts. These assumptions are rarely palatable to researchers, but there are a number of pre-processing techniques that may be incorporated to address them.

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 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 term frequency-inverse document-frequency (tf-idf) which assigns words weights based on their relative frequency between documents in a corpus. 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 down-weights 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 Relationships Between Terms

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 with lemmatizing, we still fail to capture the substitutability and similarity of words in a given context. To address this, we 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. With LSA, the meaning of a word is represented as the average of the meaning of all the documents in which it appears, and the meaning of documents are an average of the meaning of the words they contain (Landauer et al., 1998). Similar to factor analysis, LSA is based on singular value decomposition and reduces the terms in a document-term matrix into 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 100 is a common rule of thumb for many tasks (Deerwester et al., 1990).

Finally, to retain the semantic meaning of every word, a researcher may choose to incorporate word embeddings. Similar to LSA, the underlying assumption of word embeddings is that "a word is characterized by the company it keeps (Firth, 1957)". Unlike LSA, word embeddings zoom in on a word's context noting not just the document in which the word is found, but the context within that document. 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 good word embedding model will assign related words like *student* and *child* spatially close vector representations. In practice, researchers will likely use publicly available pre-trained word embeddings, but should note that the semantic relationships between words in a word embedding depend on the context in which the model was trained.

Remaining Assumptions

All of the previous techniques 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 (Gentzkow et al., 2017; Manning et al., 2008). Nonetheless, 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*. Beyond this, researchers may also make use of syntactic n-grams where words are grouped together not by their order but by their grammatical dependencies (Goldberg & Orwant, 2013) or may use extend the contextual representation given by word embeddings to full documents (Le & Mikolov, 2014)

Interpretation

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). 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; the texts likely 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 stemming 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 replication 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 looking at and comparing several modeling approaches to interpret intervention adherence and replicability scores, which we will demonstrate below in the applied example below. When looking at individual transcripts, researchers can visually inspect adherence and replication scores on a histogram to identify outliers. Similarly, they can rank transcripts to identify those with the highest and lowest adherence, or the transcripts that are most similar or dissimilar. When looking at groups of transcripts, researchers can present means and standard deviations to compare intervention adherence and replication scores across studies, sites or implementers. Again, they may also rank studies by their scores identifying which studies have the highest adherence and which studies are the best replications of one-another. We believe that these comparisons will be useful in a number of contexts, but this list of potentially relevant comparisons is not exhaustive; researchers can make whatever comparisons they want so long as the all semantic similarity measures are calculated using the same NLP techniques.

Implementing Semantic similarity Approaches

Semantic similarity methods may be implemented using a number of programming languages. In the example below, 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.

An Illustrative Example

Evaluation Context

In this section, we apply our proposed measures of intervention adherence and replication to a series of RCTs evaluating the impact of coaching on the performance of preservice teachers in a mixed-reality simulated classroom environment (see Cohen et al., 2020 for an evaluation of one of these experiments; Krishnamachari et al. for a description of the systematic replication evaluation of the coaching protocol). We analyzed implementation of a coaching protocol within and across five conceptual RCT replications that took place in a university-based teacher preparation program.

In each RCT, preservice teachers practiced a pedagogical task over two five-minute interactions with student avatars in a simulation platform. In between the two simulation sessions, treatment teachers received feedback from a coach who followed a standardized five-step conversation protocol that: 1) asked the candidate to assess their own performance; 2) affirmed an observed effective teaching practice; 3) identified one of four skills for the candidate to target in the next session; 4) engaged the candidate in skill development by asking, "Last time when Student X did/said Y, you responded by [insert teacher candidate response here]. Next time, how might you respond in a way that is more [insert targeted skill here]"; and, 5) engaged the candidate in role-play so that the candidate could practice their target skill. Coaches were expected to adhere to the same conversation protocol regardless of the pedagogical task that the candidate was practicing or the simulation scenario.

Coaching conversations were replicated over five individual experiments with systematic variations in timing, pedagogical tasks, and populations. 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. Table 1 presents summary statistics of the five RCTs. In three studies, coaching conversations focused on improving teacher candidate responses to off-task student behavior (Behavior Study 1, 2, and 3). In the other two studies, coaching conversations focused on improving the quality of instructional feedback that teacher candidates provide to students (Feedback Study 1 and 2). In four out of five studies, the participant population was drawn from methods courses for education majors. In Behavior Study 3, however, the participant population was drawn from non-majors in a course designed for students interested in exploring teaching as a profession. No matter the sample, pedagogical task, or timing, coaches were expected to follow the protocol described above. Feedback Study 1 was the first study conducted and was considered a pilot as coaches and student avatars went through very little training for this study.

The goals of applying the semantic similarity measure were to provide evaluation researchers with summary quantitative measures of the extent to which the coaching protocol was delivered with adherence and consistency within and across studies, and to allow researchers to identify outlier sessions that may indicate the need for additional coach training.

Data

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. 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]*).

To assess adherence, the research team also created a set of eight ideal scripts representing perfect implementation of the treatment protocol. Each script was tailored to one of four skills. In the feedback scenario, these skills were probing for textual evidence, scaffolding student understanding, providing descriptive feedback, or probing for a warrant. In the behavior scenario, these skills were providing redirections that are timely, specific, succinct, or calm. Depending on the scenario, a coach was expected to target one of these skills and then follow the treatment protocol for that skill.

To produce our analytic dataset, we cleaned plain text transcripts to exclude speaker tags, time tags, and any formatting characters (for example newline, *n*). We also excluded teacher candidate text to focus our analysis on coaches' implementation of the protocol rather than participant responses to coach prompts[1].

The final cleaned transcripts were stored as a variable in a dataset containing the participant ID, study, coach ID, and targeted skill.

Methods

Pre-processing. To facilitate comparisons across studies and measures, and to test the robustness of our results to the techniques employed, we created multiple document-term matrices using our full corpus of documents, including all transcripts and ideal scripts[2]. The first matrix includes all of terms in the corpus with no pre-processing, 3501 columns in total. In the second matrix, we excluded stop words from a popular pre-specified list (Python's Natural Language Toolkit – NLTK) supplemented with a set of common pause fillers and vocal ticks like "uh" and "um". Removing stop words reduced the number of columns in the document-term matrix from 3501 to 3364.

In the third matrix, we replaced all word derivatives with a single stem, further reducing the size of the matrix to 2573 columns. In the fourth matrix, we weighted each term using tf-idf weighting. Finally, in our fifth matrix, we performed LSA on a document-term matrix with stop word removal and tf-idf weighting keeping the 100 most common concepts.

[1] 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.
[2] As explained in the section titled, Calculation and Interpretation, if we had instead created separate document-term matrices for each study, we could not draw comparisons across studies as the scale and meaning of the semantic similarity measures would not be stable.

This final matrix, which incorporates LSA, stop word removal, and tf-idf weighting, is our preferred approach because LSA gracefully handles the problem of word derivatives and similarities (for example, because *teacher* and *instructor* are likely used in the same contexts, they will be a part of the same topics), while tf-idf weighting helps ensure that the abstract concepts created by LSA contain meaning (no concepts will be driven by stop words). As a robustness check, however, we present intervention adherence and replicability scores using each of the above approaches and hope to see consistent patterns when comparing transcripts and studies.

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 similarity of every transcript 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.

Result

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 when there are 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 our preferred method of text processing: LSA with stop word removal and tf-idf weighting.

Intervention Adherence. We present adherence scores in two formats: in summary tables displaying means (Tables 1 and 2) and in figures displaying the distributions (Figures 1, 2, and 3). Table 1 provides an example of how adherence scores might be included in summary tables alongside other study statistics like sample sizes and participant characteristics. We present adherence scores, listed at the bottom and calculated using LSA with stop word removal and tf-idf weighting, alongside other study descriptors. Like the other information in Table 1, the adherence

scores allow readers to quickly compare a key characteristic across studies. For example, Table 1 shows readers that they should pay close attention to Feedback Study 2 if they are interested in a high-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 ranks each study according to their average adherence scores across each of the five pre-processing approaches. A higher ranking in the table indicates higher adherence. We look at this table to determine the robustness of patterns to analytic decisions. Results in our case are largely robust; after stop words have been removed, study ranks are relatively stable no matter the pre-processing techniques employed[3]. Across techniques, Feedback Study 2 has the highest average adherence while transcripts from our pilot Feedback Study 1, have the lowest average adherence. Table 2 also shows that adherence improved the second time the coaches implemented a protocol: Feedback Study 2 has higher adherence than Feedback Study 1 and Behavior Study 2 has higher adherence than Behavior Study 1. We do not, however, see an increase in adherence between Behavior Studies 2 and 3. This may be because Behavior Study 3 was implemented with a different target population of participants, non-Education majors enrolled in class exploring teaching as a profession.

<Insert Tables 1 and 2>

In evaluation studies, it is often useful to describe the overall distributions of adherence scores. In a single study, we might look to the distribution to gain an understanding of the shape and spread of intervention adherence. We can determine if adherence is left-skewed (the mean is driven lower by a few low-adherence sessions) or right-skewed (the mean is driven by a few high-adherence sessions, and whether the distribution is tightly centered (indicating treatment is consistently delivered) or flat (indicating inconsistency). We can also compare distributions across multiple studies or sites, as in Figure 1. When sites and studies are highly standardized to the same protocol, the distribution of scores across sites or studies are highly overlapping; but when sites and studies have not been well-standardized, the distribution of scores may vary widely. Figure 1 displays the

[3] 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.

distributions of adherence scores by study and indicates that despite differences in means, the distributions of our studies are mostly overlapping. However, Feedback Study 2 appears to have a much wider distribution of adherence scores than other studies. We explore this distribution further in Figure 2 which disaggregates adherence scores by coach. In Figure 2, we see that the lowest end of the distribution in Feedback Study 2 is driven by a single coach – coach A – whose median score is much lower than that of the other coaches (0.32 compared to 0.49, 0.47, and 0.44). This suggests that Coach A could benefit from additional training.

<Insert Figures 1 and 2>

Finally, Figure 3 provides an example of how researchers might use visualization to identify abnormal transcripts that fall outside of the distribution of adherence scores. These are the transcripts we would recommend that researchers read to determine if there are any transcription errors, any implementer misunderstandings that need to be corrected, or any conditions which result in particularly high adherence. In Behavior Study 2, we highlight a low-adherence transcript and in Behavior Study 3 and Feedback Studies 1 and 2, we highlight transcripts with particularly high adherence. In Figure 3, we can also see an interesting bi-model distribution in Feedback Study 2; this distribution is explained by the low-adherence coach we observe in Figure 2.

<Insert Figure 3>

Intervention Replicability

A key assumption for systematic replication studies is that intervention conditions are delivered consistently across studies (Wong, Steiner, & Anglin, under review). 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 would 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. In other words, Behavior Studies 2 and 3 are closer replications of Behavior Study 1 than the feedback

studies. 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 script will necessarily be close to one-another. Transcripts that are far from the script, however, 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 way.

<Insert Table 3>

Discussion

A semantic similarity approach to measuring intervention adherence and replicability brings many potential advantages. First, although traditional implementation research is important to understanding intervention delivery, it requires significant resources. Automated approaches may lower this barrier. 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 replication will have perfect reliability; if the same method is applied to the same transcript, the same measure will result each time. This favorable quality is 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 have begun to drift 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. While there may be times when researchers are only interested in documenting changes that are relevant to the theory of change, there may be other times researchers simply need to measure adaptation. Transcript similarity offers a solution in these cases. Further, because transcript similarity is agnostic to the theory of change, the same measure may be applied across a variety of intervention contexts, from educational interventions to cognitive behavioral therapy.

Despite these advantages, semantic similarity measures are not one-size-fits all solution to questions of implementation. 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 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.

The second consideration is construct validity. To provide an appropriate measure of adherence and replication, 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 intervention are highly standardized and scripted. However, even in these cases, semantic similarity methods may create measures which are too blunt to satisfy researcher's needs. Rubrics are capable of measuring multiple components of a theory of change while script similarity measures a only single construct - the similarity between a treatment transcript and an ideal script. Thus, script similarity may both underrepresent some components 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 component in a theory of change has the strongest relationship with effect sizes, semantic similarity is unlikely to be helpful.

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 standardized, 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. An examination and illustration of classification methods for fidelity is beyond the scope of this paper, but we point readers to Anglin (2019) and Fesler et al. (2019) for approachable overviews of natural language classifiers in education and to Atkins et al. (2014) and Can et al. (2016) for an example application of machine learning classifiers to fidelity in evaluations of motivational interviewing.

Finally, a potential solution to addressing some of the challenges discussed above is to combine semantic similarity measures with other measures of implementation constructs. For example, researchers may evaluate the fidelity of a sample of intervention sessions using rubrics as a formal measure and then informally monitor fidelity for the rest of the sample using semantic similarity. In multi-stage trials, researchers may employ both a semantic similarity and rubric approach in the pilot study in order to validate the semantic similarity measure and to determine appropriate cut-offs for manual inspection of transcripts. Finally, researchers may use rubrics to assess intervention fidelity and semantic similarity to assess replicability. When implementation has a low adhereance score, researchers can determine with the replicability score whether implementers diverge in similar ways from the protocol (low-adherence but high-replicability), or stray in a similar manner (low-adherence and low-replicability).

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, and reliable method for assessing intervention adherence and replicability for standardized scripted 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. However, a semantic similarity approach to implementation research can be adopted to a variety of

implementation constructs across a broad array of intervention-types and contexts. 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

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

Angrist, J. D., Pathak, P. A., & Walters, C. R. (2013). Explaining Charter School Effectiveness. *American Economic Journal: Applied Economics*, *5*(4), 1–27.

Atkins, D. C., Steyvers, M., Imel, Z. E., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, *9*(1), 49. https://doi.org/10.1186/1748-5908-9-49

Can, D., Marín, R. A., Georgiou, P. G., Imel, Z. E., Atkins, D. C., & Narayanan, S. S. (2016). "It sounds like...": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology*, *63*(3), 343. https://doi.org/10.1037/cou0000111

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

Cordray, D. S., & Pion, G. M. (2007). Treatment Strength and Integrity: Models and Methods. *Strengthening Research Methodology: Psychological Measurement and Evaluation.*, *4*, 103–124. https://doi.org/10.1037/11384-006

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

Deerwester, S., Dumais, S., Furnas, G., & Landauer. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

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

Fesler, L., Dee, T., Baker, R., & Evans, B. (2019). Text as Data Methods in Education. *Journal of Research on Educational Effectiveness*, 707–727.

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

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.

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

Goldberg, Y., & Orwant, J. (2013). A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books. *Second Joint Conference on Lexical and Computational Semantics*, *Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, 6.

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.

Hulleman, C. S., & Cordray, D. S. (2009). Methodological studies: 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. (2020). *Standards for Excellence in Education Research*. https://ies.ed.gov/seer/index.asp

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

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning*, *32*, 9.

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

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*.

Munter, C., Wilhelm, A. G., Cobb, P., & Cordray, D. S. (2014). Assessing Fidelity of Implementation of an Unprescribed, Diagnostic Mathematics Intervention. *Journal of Research on Educational Effectiveness*, *7*(1), 83–113. https://doi.org/10.1080/19345747.2013.809177

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

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.

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

Sarama, J., Clements, D. H., Starkey, P., Klein, A., & Wakeley, A. (2008). Intervention, evaluation, and policy studies: Scaling up the implementation of a pre-kindergarten mathematics curriculum: Teaching for understandingwith trajectories and technologies. *Journal of Research on Educational Effectiveness*, *1*(2), 89–119. https://doi.org/10.1080/19345740801941332

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. 50–50). http://search.proquest.com/docview/1651827259?accountid=13042

Spillane, J. P. (1998). A Cognitive Perspective on the Role of the Local Educational Agency in Implementing Instructional Policy: Accounting for Local Variability. In *Educational Administration Quarterly* (Vol. 34, Issue 1, pp. 31–57). https://doi.org/10.1177/0013161X98034001004

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.

Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2006). Code-oriented instruction for kindergarten students at risk for reading difficulties: A randomized field trial with paraeducator implementers. *Journal of Educational Psychology*, *98*(3), 508–528. https://doi.org/10.1037/0022-0663.98.3.508

Wong, V. C., & Steiner, P. M. (2018). Replication Designs for Causal Inference. In *EdPolicyWorks Working Paper Series* (Issue 62). http://curry.virginia.edu/uploads/epw/62_Replication_Designs.pdfhttp://curry.virginia.edu/edpolicyworks/wp

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
Intervention Delivery Setting					
Pedagogical Task in Simulation Scenario	Beahvior Management	Behavior Management	Behavior Management	Providing Feedback	Providing Feedback
Timing	Spring 2018	Spring 2019	Fall 2019	Fall 2017	Fall 2018
N (treatment and control)	99	98	93	122	93
N (treatment transcriptions)	68	45	47	76	46
Adherence Score from Semantic Similarity Measure	0.26 [.05]	0.31 [.05]	0.26 [.06]	0.20 [.06]	0.42 [.10]

Table 2

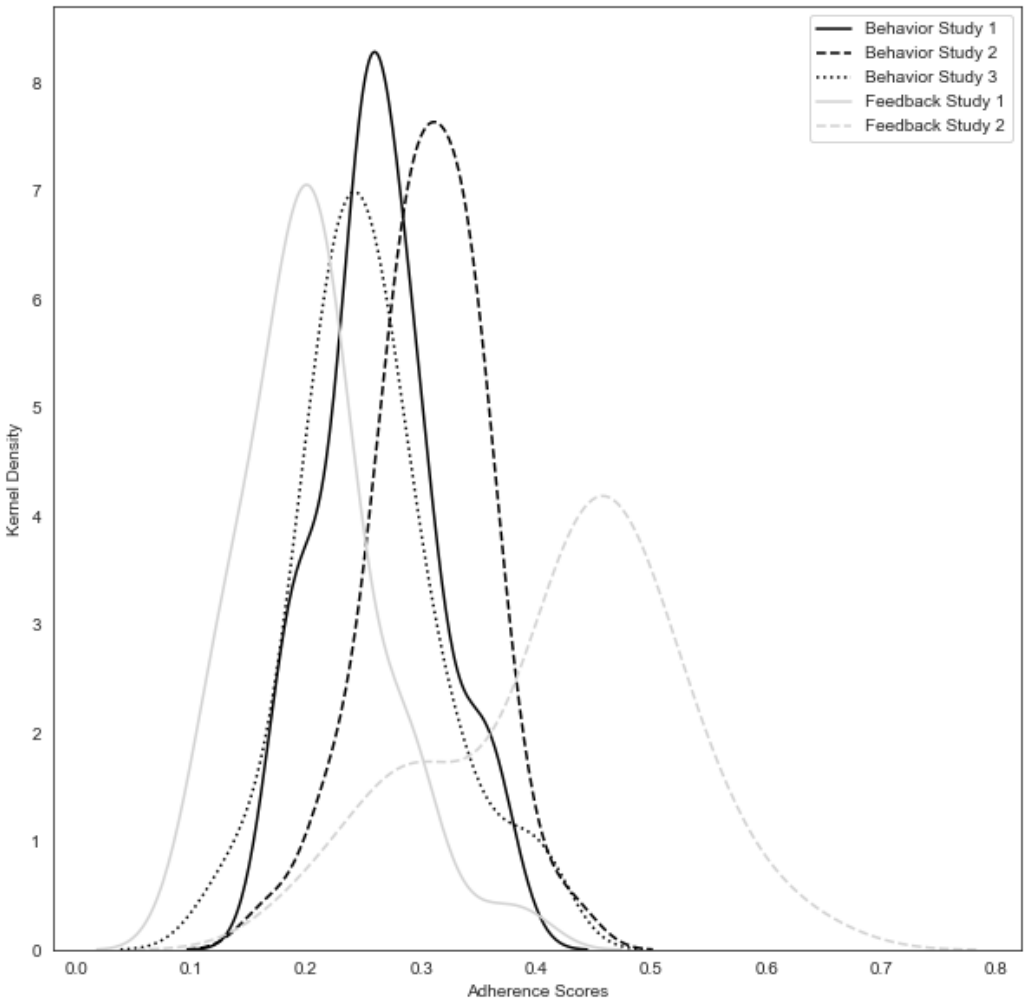
Study Adherence Rankings

	(1)	(2)	(3)	(4)	(5)
Behavior Study 1	4	4	3	3	3
Behavior Study 2	1	2	2	2	2
Behavior Study 3	2	3	4	3	3
Feedback Study 1	5	5	5	5	5
Feedback Study 2	3	1	1	1	1
LSA					X
Remove Stop Words		X	X	X	X
TF-IDF Weighting			X	X	X
Stemming				X	

Note. Adherence scores were estimated by calculating the cosine similarity between each transcript and a benchmark script.Values in the rows represent ranking by the average adherence score for each study. A higher ranking indicates greater adherence to the script.

Figure 1

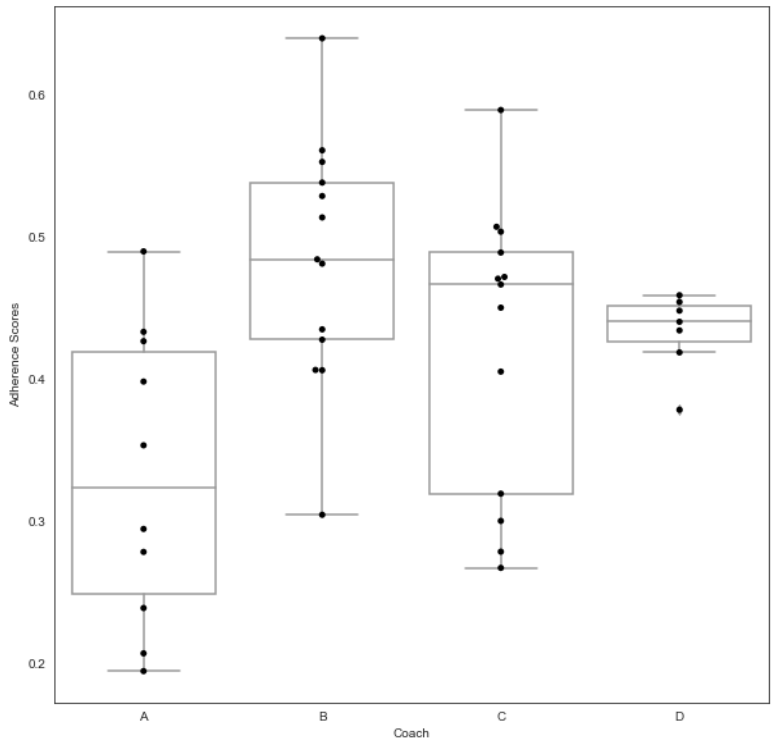
Distribution of Adherence Scores for Five Studies



Note. Adherence scores were estimated by calculating the cosine similarity between each transcript and an ideal script using a document-term matrix with latent-semantic analysis, no stop words, and term-frequency-inverse-document-frequency weighting. A high score indicates higher adherence to the script.

Figure 2

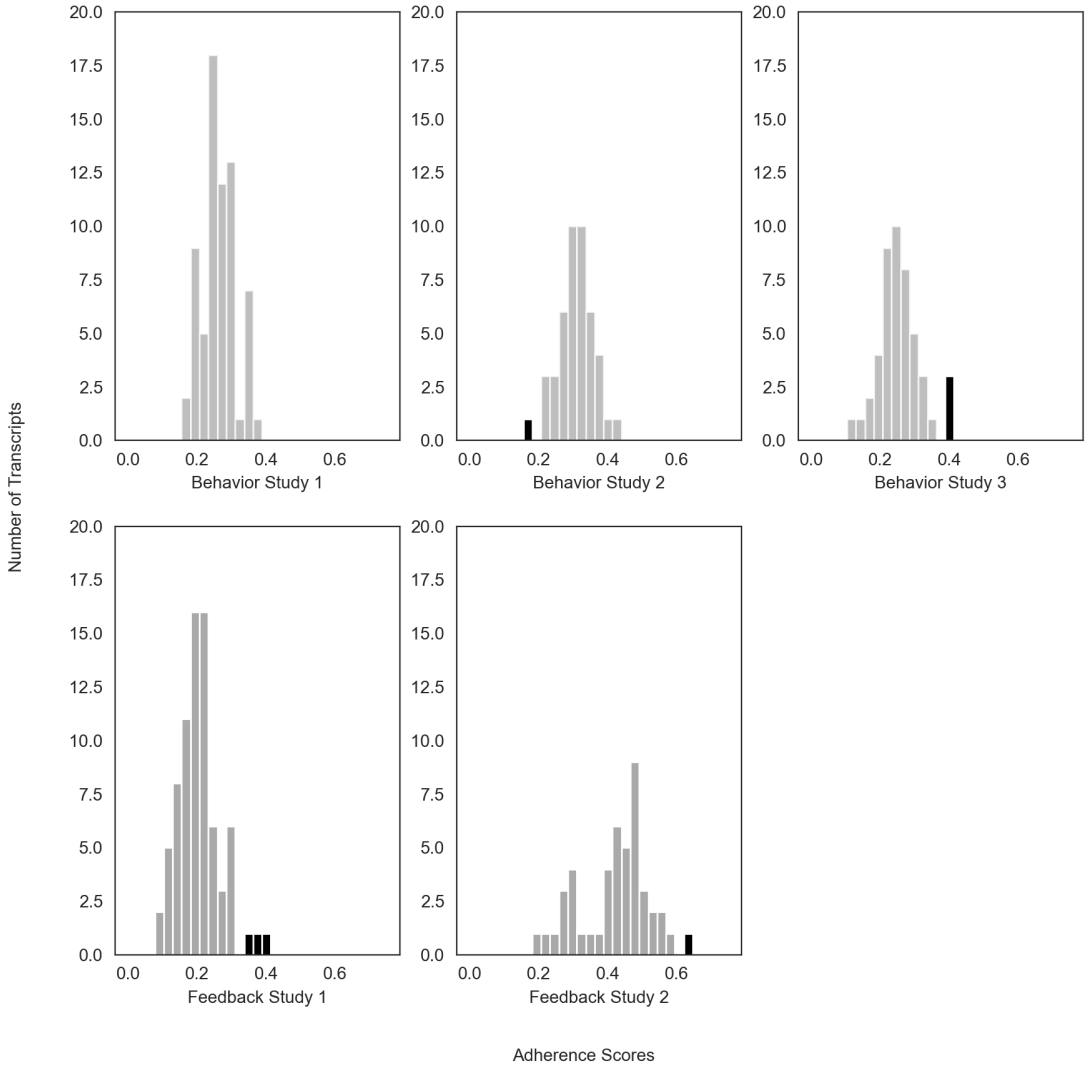
Distribution of Adherence Scores by Coaches within Feedback Study 2



Note. Adherence scores were estimated by calculating the cosine similarity between each transcript and a benchmark script using a document-term matrix with latent-semantic analysis, no stop words, and term-frequency-inverse-document-frequency 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 by Study, with Unusual Transcripts Highlighted



Note. Adherence scores were estimated by calculating the cosine similarity between each transcript and a benchmark script using a document-term matrix with latent-semantic analysis, no stop words, and term-frequency-inverse-document-frequency weighting. A high 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.

Table 3

Replicability Matrix

	Behavior Study 1	Behavior Study 2	Behavior Study 3	Feedback Study 1	Feedback Study 2
Behavior Study 1	**0.40**				
Behavior Study 2	0.31	**0.40**			
Behavior Study 3	0.27	0.33	**0.37**		
Feedback Study 1	0.20	0.20	0.20	**0.35**	
Feedback Study 2	0.22	0.24	0.23	0.30	**0.50**

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 latent-semantic analysis, no stop words, and term-frequency-inverse-document-frequency weighting. Cells shaded in dark gray (on the diagonal) display the similarity of transcripts to other transcripts within the same study. Cells shaded in light gray display the similarity of transcripts to other studies within the same context (behavior management or feedback).