



A machine learning approach to improve implementation monitoring of family-based preventive interventions in primary care

Implementation Research and Practice
Volume 4: Jan-Dec 2023 1–13
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/26334895231187906
journals.sagepub.com/home/irp



Cady Berkel^{1,2} , Dillon C. Knox³, Nikolaos Flemotomos³, Victor R. Martinez³, David C. Atkins⁴, Shrikanth S. Narayanan³, Lizeth Alonso Rodriguez², Carlos G. Gallo⁵ and Justin D. Smith⁶

Abstract

Background

Evidence-based parenting programs effectively prevent the onset and escalation of child and adolescent behavioral health problems. When programs have been taken to scale, declines in the quality of implementation diminish intervention effects. Gold-standard methods of implementation monitoring are cost-prohibitive and impractical in resource-scarce delivery systems. Technological developments using computational linguistics and machine learning offer an opportunity to assess fidelity in a low burden, timely, and comprehensive manner.

Methods

In this study, we test two natural language processing (NLP) methods [i.e., Term Frequency-Inverse Document Frequency (TF-IDF) and Bidirectional Encoder Representations from Transformers (BERT)] to assess the delivery of the Family Check-Up 4 Health (FCU4Health) program in a type 2 hybrid effectiveness-implementation trial conducted in primary care settings that serve primarily Latino families. We trained and evaluated models using 116 English and 81 Spanish-language transcripts from the 113 families who initiated FCU4Health services. We evaluated the concurrent validity of the TF-IDF and BERT models using observer ratings of program sessions using the COACH measure of competent adherence. Following the Implementation Cascade model, we assessed predictive validity using multiple indicators of parent engagement, which have been demonstrated to predict improvements in parenting and child outcomes.

Results

Both TF-IDF and BERT ratings were significantly associated with observer ratings and engagement outcomes. Using mean squared error, results demonstrated improvement over baseline for observer ratings from a range of 0.83–1.02 to 0.62–0.76, resulting in an average improvement of 24%. Similarly, results demonstrated improvement over baseline for parent engagement indicators from a range of 0.81–27.3 to 0.62–19.50, resulting in an approximate average improvement of 18%.

Conclusions

These results demonstrate the potential for NLP methods to assess implementation in evidence-based parenting programs delivered at scale. Future directions are presented.

¹ College of Health Solutions, Arizona State University, Phoenix, AZ, USA

² Ming Hsieh Department of Electrical Engineering, USC Viterbi School of Engineering, REACH Institute, Arizona State University, Tempe, AZ, USA

³ Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, CA, USA

⁴ Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA

⁵ Department of Psychiatry and Behavioral Sciences, Northwestern University, Chicago, IL, USA

⁶ Department of Population Health Sciences, Spencer Fox Eccles School of Medicine, University of Utah, Salt Lake City, UT, USA

Corresponding author:

Cady Berkel, College of Health Solutions, Arizona State University, 425 N. 5th St, Phoenix, AZ 85004 | ABC 226, USA.

Email: cady.berkel@asu.edu



Trial registration

NCT03013309 ClinicalTrials.gov.

Plain Language Summary: Research has shown that evidence-based parenting programs effectively prevent the onset and escalation of child and adolescent behavioral health problems. However, if they are not implemented with fidelity, there is a potential that they will not produce the same effects. Gold-standard methods of implementation monitoring include observations of program sessions. This is expensive and difficult to implement in delivery settings with limited resources. Using data from a trial of the Family Check-Up 4 Health program in primary care settings that served Latino families, we investigated the potential to make use of a form of machine learning called natural language processing (NLP) to monitor program delivery. NLP-based ratings were significantly associated with independent observer ratings of fidelity and participant engagement outcomes. These results demonstrate the potential for NLP methods to monitor implementation in evidence-based parenting programs delivered at scale.

Keywords

implementation, implementation outcomes, integrated care, preventive intervention, provider, health care, treatment fidelity, validity

Introduction

Evidence-based parenting programs have an extensive body of literature supporting their potential to prevent the onset and escalation of a broad array of negative behavioral health outcomes for children and adolescents, including anxiety, depression, suicide, conduct problems, substance use, sexual risk, and academic underachievement (O'Connell et al., 2009). Research on the implementation of these preventive interventions has concluded that factors related to the implementers' delivery (e.g., fidelity/adherence, quality) and the participants' responsiveness to the program (e.g., attendance, active in-session participation, and home practice) are key drivers of program success (Berkel, Mauricio, et al., 2018; Dane & Schneider, 1998; Durlak & DuPre, 2008; Dusenbury et al., 2005). When evidence-based programs have been scaled out to community settings, declines in implementation have translated to diminished intervention effects, sometimes referred to as the "voltage drop" (Chambers et al., 2013; Henggeler, 2004; Kilbourne et al., 2007).

Gold-standard methods of implementation monitoring used in efficacy and effectiveness trials involve independent observations (Kazdin, 2003; Perepletchikova et al., 2007), which require extensive training and supervision of coders to maintain interrater reliability (Berkel et al., 2019; Schoenwald et al., 2011). Even though these gold-standard approaches have superior validity relative to other commonly used methods (e.g., facilitator self-report; Dusenbury et al., 2003; Mauricio et al., 2017), they are cost-prohibitive and impractical in resource-scarce community-based delivery systems (Berkel et al., 2019; Brown et al., 2013; Hanson et al., 2014). Because mechanisms to reimburse supervision time in community organizations are lacking, observational assessment of program delivery is rarely done in community settings (Schoenwald et al., 2011). Supervision is typically limited to discussions about administrative issues, and sometimes includes

problem solving for difficult cases (Aarons & Sawitzky, 2006). In addition to the clinical need for scalable implementation measurement, the field of implementation science is also lacking efficient implementation outcome measures that would facilitate the testing of implementation strategies at the organizational level (Proctor et al., 2013; Waltz et al., 2019). The need to use observational methods to assess implementation outcomes places a severe strain on already overstretched research budgets. Thus, the development of feasible and scalable measures of program delivery can address both clinical and scientific needs.

In recognition of the limited evidence for the validity of self-report, and the high burden of gold-standard observational techniques, multiple strategies have been proposed to balance feasibility and validity of such assessments. For example, Beidas and colleagues (2016) conducted a randomized trial comparing goal standard behavioral observations with self-report and two innovative approaches for assessing fidelity in CBT. One strategy was chart-stimulated recall, in which a trained rater conducts a structured interview with a provider using a patient chart to improve recall of what happened during the intervention. The second approach, which is often used during training, was behavioral rehearsal, in which providers role-played the intervention during a session with a trained rater. Results demonstrated that behavioral rehearsal was on par with behavioral observations; however, chart-stimulated recall significantly overestimated adherence, as did self-report (Becker-Haimes et al., 2022). A potential drawback is that both approaches depend on the availability of a trained rater, which may limit feasibility in resource-poor implementation environments.

Other innovative approaches have sought to make use of intervention documents to enhance feasibility. Whiltsey Stirman and colleagues (2018) developed a coding system that made use of fidelity via worksheets used in CBT sessions to schedule behavioral activities

and track effects on mood. Berkel, Sandler, and colleagues (2018) used home practice worksheets in a preventive intervention for divorcing families to assess the quantity and quality of engagement in program skills. They found that the quality of home practice was associated with improvements in parenting and child mental health outcomes; each 1-unit increase in home practice competence, rated by the program implementer, was associated with a reduction for the risk of borderline diagnosis of internalizing by one-half and externalizing by one-third (Berkel, Mauricio, et al., 2018; Berkel, Sandler, et al., 2018). Moreover, independent ratings of fidelity to program content were significantly associated with higher levels of home practice competence (Berkel, Mauricio, et al., 2018). There were also significant indirect effects of delivery quality on home practice competence, mediated through attendance. Based on these findings, a Dynamic Implementation Monitoring and Feedback System (DIMFS) was proposed, which would make use of home practice data to broadly monitor implementation and prompt a more in-depth review of sessions if home practice scores fell below an acceptable threshold (Berkel et al., 2019).

Within the DIMFS, natural language processing (NLP) was proposed as a potentially feasible and valid way to facilitate these more in-depth reviews. NLP combines computational linguistics and machine learning, offering an opportunity to assess delivery in a low burden, timely, and comprehensive manner (Berkel et al., 2019; Brown et al., 2013; Flemotomos et al., 2021; Gallo et al., 2021). Automated ratings of written or spoken language have been used in private sectors and education for decades (e.g., Shermis et al., 2015), and recently have been used to monitor the delivery of evidence-based interventions. For example, Atkins, Narayanan, and colleagues applied NLP to assess linguistically-based indicators of the quality of motivational interviewing (MI) in individual psychotherapy (Atkins et al., 2014; Can et al., 2016; Flemotomos et al., 2022; Imel et al., 2014; Xiao et al., 2015). They found high degrees of sensitivity and specificity (.63–.86) between human ratings and text classification of linguistically based quality indicators, including open-ended questions, reflections, and empathy. In the field of prevention, Gallo, Pantin, et al. (2015) used NLP to assess facilitators' use of open-ended questions in the Familias Unidas program and found high correspondence between human and machine ratings for open-ended questions. Validated NLP methods to assess indicators of program delivery could reduce the cost of objective assessment, improve the reliability of assessments, and enable a rapid feedback system to support sustained, high-quality implementation when programs are delivered at scale in community settings (Berkel et al., 2019; Gallo et al., 2021).

In the current study, we test machine learning methods in the assessment of the delivery of the Family Check-Up 4 Health (FCU4Health) program in the Raising Healthy Children study (Smith, Berkel, Jordan, et al., 2018). The

FCU4Health is an adaptation of the Family Check-Up (FCU; Dishion & Stormshak, 2007) developed to fit within the context of primary care (Berkel, Rudo-Stern, et al., 2020; Smith, Berkel, Rudo-Stern, et al., 2018). The original FCU is an evidence-based preventive intervention that was originally designed to prevent conduct problems and substance use through an individually tailored approach. An ecological assessment and MI skills are used to engage parents in a menu of follow-up support, including parenting modules and referrals to community-based resources, to improve family management and child behavioral health outcomes.

Multiple randomized trials have demonstrated the effects of the FCU on parenting, parental depression, child self-regulation, child conduct problems, and adolescent substance use (e.g., Dishion, Brennan, et al., 2014; Fosco et al., 2014; Shaw et al., 2009; Van Ryzin & Nowicka, 2013). Nonetheless, as with many preventive interventions, scale-up has been limited (O'Connell et al., 2009). Primary care was identified as a setting that could overcome many of the barriers to implementation, including a trusting, longitudinal relationship with pediatricians, the lack of stigma associated with primary care, and sustainable billing mechanisms (Leslie et al., 2016; Perrin et al., 2016). At the time, integration of behavioral health support in healthcare settings was limited, and while our primary care partners recognized the importance of parenting and behavioral health, it was seen as secondary to the focus on physical health—in particular, nutrition and pediatric obesity (Berkel, Rudo-Stern, et al., 2020). With emerging findings that the FCU had spillover effects on obesity, nutrition, and physical activity (Rudo-Stern et al., 2016; Smith et al., 2015; Van Ryzin & Nowicka, 2013), our primary care partners' perceptions about the appropriateness of the FCU for their context increased. We thus enhanced the program, now called the FCU4Health, to add a more explicit focus on family health routines and child health behaviors, such as nutrition, sleep, physical activity, and screen time, resulting in a whole-child health approach (Berkel, Rudo-Stern, et al., 2020; Smith, Berkel, Rudo-Stern, et al., 2018). In a randomized trial conducted with three primary care organizations, the FCU4Health program has demonstrated effects on parenting, parental depression, child self-regulation, internalizing, and externalizing (e.g., Berkel, Fu, et al., 2021), as well as family health routines and child health behaviors (e.g., physical activity, dietary choices; Smith, Berkel, et al., 2021; Smith, Carroll, et al., 2023).

Fidelity to the content and clinical processes (i.e., competent adherence) of the FCU and FCU4Health is assessed via the COACH observational rating system (Dishion, Smith, et al., 2014). Domains of the COACH include: (a) Conceptual accuracy and adherence to the FCU model, (b) Observant and responsive to client needs, (c) Actively structuring sessions to optimize effectiveness, (d) Careful and appropriate teaching, and (e) Hope and

motivation are generated (see Smith et al., 2013, for a full description). Following the implementation cascade model (Berkel et al., 2011; Berkel, Mauricio et al., 2018), research on the implementation of the FCU and FCU4Health has demonstrated that program implementers' delivery of the program influences participant responsiveness, which in turn is associated with improvements in parenting and child outcomes (Berkel, Mauricio, et al., 2021; Chiapa et al., 2015; Smith et al., 2013). In the Early Steps efficacy trial of the FCU, COACH scores predicted change in parenting skills at the 1-year follow-up; this relation was mediated by parent engagement (Smith et al., 2013). Drift in COACH ratings over a 4-year period (child ages 2, 3, 4, and 5), although fairly minimal, predicted distal changes in child oppositional and aggressive behaviors 3 years later (Chiapa et al., 2015). In the FCU4Health, COACH scores were associated with in-session active participation, engagement in parenting modules, and increases in parents' self-reported motivation to improve family management and child health (Berkel, Mauricio, et al., 2021). Moreover, COACH ratings were unrelated to parent language preference (Spanish vs. English), or baseline parent motivation or depression.

The Current Study

This study sought to validate the use of NLP-based ratings as a method of evaluating implementation in the FCU4Health program. We analyzed transcripts from the 113 English and Spanish-speaking families in the intervention condition of the Raising Healthy Children trial who participated in at least one program session to train and evaluate NLP models of program delivery. To assess concurrent validity, we compared NLP-based ratings to COACH scores rated by observers. To assess predictive validity, we also compared them to multiple indicators of parent engagement (in-session participation, attendance at follow-up parenting sessions, home practice ratings, and parent motivation to change) that previous research has found to be associated with human ratings of program delivery and predictive of improvements in parenting and child behavioral health (Berkel, Mauricio, et al., 2021; Chiapa et al., 2015; Smith et al., 2013).

Methods

Study Procedures

This study made use of data from the Raising Healthy Children study, a type 2 hybrid effectiveness-implementation trial of the FCU4Health conducted in partnership with three primary care systems; two federally qualified health centers (FQHCs) and one hospital-based outpatient primary care clinic (Smith, Berkel, Jordan, et al., 2018). These clinics are located in the urban center of Phoenix and serve primarily Mexican American families with low

income. The majority of children received Medicaid, and more than half of families struggled with food insecurity. Human subject's involvement was overseen by the Institutional Review Boards of Arizona State University and Phoenix Children's Hospital. All other institutions participating in this research provided signed reliance agreements ceding to the IRB of Arizona State University. Children aged 5.5–12 years with elevated BMI ($\geq 85^{\text{th}}$ percentile for age and gender) were identified by primary care providers, primarily during regular well-checks. Prior to participation, parents provided written consent for themselves and their children; children provided assent. Enrolled families ($n = 240$) completed baseline assessments, which included validated and normed surveys related to social determinants of health, parenting, child behavioral health, family health routines, and child health behaviors (Berkel & Smith, 2017). The assessments were conducted either in English or Spanish with native speakers, based on the preference of the family. Subsequent to the baseline assessment, families were randomly assigned to the FCU4Health ($n = 141$) or usual care ($n = 99$). Follow-up assessments were conducted at 3 months, 6 months, and 12 months after baseline. Families were compensated for participation in each wave of data collection (\$40 at baseline, \$25 at the 3-month follow-up, \$30 at the 6-month follow-up, and \$55 at the 1-year follow-up).

Families begin the FCU4Health with the first Feedback Session, in which FCU4Health Coordinators use MI skills to learn about family context and priorities, share the results of the baseline assessment, and work with the families to set goals for tailored follow-up supports. Depending on the identified needs, FCU4Health Coordinators offered parenting modules from the Everyday Parenting curriculum (Dishion et al., 2011) and/or coordination with community resources to address needs related to social determinants of health. In two follow-up Feedback Sessions (after the 3-month and 6-month follow-ups), FCU4Health Coordinators checked-in with families about their progress, addressed potential barriers, and set new goals. During the 6-month Feedback Session, families were also linked with ongoing support as needed. The 12-month assessment was for data collection only. Of the 141 families randomly assigned to the FCU4Health, 113 participated in the first Feedback Session. Feedback Sessions were video-recorded for clinical supervision and fidelity monitoring. These sessions were professionally transcribed for the purposes of this study ($n = 197$; 116 in English and 81 in Spanish).

Participants

Because the focus of the study is on the implementation fidelity of the FCU4Health, only families in the intervention condition were included. At baseline, children's mean age was 9.5 (SD = 1.9) years, and gender was 52% male and 48% female. Mean child BMI percentile was 115.6% above the 95th percentile. Caregivers were

predominantly female ($n = 130$; 92%). Caregivers' racial/ethnic background was: 68% Latino, 14% non-Latino White, 5% African American, 4% American Indian/Alaska Native, 2% Asian, and 5% multiple racial/ethnic categories. A majority of caregivers (62%) completed baseline assessments in English, with the remainder (38%) completing them in Spanish.

Measures and Coding Procedures

NLP-Based Ratings of Program Delivery

We conducted analyses to compare two commonly used NLP models: Term Frequency-Inverse Document Frequency (TF-IDF) and Bidirectional Encoder Representations from Transformers (BERT). TF-IDF focuses on word and phrase usage and how these vary across "documents" (here, session transcripts) as inputs to machine learning models. TF-IDF reduces the influence of generally common words and phrases and increases words and phrases that vary across sessions, which are likely to be more discriminative, especially those relying on small datasets (Bafna et al., 2016; Kadhim, 2019). Because these n -gram approaches can lead to thousands of potential features (i.e., regression inputs), features were reduced a priori using feature extraction, performed using the scikit-learn library (Pedregosa et al., 2011). Dimensional reduction was then performed using Latent Semantic Analysis. The reduced set of TF-IDF features was used as inputs to a support vector regression (SVR) with a Gaussian kernel, which is a traditional machine learning model that can yield robust prediction results with thousands of inputs. In the current application, individual SVRs were trained for each outcome.

TF-IDF approaches have a simplistic view of language in that they fundamentally are looking at word-by-document co-occurrence and do not bring any information about the meaning of individual words or phrases to a given prediction task. More recent NLP models, like BERT, make use of word embeddings. The basic idea of word embeddings is that a word's meaning can, in part, be derived from the words that frequently occur in close proximity to it across a great number of text documents. The notion of word embeddings can be generalized to phrases, sentences, or other linguistic units. Approaches using such embeddings have two separate steps: (a) embeddings are first learned in a large, general corpus and (b) the embeddings are then potentially adapted to a given domain and then used as inputs to a specific prediction task. For the present work, we employed a pre-trained multilingual BERT model (Devlin et al., 2019) to extract linguistic embeddings. This model has been trained on large-scale Wikipedia data for 104 languages, including English and Spanish. We adapted BERT using a corpus of general psychotherapy transcripts (Alexander Street, n.d.; Imel et al., 2015) to provide better, domain-relevant embeddings for the provider utterances. The corpus contains 330,050 therapist utterances from 3,642

session transcripts. The sequence of the adapted BERT embeddings then served as input into a long short term memory (LSTM)-based model, which is a common, deep learning model for NLP. The model was built using Tensorflow (Abadi et al., 2016) and trained using an Adam optimizer (Kingma & Ba, 2015), with a learning rate of $1e-3$ and a batch size of 16. Each experiment was run for a maximum of 100 epochs with early stopping with a patience of 10 epochs, based on validation loss. A unique model was trained for each outcome.

Observer Ratings of Program Delivery

The COACH observational rating system (Dishion, Smith et al., 2014) was used to rate fidelity to the FCU4Health feedback session protocol. The COACH assesses five dimensions of observable coordinator skill in the FCU4Health, which are rated separately on a 9-point scale: 1–3 (*needs work*), 4–6 (*competent work*), and 7–9 (*excellent work*). Coders were four family interventionists with training in the FCU4Health and experience delivering the program to families in the trial. Three bilingual coders were assigned sessions in Spanish and English, and one monolingual English coder was assigned only sessions in English. They did not rate any of their own sessions. They received approximately 20 h of training on COACH coding. First, the coding team conducted ratings of Feedback Sessions as a group led by one of the program developers (JDS). Next, they each independently rated 3–5 sessions to evaluate the reliability across coders. As was the case in previous trials, "agreement" was achieved if raters' scores were within one point of the gold standard on each dimension. The reliability criterion required scoring three sessions in a row with 85% agreement with gold standard ratings. Once the reliability criterion was met, the coders were assigned sessions and attended bi-weekly meetings to maintain reliability and minimize coder drift over time. Coders first reviewed the assessment results to establish familiarity with the family and develop a case conceptualization. This step has been shown to improve the reliability of COACH ratings (Smith et al., 2016). Coders then viewed the entire FCU4Health feedback session. To calculate interrater reliability, 20% of the sessions were randomly selected for independent rating by two different members of the team. The same 1-point criterion was used to calculate the percent agreement between coders. Agreement was in the acceptable range ($IRR = .74$).

Parent Motivation

At posttest, caregivers reported on their motivation to achieve seven goals for their families. Three of these goals, related to parenting and family dynamics, were taken from the original FCU (Fosco et al., 2014). Four new goals that relate to child health behaviors (e.g., nutrition, physical activity, sleep, and screen time) were added for the FCU4Health. Parents rated each of these goals on a

5-point scale with anchors informed by the transtheoretical model (Prochaska & DiClemente, 1983): 1 = *no change needed*, 2 = *thinking about change*, 3 = *wanting to change*, 4 = *taking steps to change*, and 5 = *working hard to change*. Cronbach's α for the full 7-item scale was .91.

Parent Engagement Outcomes

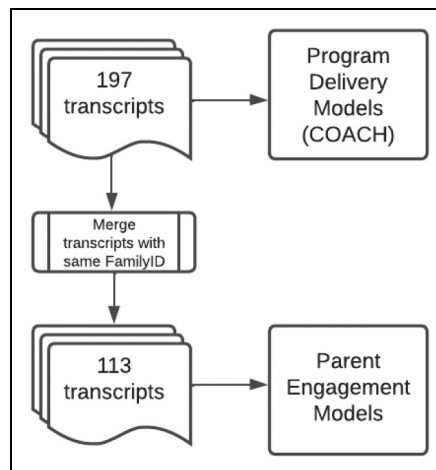
In-Session Engagement. Coders also rated caregivers' in-session engagement during the feedback session as: 1–3 (*low*, caregiver is inattentive or disengaged), 4–6 (*medium*, modest signs of engagement), and 7–9 (*high*, caregiver actively participates and is attentive and responsive). As with the COACH dimensions, the engagement dimension includes participant behaviors that reflect positive and negative indicators of engagement, including "Engages in 'change talk' by reflecting on the past and future" (positive), "Actively participates, nods head, and stays on topic during feedback" (positive), and "Angry or defensive during feedback session" (negative). Interrater reliability for the caregiver engagement item has been fair to excellent in previous studies (Chiapa et al., 2015; Smith, Dishion et al., 2013; Smith, Rudo-Stern, et al., 2019). Agreement was in the acceptable range (IRR = .73).

Attendance at Follow-Up Parenting Sessions. The FCU4Health activities checklist was adapted from a form used in the original FCU trials (Winter & Dishion, 2007) to capture administrative data used as part of program delivery. This form was used to capture the number of follow-up parenting sessions attended.

Home Practice Competence. Like many parenting programs, a core component of the FCU4Health is the practice of parenting skills between program sessions. Home practice was assigned at each parenting session. To reinforce the use of the skills and troubleshoot any changes, FCU4Health Coordinators asked parents to report back on how their skills practice went during the following session. To capture the quality of parents' home practice of program skills, we made use of a measure of home practice competence that we developed for an effective trial of the New Beginnings Program for divorcing families (Berkel et al., 2019). In that trial, program facilitators rated parents on a 5-point scale for each skill assigned during the week. These ratings were associated with improvements in multiple domains of parenting and child behavioral health outcomes; for every one-unit increase in home practice competence, the risk of child internalizing was cut in half and the risk for externalizing was cut by one-third (Berkel, Mauricio, et al., 2018; Berkel, Sandler, et al., 2018). We followed the same procedures in this trial. Specifically, on the subsequent meeting after each home practice was assigned, the FCU4Health Coordinators assigned the home practice and then rated parents' competence with using the home practice skills. For the purposes of this study, a mean was taken across all of the home practice ratings.

Figure 1

Number of Transcripts for the Assessment of Concurrent and Predictive Validity



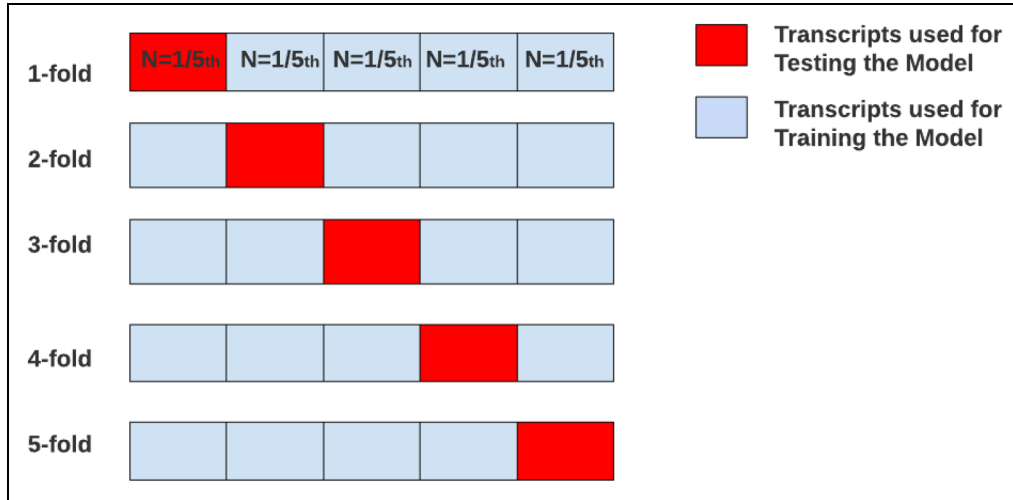
Analytic Approach

We ran analyses to assess the concurrent validity of the NLP-based ratings with observer ratings on the COACH and predictive validity to assess the extent to which NLP-based ratings were associated with measures of parent engagement established in previous studies noted above. For the former case, since annotations from multiple coders were available, we fused the multiple annotations using a majority vote, triplet embedding scheme (Booth et al., 2018, October). Such a method alleviates personal annotator biases and helps us generate more robust, "ground truth" target values of the COACH ratings. For the latter, to address the nesting of multiple Feedback Sessions within families, all transcripts for each family were combined into a single transcript (see Figure 1).

For each analysis, we report results using the TF-IDF and BERT embedding approaches, as detailed above. As is common in machine learning, models were trained on a subset of data, where the resulting model accuracy is then tested on the held-out data, not included in training. The dataset used to train and test the models was composed of both English and Spanish language sessions. We considered conducting separate analyses for English and Spanish transcripts. However, we elected to combine them for two primary reasons:

1. Although we define language preference based on a dichotomous indicator of the language parents chose to complete their assessments, this is an oversimplification. The Phoenix area is a bicultural community (Basilio et al., 2014) and even when parents choose English, they often communicate through a mix of English and Spanish (and vice versa for the parents who choose Spanish).

Figure 2
The 5-Fold Method for NLP Model Training and Testing



Note. Assignment was stratified by language to ensure equal representation for English and Spanish transcripts.

- The ultimate goal is to develop a system that can be used to monitor implementation for the entire population of patients served by the FCU4Health. Separating out English and Spanish transcripts would result in one system for English speakers and a separate system for Spanish speakers; each of which would need to be maintained and updated over time. Moreover, given the preference of many families to use both English and Spanish noted above, it is unclear how this would be accomplished.

We report the average results using 5-fold cross-validation, where the transcripts were divided into five non-overlapping groups (see Figure 2). The random assignment of transcripts to groups was stratified by language, and as a result, the proportion of Spanish:English transcripts was equal across groups. Each fold was independently held out as a test set, and a model was trained using the remaining folds. Mean squared error (MSE) on the held-out test set of data is the primary focus of evaluation. Here, MSE is given as follows:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2,$$

where n is the number of data points, Y_i are the observed values, and \hat{Y}_i are the predicted values. Lower MSEs indicate a better fit between the predicted values and the observed values. We compared our results against a simple baseline where the prediction every time is the numeric average of the target prediction for all the sessions in the training set. We conducted paired bootstrap tests to determine whether the models significantly improved prediction over baseline (Berg-Kirkpatrick et al., 2012).

Results

Concurrent Validity: Associations Between Machine Ratings and Human Ratings of Delivery

The results for each of the five COACH ratings of program delivery are provided in Table 1. All models predicted outcomes significantly better than the baseline ($p < .05$). Neither model was consistently superior to the other in predicting COACH scores: the LSTM model using BERT embeddings as the input representation performed best on three of the scores, while the support vector regressor using TF-IDF features produced the highest performance on the other two. We observed decreases in MSE relative to the baseline of 25.6% for *Conceptually Accurate*, 14.0% for *Observant and Responsive*, 25.9% for *Active Structuring*, 26.2% for *Careful and Appropriate Teaching*, and 26.0% for *Hope and Motivation*, for an average performance increase of 23.5% across the five scores.

Predictive Validity: Associations Between Machine Ratings and Participant Engagement

The results for the measures of parent engagement are provided in Table 2. Again, all models were found to produce results that were significantly better than baseline ($p < .05$) and neither model was consistently superior to the other. The LSTM model using BERT embeddings as the input representation performed best on three of the measures, while the support vector regressor using TF-IDF features produced the highest

Table 1*Concurrent Validity: Associations Between Machine Ratings and Observer Ratings of Delivery (Mean Squared Error)*

	Conceptually accurate	Observant and responsive	Actively structures	Careful and appropriate teaching	Hope and motivation
Baseline	0.83	0.76	1.02	0.88	1.00
TF-IDF SVM	0.64	0.66	0.77	0.65	0.75
BERT LSTM	0.62	0.69	0.76	0.67	0.74

Note. All models reflect significant improvement over baseline; best-performing model for each outcome is bolded. TF-IDF SVM = Term Frequency-Inverse Document Frequency; BERT LSTM = bidirectional encoder representations from transformers long- and short-term memory.

Table 2*Predictive Validity: Associations Between Machine Ratings and Participant Engagement Indicators (Mean Squared Error)*

	Observer ratings of in-session active participation	Number of follow-up parenting sessions attended	FCU4Health coordinator ratings of home practice competence	Parent-reported motivation
Baseline	0.77	5.46	0.81	1.48
TF-IDF SVM	0.77	4.86	0.80	1.46
BERT LSTM	0.62	5.16	0.64	1.34

Note. All models reflect significant improvement over baseline; best-performing model for each outcome is bolded. TF-IDF SVM = Term Frequency-Inverse Document Frequency; BERT LSTM = bidirectional encoder representations from transformers long- and short-term memory.

performance on the other two. The increase in absolute performance over the baseline for the best-performing model is 19.7% for in-session active participation ratings, 11.0% for the number of follow-up parenting sessions attended, 21.2% for home practice ratings, and 9.5% for motivation, for an average absolute performance increase of 18.0% across the five measures.

Discussion

As evidence-based parenting interventions move towards scale-up, feasible and valid methods to monitor delivery and provide timely feedback are needed to ensure public health impact (Berkel et al., 2019). There is a growing focus on developing innovative methods to address the limited feasibility of gold-standard observational methods and the limited validity of self-report (Beidas et al., 2016; Berkel et al., 2019; Wiltsey Stirman et al., 2018). NLP methods may offer an opportunity to assess program delivery in a low burden, timely, and comprehensive manner (Berkel et al., 2019; Gallo et al., 2021), which may surpass other approaches in terms of feasibility and validity. Although widely used in other settings, NLP has rarely been applied to behavioral health interventions, and more rarely still with evidence-based parenting programs (see Flemotomos et al., 2022; Gallo, Berkel, et al., 2015; Gallo, Pantin, et al., 2015 for exceptions). This study made use of transcripts from the FCU4Health program to assess the concurrent and predictive validity of automated ratings of program

delivery. Results demonstrated that both NLP methods tested within the current study (TF-IDF and BERT embeddings) were significantly associated with observer ratings using the COACH measure. The COACH has been used in both FCU and FCU4Health trials, and previous analyses have linked ratings on the COACH to parent engagement and improvements in parenting and child behavioral health outcomes (Berkel, Mauricio, et al., 2021; Chiapa et al., 2015; Smith et al., 2013). More importantly, they also significantly predicted multiple theoretically important domains of parent engagement in the program, including in-session active participation, the number of follow-up parenting sessions attended, ratings of home practice competence, and parent motivation to make a change. It should be noted that both models utilized different representations of the transcripts and are potentially capturing information in different ways that make one model better for predicting a given rating over the other. A benefit of NLP models is that it is quite feasible to use a combination of approaches that result in the best outcome.

Limitations and Implications for Future Research

We acknowledge a number of limitations in this study and provide suggestions about future directions for research in using NLP for implementation monitoring in evidence-based practice. First, the sample size was relatively small ($n = 197$ transcripts) compared to other NLP studies, and

there is a potential for results to be influenced by the specific context. In particular, it is important to consider the demographic breakdown of the sample (68% Latino; 38% Spanish preference) and transcripts (59% Spanish; 41% English). Even with other populations with high representation of Latinos and Spanish-speakers, dialects differ greatly across locations of origin (e.g., by country, rural vs. urban regions). These models should be replicated in other regions with different populations and dialects.

It should also be noted that for this study, we made use of human-created transcripts, rather than coding directly from the audio recordings using machine transcription methods. We attempted to use the audio directly and encountered challenges related to external noise (e.g., air conditioning) and multiple speakers (e.g., interruptions from children or other family members), particularly when sessions were conducted in families' homes. These challenges are more easily overcome by human transcription than machine transcription. In a current trial, FCU4Health sessions have moved entirely online to telehealth delivery (via zoom) due to the COVID-19 pandemic mitigation measures (Berkel, Smith, et al., 2020). We expect that this will result in better audio quality than the in-person meetings and facilitate coding directly from the audio. It should be considered that limited access to internet in rural areas could promote implementation inequities with this type of approach. The transition to telehealth during the COVID pandemic has shown a light on internet access as a social determinant of health, and initiatives are underway to increase access in rural communities across the country.

Finally, this was the first phase in a process to develop and test NLP methods for monitoring the implementation of evidence-based parenting programs like the FCU4Health. Several subsequent phases will be needed before this method can be included as part of an implementation support package to promote scale-up. First, as noted above, it will be critical to establish effective and feasible methods of coding directly from audio, as human transcription adds a substantial delay in providing feedback. We expect to be able to achieve this given the transition to telehealth delivery (noted above) and growing examples of AI-based analysis and feedback for evidence-based practices. Co-authors, Atkins and Narayanan, have developed Lyssn as a commercial platform (<https://www.lyssn.io/>), which grew out of their program of research for the assessment of MI skills in substance use therapy (Lyon et al., 2019). Issues of privacy and storage are also especially important in the primary care context. An external platform, like Lyssn's, with data sharing agreements in place, is probably the most feasible approach, given the large size of videos and the limited space within electronic health records. Moreover, privacy concerns are reduced through computer-based ratings. Second, analyses to establish thresholds for high vs. low delivery are needed to provide feedback to providers about areas

for improvement. Third, it will be important to test the potential acceptability and effectiveness of NLP-based feedback on program delivery on improvement or sustainment of high-quality implementation over time. Program facilitators in our previous work reported less performance anxiety about computer-based ratings of their delivery compared to supervisor ratings (Berkel et al., 2019). Improvement in delivery would only be expected to the extent that the system provided actionable information that was trusted by program implementers.

Once NLP methods are established and widely accessible, they can be used to assess fidelity, in situ adaptations, quality and cultural competence, and participant responsiveness at scale—to test, for example, the effectiveness of training and supervision models or compare implementation across cultural contexts. They can also facilitate in-depth questions about the nature of program delivery that were previously only available via highly intensive micro-coding studies, such as indicators of participant engagement and adherence to care, how program implementers address participant resistance, what is the role of emotion in program delivery (e.g., Berkel et al., 2013; Gallo et al., 2021; Gallo et al., under review).

Conclusions

Although it is well established that evidence-based parenting programs preempt the onset and escalation of child behavioral health problems (Ladis et al., 2018; Leijten et al., 2019; O'Connell et al., 2009; van Mourik et al., 2017), the lack of feasible and effective methods for monitoring the delivery and provide feedback in regular service settings may result in limited public health impact once taken to scale (Schoenwald et al., 2011). This study was one of the first to evaluate NLP methods for monitoring the delivery of an evidence-based parenting program. Results demonstrated support for the concurrent and predictive validity of two commonly used NLP methods (TF-IDF and BERT), establishing their potential for use in supporting the high-quality delivery of evidence-based parenting programs in primary care and other service delivery settings. Future research is needed to evaluate feasibility, acceptability, and effectiveness of NLP methods; however, the clinical and research potential of NLP in implementation research seems almost limitless.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding


This study was supported by grant U18 DP006255 from the National Center for Chronic Disease Prevention and Health Promotion of the Centers of Disease Control and Prevention,

under the Childhood Obesity Research Demonstration Project 2.0 (CORD), awarded to Cady Berkel and Justin D. Smith. The opinions expressed herein are the views of the authors and do not necessarily reflect the official policy or position of the Centers for Disease Control and Prevention.

Ethical Approval and Consent to Participate

This study was conducted in accordance with the basic ethical principles of autonomy, beneficence, justice, and nonmaleficence and will be conducted in accordance with the rules of Good Clinical Practice outlined in the most recent Declaration of Helsinki. The project was approved by the Institutional Review Board of Arizona State University on July 14, 2016 (Protocol 00004530), and by the Institutional Review Board of the Phoenix Children's Hospital on May 30, 2017 (Protocol 17-001). All other institutions participating in this research provided signed reliance agreements ceding to the Institutional Review Board of Arizona State University. Written informed consent of patients will be required. Data confidentiality and anonymity will be ensured, according to the provisions of United States law, both during the implementation phase of the project and in any resulting presentations or publications.

ORCID iD

Cady Berkel  <https://orcid.org/0000-0001-9664-9485>

References

- Aarons, G. A., & Sawitzky, A. C. (2006). Organizational climate partially mediates the effect of culture on work attitudes and staff turnover in mental health services. *Administration and Policy in Mental Health and Mental Health Services Research, 33*(3), 289–301. <https://doi.org/10.1007/s10488-006-0039-1>
- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., & Warden, P., ... Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation, Savannah, GA.
- Alexander Street. (n.d.). *Counseling and psychotherapy transcripts*. <https://alexanderstreet.com/products/counseling-and-psychotherapy-transcripts-series>
- 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*(49), 1–11. <https://doi.org/10.1186/1748-5908-9-49>
- Bafna, P., Pramod, D., & Vaidya, A. (2016). Document clustering: TF-IDF approach. International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, India.
- Basilio, C. D., Knight, G. P., O'Donnell, M., Roosa, M. W., Gonzales, N. A., Umaña-Taylor, A. J., & Torres, M. (2014). The Mexican American Biculturalism Scale: Bicultural comfort, facility, and advantages for adolescents and adults. *Psychological Assessment, 26*(2), 539–554. <https://doi.org/10.1037/a0035951>
- Becker-Haimes, E. M., Marcus, S. C., Klein, M. R., Schoenwald, S. K., Fugo, P. B., McLeod, B. D., Dorsey, S., Williams, N. J., Mandell, D. S., & Beidas, R. S. (2022). A randomized trial to identify accurate measurement methods for adherence to cognitive-behavioral therapy. *Behavior Therapy, 53*(6), 1191–1204. <https://doi.org/https://doi.org/10.1016/j.beth.2022.06.001>
- Beidas, R. S., Maclean, J. C., Fishman, J., Dorsey, S., Schoenwald, S. K., Mandell, D. S., Shea, J. A., McLeod, B. D., French, M. T., Hogue, A., Adams, D. R., Lieberman, A., Becker-Haimes, E. M., & Marcus, S. C. (2016). A randomized trial to identify accurate and cost-effective fidelity measurement methods for cognitive-behavioral therapy: Project FACTS study protocol. *BMC Psychiatry, 16*(1). <https://doi.org/10.1186/s12888-016-1034-z>
- Berg-Kirkpatrick, T., Burkett, D., & Klein, D. (2012, July). An empirical investigation of statistical significance in NLP. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 995–1005.
- Berkel, C., Fu, E., Carroll, A. J., Wilson, C., Tovar-Huffman, A., Mauricio, A., Rudo-Stern, J., Grimm, K. J., Dishion, T. J., & Smith, J. D. (2021). Effects of the Family Check-Up 4 Health on parenting and child behavioral health: A randomized clinical trial in primary care. *Prevention Science, 22*(4), 464–474. <https://doi.org/10.1007/s11121-021-01213-y>
- Berkel, C., Gallo, C. G., Sandler, I. N., Mauricio, A. M., Smith, J. D., & Brown, C. H. (2019). Redesigning implementation measurement for monitoring and quality improvement in community delivery settings. *The Journal of Primary Prevention, 40*(1), 111–127. <https://doi.org/10.1007/s10935-018-00534-z>
- Berkel, C., Mauricio, A. M., Rudo-Stern, J., Dishion, T. J., & Smith, J. D. (2021). Motivational interviewing and caregiver engagement in the Family Check-Up 4 Health. *Prevention Science, 22*(6), 737–746. <https://doi.org/10.1007/s11121-020-01112-8>
- Berkel, C., Mauricio, A. M., Sandler, I. N., Wolchik, S. A., Gallo, C. G., & Brown, C. H. (2018). The cascading effects of multiple dimensions of implementation on program outcomes: A test of a theoretical model. *Prevention Science, 19*(6), 782–794. <https://doi.org/10.1007/s11121-017-0855-4>
- Berkel, C., Mauricio, A. M., Schoenfelder, E., & Sandler, I. N. (2011). Putting the pieces together: An integrated model of program implementation. *Prevention Science, 12*(1), 23–33. <https://doi.org/10.1007/s11121-010-0186-1>
- Berkel, C., Murry, V. M., Roulston, K. J., & Brody, G. H. (2013). Understanding the art and science of implementation in the SAAF efficacy trial. *Health Education, 113*(4), 297–323. <https://doi.org/10.1108/09654281311329240>
- Berkel, C., Rudo-Stern, J., Abraczinskas, M., Wilson, C., Lokey, F., Flanigan, E., Villamar, J. A., Dishion, T. J., & Smith, J. D. (2020). Translating evidence-based parenting programs for primary care: Stakeholder recommendations for sustainable implementation. *Journal of Community Psychology, 48*(4), 1178–1193. <https://doi.org/10.1002/jcop.22317>
- Berkel, C., Sandler, I. N., Wolchik, S. A., Brown, C. H., Gallo, C. G., Chiapa, A., Mauricio, A. M., & Jones, S. (2018). “Home practice is the program:” Parents’ practice of program skills as predictors of outcomes in the new beginnings program effectiveness trial. *Prevention Science, 19*(5), 663–673. <https://doi.org/10.1007/s11121-016-0738-0>

- Berkel, C., & Smith, J. D. (2017). *Raising health children study battery*. Arizona State University.
- Berkel, C., Smith, J. D., Bruening, M. M., Jordan, N., Fu, E., Mauricio, A. M., Grimm, K. J., Winslow, E., Ray, K., Bourne, A., & Dishion, T. J. (2020). The Family Check-Up 4 Health: Study protocol of a randomized type II hybrid effectiveness-implementation trial in integrated primary care (the healthy communities 4 healthy students study). *Contemporary Clinical Trials*, *96*, 1–8. <https://doi.org/https://doi.org/10.1016/j.cct.2020.106088>
- Booth, B. M., Mundnich, K., & Narayanan, S. (2018, October). Fusing annotations with majority vote triplet embeddings. Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, pp.83–89.
- Brown, C. H., Mohr, D. C., Gallo, C. G., Mader, C., Palinkas, L., Wingood, G., Prado, G., Kellam, S. G., Pantin, H., Poduska, J., Gibbons, R., McManus, J., Ogihara, M., Valente, T., Wulczyn, F., Czaja, S., Sutcliffe, G., Villamar, J., & Jacobs, C. (2013). A computational future for preventing HIV in minority communities: How advanced technology can improve implementation of effective programs. *Journal of Acquired Immune Deficiency Syndromes*, *63*(Supp. 1), S72–S84. <https://doi.org/10.1097/QAI.0b013e31829372bd>
- 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), 343–350. <https://doi.org/10.1037/cou0000111>
- Chambers, D. A., Glasgow, R. E., & Stange, K. C. (2013). The dynamic sustainability framework: Addressing the paradox of sustainment amid ongoing change. *Implementation Science*, *8*(1), 117. <https://doi.org/10.1186/1748-5908-8-117>
- Chiapa, A., Smith, J. D., Kim, H., Dishion, T. J., Shaw, D. S., & Wilson, M. N. (2015). The trajectory of fidelity in a multiyear trial of the family check-up predicts change in child problem behavior. *Journal of Consulting and Clinical Psychology*, *83*(5), 1006–1011. <https://doi.org/10.1037/ccp0000034>
- 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. <http://ejournals.ebsco.com/direct.asp?ArticleID=5GR239V6912EURV1T3VT> [https://doi.org/10.1016/S0272-7358\(97\)00043-3](https://doi.org/10.1016/S0272-7358(97)00043-3)
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv, abs/1810.04805*.
- Dishion, T. J., Brennan, L. M., Shaw, D. S., McEachern, A. D., Wilson, M. N., & Jo, B. (2014). Prevention of problem behavior through annual family check-ups in early childhood: Intervention effects from home to early elementary school. *Journal of Abnormal Child Psychology*, *42*(3), 343–354. <https://doi.org/10.1007/s10802-013-9768-2>
- Dishion, T. J., Smith, J. D., Knutson, N., Brauer, L., Gill, A., & Risso, J. (2014). *Family check-up: COACH ratings manual. Version 2*. Available from the Child and Family Center, 6217 University of Oregon, Eugene, OR 97403.
- Dishion, T. J., & Stormshak, E. A. (2007). *Intervening in children's lives: An ecological, family-centered approach to mental health care*. American Psychological Association.
- Dishion, T. J., Stormshak, E. A., & Kavanagh, K. (2011). *Everyday parenting: A professional's guide to building family management skills*. Research Press.
- Durlak, J., & DuPre, E. (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. A., 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>
- Dusenbury, L. A., Brannigan, R., Hansen, W. B., Walsh, J., & Falco, M. (2005). Quality of implementation: Developing measures crucial to understanding the diffusion of preventive interventions. *Health Education Research*, *20*(3), 308–313. <https://doi.org/10.1093/her/cyg134>
- Flemotomos, N., Martinez, V. R., Chen, Z., Creed, T. A., Atkins, D. C., & Narayanan, S. (2021). Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations. *PLoS One*, *16*(10), e0258639. <https://doi.org/10.1371/journal.pone.0258639>
- Flemotomos, N., Martinez, V. R., Chen, Z., Singla, K., Ardulov, V., Peri, R., Caperton, D. D., Gibson, J., Tanana, M. J., Georgiou, P., Van Epps, J., Lord, S. P., Hirsch, T., Imel, Z. E., Atkins, D. C., & Narayanan, S. (2022). Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods*, *54*(2), 690–711. <https://doi.org/10.3758/s13428-021-01623-4>
- Fosco, G. M., Van Ryzin, M., Stormshak, E. A., & Dishion, T. J. (2014). Putting theory to the test: Examining family context, caregiver motivation, and conflict in the Family Check-Up model. *Development and Psychopathology*, *26*(2), 305–318. <https://doi.org/10.1017/S0954579413001004>
- Gallo, C. G., Berkel, C., Mauricio, A., Sandler, I., Wolchik, S., Villamar, J. A., Mehrotra, S., & Brown, C. H. (2021). Implementation methodology from a social systems informatics and engineering perspective applied to a parenting training program. *Familie, Systems & Health*, *39*(1), 7–18. <https://doi.org/10.1037/fsh0000590>
- Gallo, C. G., Berkel, C., Sandler, I. N., & Brown, C. H. (2015). Improving implementation of behavioral interventions by monitoring quality of delivery in speech. Annual Conference on the Science of Dissemination & Implementation, Washington, DC.
- Gallo, C. G., Li, Y., Berkel, C., Mehrotra, S., Liu, L., Benbow, N., & Brown, C. H. (under review). Recognizing emotion in speech for assessing the implementing behavioral interventions.
- Gallo, C. G., Pantin, H., Villamar, J., Prado, G., Tapia, M. I., Ogihara, M., & Brown, C. H. (2015). Blending qualitative and computational linguistics methods for fidelity assessment: Experience with the Familias Unidas preventive intervention. *Administration and Policy in Mental Health and Mental Health Services Research*, *42*(5), 574–585. <https://doi.org/10.1007/s10488-014-0538-4>
- Hanson, R. F., Gros, K. S., Davidson, T. M., Barr, S., Cohen, J., Deblinger, E., Mannarino, A. P., & Ruggiero, K. J. (2014). National trainers' perspectives on challenges to implementation of an empirically-supported mental health treatment. *Administration and Policy in Mental Health and Mental Health Services Research*, *41*(4), 522–534. <https://doi.org/10.1007/s10488-013-0492-6>
- Henggeler, S. W. (2004). Decreasing effect sizes for effectiveness studies - Implications for the transport of evidence-based

- treatments: Comment on Curtis, Ronan, and Borduin (2004). *Journal of Family Psychology*, 18(3), 420–423. <https://doi.org/10.1037/0893-3200.18.3.420>
- Imel, Z. E., Barco, J. S., Brown, H. J., Baucom, B. R., Baer, J. S., Kircher, J. C., & Atkins, D. C. (2014). The association of therapist empathy and synchrony in vocally encoded arousal. *Journal of Counseling Psychology*, 61(1), 146–153. <https://doi.org/10.1037/a0034943>
- Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Computational psychotherapy research: Scaling up the evaluation of patient-provider interactions. *Psychotherapy*, 52(1), 19–30. <https://doi.org/10.1037/a0036841>
- Kadhim, A. I. (2019). Term weighting for feature extraction on Twitter: A comparison between BM25 and TF-IDF. 2019 International Conference on Advanced Science and Engineering (ICOASE), pp.124–128.
- Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Allyn and Bacon.
- Kilbourne, A. M., Neumann, M. S., Pincus, H. A., Bauer, M. S., & Stall, R. (2007). Implementing evidence-based interventions in health care: Application of the replicating effective programs framework. *Implementation Science*, 2(1), 1–10. <https://doi.org/10.1186/1748-5908-2-42>
- Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. 3rd International Conference for Learning Representations, San Diego.
- Ladis, B. A., Macgowan, M., Thomlison, B., Fava, N. M., Huang, H., Trucco, E. M., & Martinez, M. J. (2018). Parent-focused preventive interventions for youth substance use and problem behaviors: A systematic review. *Research on Social Work Practice*, 29(4), 420–442. <https://doi.org/10.1177/1049731517753686>
- Leijten, P., Gardner, F., Melendez-Torres, G. J., van Aar, J., Hutchings, J., Schulz, S., Knerr, W., & Overbeek, G. (2019). Meta-analyses: Key parenting program components for disruptive child behavior. *Journal of the American Academy of Child & Adolescent Psychiatry*, 58(2), 180–190. <https://doi.org/https://doi.org/10.1016/j.jaac.2018.07.900>
- Leslie, L. K., Mehus, C. J., Hawkins, J. D., Boat, T., McCabe, M. A., Barkin, S., Perrin, E. C., Metzler, C. W., Prado, G., Tait, V. F., Brown, R., & Beardslee, W. (2016). Primary health care: Potential home for family-focused preventive interventions. *American Journal of Preventive Medicine*, 51(4), S106–S118. <https://doi.org/10.1016/j.amepre.2016.05.014>
- Lyon, A. R., Munson, S. A., Renn, B. N., Atkins, D. C., Pullmann, M. D., Friedman, E., & Areán, P. A. (2019). Use of human-centered design to improve implementation of evidence-based psychotherapies in low-resource communities: Protocol for studies applying a framework to assess usability. *JMIR Research Protocols*, 8(10), e14990. <https://doi.org/10.2196/14990>
- Mauricio, A. M., Berkel, C., Gallo, C. G., Sandler, I. N., Wolchik, S. A., Tein, J.-Y., & Brown, C. H. (2017). Concordance between provider and independent observer ratings of quality of delivery in the New Beginnings Program. Annual meeting of the Society for Prevention Research, Washington, DC.
- O'Connell, M. E., Boat, T., & Warner, K. E. (Eds.). (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities*. National Academies Press. <https://www.nap.edu/catalog/12480/preventing-mental-emotional-and-behavioral-disorders-among-young-people-progress>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12(85), 2825–2830. <https://doi.org/10.48550/arXiv.1201.0490>
- Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology*, 75(6), 829–841. <https://doi.org/10.1037/0022-006X.75.6.829>
- Perrin, E. C., Leslie, L. K., & Boat, T. (2016). Parenting as primary prevention. *JAMA Pediatrics*, 170(7), 637–638. <https://doi.org/10.1001/jamapediatrics.2016.0225>
- Prochaska, J. O., & DiClemente, R. J. (1983). Towards an integrative model of change. *Journal of Consulting and Clinical Psychology*, 51(3), 390–395. <https://doi.org/10.1037/0022-006X.51.3.390>
- Proctor, E. K., Powell, B. J., & McMillen, J. C. (2013). Implementation strategies: Recommendations for specifying and reporting. *Implementation Science*, 8(1), 1–11. <https://doi.org/10.1186/1748-5908-8-139>
- Rudo-Stern, J., Dishion, T. J., Aiken, L. S., & Wolchik, S. A. (2016). *Collateral effect of the Family Check-Up on physical activity: A randomized control trial*. Social Science Research Graduate Student Poster Session, Tempe, AZ.
- Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, 38(1), 32–43. <https://doi.org/10.1007/s10488-010-0321-0>
- Shaw, D. S., Connell, A. M., Dishion, T. J., Wilson, M. N., & Gardner, F. (2009). Improvements in maternal depression as a mediator of intervention effects on early childhood problem behavior. *Development and Psychopathology*, 21(2), 417–439. <https://doi.org/10.1017/S0954579409000236>
- Shermis, M., Burstein, J., Elliot, N., Miel, S., & Foltz, P. (2015). Automated writing evaluation: A growing body of knowledge. In MacArthur, C., Graham, S., & Fitzgerald, J. (Eds.), *Handbook of writing research*. Guilford Press.
- Smith, J. D., Berkel, C., Carroll, A. J., Fu, E., Grimm, K. J., Mauricio, A. M., Rudo-Stern, J., Winslow, E., Dishion, T. J., Jordan, N., Atkins, D. C., Narayanan, S. S., Gallo, C., Bruening, M. M., Wilson, C., Lokey, F., & Samaddar, K. (2021). Health behaviour outcomes of a family based intervention for paediatric obesity in primary care: A randomized type II hybrid effectiveness-implementation trial. *Pediatric Obesity*, 16(9), e12780. <https://doi.org/https://doi.org/10.1111/ijpo.12780>
- Smith, J. D., Berkel, C., Jordan, N., Atkins, D. C., Narayanan, S. S., Gallo, C., Grimm, K. J., Dishion, T. J., Mauricio, A. M., Rudo-Stern, J., Meachum, M. K., Winslow, E., & Bruening, M. M. (2018). An individually tailored family-centered intervention for pediatric obesity in primary care: Study protocol of a randomized type II hybrid effectiveness-implementation trial (Raising Healthy Children study). *Implementation Science*, 13(1), 1–15. <https://doi.org/10.1186/s13012-017-0697-2>

- Smith, J. D., Berkel, C., Rudo-Stern, J., Montañó, Z., St. George, S. M., Prado, G., Mauricio, A. M., Chiapa, A., Bruening, M. M., & Dishion, T. J. (2018). The Family Check-Up 4 Health (FCU4Health): Applying implementation science frameworks to the process of adapting an evidence-based parenting program for prevention of pediatric obesity and excess weight gain in primary care. *Frontiers in Public Health*, 6, 293. <https://doi.org/10.3389/fpubh.2018.00293>
- Smith, J. D., Carroll, A. J., Fu, E., & Berkel, C. (2023). Baseline targeted moderation in a trial of the Family Check-Up 4 Health: Potential explanations for finding few practical effects. *Prevention Science*, 24(2), 226–236. <https://doi.org/10.1007/s11121-021-01266-z>
- Smith, J. D., Dishion, T. J., Brown, K., Ramos, K., Knoble, N. B., Shaw, D. S., & Wilson, M. N. (2016). An experimental study of procedures to enhance ratings of fidelity to an evidence-based family intervention. *Prevention Science*, 17(1), 62–70. <https://doi.org/10.1007/s11121-015-0589-0>
- Smith, J. D., Dishion, T. J., Shaw, D. S., & Wilson, M. N. (2013). Indirect effects of fidelity to the family check-up on changes in parenting and early childhood problem behaviors. *Journal of Consulting and Clinical Psychology*, 81(6), 962–974. <https://doi.org/10.1037/a0033950>
- Smith, J. D., Montañó, Z., Dishion, T. J., Shaw, D. S., & Wilson, M. N. (2015). Preventing weight gain and obesity: Indirect effects of a family-based intervention in early childhood. *Prevention Science*, 16(3), 408–419. <https://doi.org/10.1007/s11121-014-0505-z>
- Smith, J. D., Rudo-Stern, J., Dishion, T. J., Stormshak, E. A., Montag, S., Brown, K., Ramos, K., Shaw, D. S., & Wilson, M. N. (2019). Effectiveness and efficiency of observationally assessing fidelity to a family-centered child intervention: A quasi-experimental study. *Journal of Clinical Child & Adolescent Psychology*, 48(1), 16–28. <https://doi.org/10.1080/15374416.2018.1561295>
- van Mourik, K., Crone, M. R., de Wolff, M. S., & Reis, R. (2017). Parent training programs for ethnic minorities: A meta-analysis of adaptations and effect. *Prevention Science*, 18(1), 95–105. <https://doi.org/10.1007/s11121-016-0733-5>
- Van Ryzin, M. J., & Nowicka, P. (2013). Direct and indirect effects of a family-based intervention in early adolescence on parent–youth relationship quality, late adolescent health, and early adult obesity. *Journal of Family Psychology*, 27(1), 106–116. <https://doi.org/10.1037/a0031428>
- Waltz, T. J., Powell, B. J., Fernández, M. E., Abadie, B., & Damschroder, L. J. (2019). Choosing implementation strategies to address contextual barriers: Diversity in recommendations and future directions. *Implementation Science*, 14(1), 42. <https://doi.org/10.1186/s13012-019-0892-4>
- Wiltsey Stirman, S., Marques, L., Creed, T. A., Gutner, C. A., DeRubeis, R., Barnett, P. G., Kuhn, E., Suvak, M., Owen, J., Vogt, D., Jo, B., Schoenwald, S., Johnson, C., Mallard, K., Beristianos, M., & La Bash, H. (2018). Leveraging routine clinical materials and mobile technology to assess CBT fidelity: The Innovative Methods to Assess Psychotherapy Practices (imAPP) study. *Implementation Science*, 13(1), 69. <https://doi.org/10.1186/s13012-018-0756-3>
- Winter, C., & Dishion, T. J. (2007). *Parent consultant log*. Child and Family Center, University of Oregon.
- Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2015). “Rate my therapist”: Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLoS One*, 10(12), e0143055. <https://doi.org/10.1371/journal.pone.0143055>