



Published in final edited form as:

*Adm Policy Ment Health*. 2015 September ; 42(5): 574–585. doi:10.1007/s10488-014-0538-4.

## Blending Qualitative and Computational Linguistics Methods for Fidelity Assessment: Experience with the Familias Unidas Preventive Intervention

Carlos Gallo<sup>1</sup>, Hilda Pantin<sup>2</sup>, Juan Villamar<sup>1</sup>, Guillermo Prado<sup>2</sup>, Maria Tapia<sup>2</sup>, Mitsunori Ogiwara<sup>3</sup>, Gracelyn Cruden<sup>1</sup>, and C Hendricks Brown<sup>1</sup>

<sup>1</sup>Northwestern University, Fienberg School of Medicine, Department of Psychiatry and Behavioral Sciences

<sup>2</sup>University of Miami Miller School of Medicine, Department of Public Health Sciences

<sup>3</sup>University of Miami Miller School of Medicine, Center for Computational Sciences

### Abstract

Careful fidelity monitoring and feedback are critical to implementing effective interventions. A wide range of procedures exist to assess fidelity; most are derived from observational assessments (Schoenwald et al, 2013). However, these fidelity measures are resource intensive for research teams in efficacy/effectiveness trials, and are often unattainable or unmanageable for the host organization to rate when the program is implemented on a large scale. We present a first step towards automated processing of linguistic patterns in fidelity monitoring of a behavioral intervention using an innovative mixed methods approach to fidelity assessment that uses rule-based, computational linguistics to overcome major resource burdens. Data come from an effectiveness trial of the Familias Unidas intervention, an evidence-based, family-centered preventive intervention found to be efficacious in reducing conduct problems, substance use and HIV sexual risk behaviors among Hispanic youth. This computational approach focuses on “joining,” which measures the quality of the working alliance of the facilitator with the family. Quantitative assessments of reliability are provided. Kappa scores between a human rater and a machine rater for the new method for measuring joining reached .83. Early findings suggest that this approach can reduce the high cost of fidelity measurement and the time delay between fidelity assessment and feedback to facilitators; it also has the potential for improving the quality of intervention fidelity ratings.

### Introduction

A current finding in the emerging field of implementation science (Chambers 2012; Landsverk et al., 2012; Aarons, Hurlburt, & Horwitz, 2011) is that evidence-based interventions need to be delivered with precision, or fidelity, in order to achieve the level of

Corresponding author: Carlos G. Gallo, Department of Psychiatry and Behavioral Sciences, 680 N. Lake Shore Drive, Suite 1400 Chicago, IL 60611. carlos.gallo@northwestern.edu.

Co-Author contacts, in order listed: hpantin@med.miami.edu, jvillamar2@med.miami.edu, gprado@med.miami.edu, mtapia@med.miami.edu, mogihara@med.miami.edu, gcruden@med.miami.edu, chbrown@med.miami.edu

effects in large-scale implementation that were previously obtained in efficacy or effectiveness research trials (Durlak & DuPre, 2008; Allen, Linnan, & Emmons, 2012). Fidelity monitoring and feedback are critical parts of implementing effective behavioral interventions (Poduska et al., 2009) for without these we cannot expect acceptable delivery or intended outcome of high quality programs (Schoenwald et al., 2011). These challenges are central to implementation research, which involves “the use of strategies to adopt and integrate evidence-based health interventions and change practice patterns within specific settings” (Chambers 2008). Key determinants of the fidelity assessment process in practice includes the capacity of community based organizations, or community or state level service agencies to use such a measure for monitoring and feedback (Landsverk et al., 2012). Thus, fidelity measurement systems must be both readily available and responsive to the capacity of state level agencies and community service providers engaged in implementation.

In general terms, fidelity is the extent to which an intervention is delivered as intended. Schoenwald and colleagues (2011) describe fidelity as composed of three elements: adherence (interventionist adherence to an intervention), competence (interventionist competence), and differentiation (intervention differentiation). There is an overabundance of fidelity rating systems now being used in many behavioral interventions delivered in mental health and other social service settings (Schoenwald et al., 2011). A recent review of the fidelity measurement literature (between 1998 and 2008) identified 249 unique adherence measurement methods in 304 studies of psychosocial interventions (Schoenwald & Garland, 2013). These methods rely on direct observation of supervisors, review of videotapes and audiotapes, and even self-reports of the interventionist or facilitator. Most of these methods require a major commitment in resources and produce a time lag from the time that the rating takes place and the time the feedback is given to the facilitator.

In the prevention field, the host organization for such program implementation is often a school or community based organization, and while such prevention programs may support its overall mission, e.g. a drug abuse prevention program in schools, the host organization often does not have sufficient support for fidelity monitoring and supervision during implementation. This contrasts with effectiveness trial research projects in which a research partner or the program's purveyor generally serves to provide this monitoring, feedback, and supervision support. Fidelity monitoring in an effectiveness trial would routinely be paid by grant funds that pay for the trial. However, for implementation research or practice, the fidelity monitoring and feedback system would only be sustainable if supported outside of the grant funding mechanisms. When faced with the choice, host organizations are more likely to use methods that match the available resources, which often prohibit full fidelity monitoring, and as a consequence they often times fail to achieve the desired outcomes in the populations they serve (S. K. Schoenwald et al., 2008; Fixsen, Naoom, Blase, Friedman, & Wallace, 2005; Real & Poole, 2005). Thus, for research on implementation, we would ultimately seek to develop cost-effective and accurate fidelity monitoring systems that can easily be embedded within the available community or service systems. Hanson et al. (2013) found that implementation interventionists recognize the difficulty in maintaining fidelity and acknowledge the high resource intensity of effective intervention implementation, leading them to advocate for the increased use of technology to monitor and reduce drift. In

addition to accurate fidelity monitoring, usability requires that we would have a minimum time lag between fidelity rating and feedback to the facilitator.

Brown and colleagues (2013) propose exploring innovative methodology, such as the use of computational technology, to address these and other fundamental challenges in implementation. Our main objective in this paper is to examine whether computational approaches, and particularly computational linguistics, can be used to support automated monitoring of fidelity data in an effectiveness trial of Familias Unidas, an evidence-based preventive intervention being delivered via school counselors to Hispanic families (Prado & Pantin, 2011). Ours is a proof of concept approach, examining the extent to which one computational component applied to videotapes from the Familias Unidas family-based, adolescent substance abuse and HIV prevention intervention, shows sufficient reliability to recommend further work. We specifically examine whether machine-scored (i.e., computer generated) fidelity scores are reliable against human-scored fidelity ratings.

We first describe the Familias Unidas preventive intervention and its evaluation in an effectiveness trial. We then describe the specific challenge of obtaining cost effective, valid, and reliable fidelity ratings in Familias Unidas. In this paper we focus on one key aspect of Familias Unidas' assessed fidelity called joining, a characteristic of competence that is described below. We then develop a micro-level coding system to assess facilitator joining quality that can be assessed by both humans and through a computer algorithm, and compare the reliability of machine versus human coding on this construct. Because computational approaches, including computational linguistics, are likely to be unfamiliar to many readers, we provide a short background to this field, followed by a rationale explicating why computational approaches could be useful for fidelity monitoring. Further, we relate these computational approaches to mixed methods research. Finally, we describe what other components would be required for a fully developed automated fidelity rating system, as well as what challenges such a system would face in wide-scale implementation.

### **Familias Unidas Effectiveness trial: Joining, Facilitators**

Familias Unidas (Prado & Pantin, 2011) is a Hispanic specific, family-based preventive intervention guided by eco-developmental theory (Szapocznik & Coatsworth, 1999). The intervention is designed to reduce risk for behavioral problems, substance use, and risky sexual behaviors in Hispanic adolescents by improving family support of the adolescent, parental involvement, parental monitoring of peer and school activities, as well as parent and adolescent effective communication. The Familias Unidas intervention has been evaluated in three completed randomized clinical trials and found to be efficacious in reducing substance use and sexual risk behavior among Hispanic youth (Pantin, Schwartz, Sullivan, Prado, & Szapocznik, 2004; Pantin et al. 2009; Prado et al., 2007; Prado et al., 2012). Familias Unidas is delivered through eight family-centered, multi-parent groups and through four family visits that place parents in charge and in a role that can promote change in their families and adolescents.

In these sessions, parents most often communicate in Spanish while the adolescent often converses in English, so facilitators often speak in both Spanish and English in the same session. In order to assess implementation fidelity in the Familias Unidas program, a fidelity

rating system of randomly selected videotapes of family visits was implemented (Prado, Pantin, Schwartz, Lupei, & Szapocznik, 2006). This rating system has been used in the previous trials (Prado et al., 2007). To assess fidelity, 10% of all family visits (N = 111) in the effectiveness trial were selected randomly for videotape rating by adherence raters. Raters used a standard adherence form to record the presence or absence of prescribed (e.g., joining) and proscribed (e.g., acts as a switchboard and/or speaks for long periods of time) facilitator behaviors. Adherence raters were trained to achieve an interrater reliability (intraclass correlation) of .80 or above to a senior Familias Unidas rater, the “gold standard.” Interrater reliability between the adherence raters and with the Gold Standard was reevaluated monthly to control for rating drift, and any adherence problems identified by raters were discussed with the Principal Investigator and Clinical Supervisor in weekly intervention integrity meetings. When adherence ratings fell below 70% for 2 out of 4 consecutively rated sessions, the information was brought to the attention of the clinical supervisor who met with the facilitators and actively retrained.

### **Familias Unidas and the Joining Process**

In this paper we discuss automating one component of “joining,” a key dimension of Familias Unidas’ fidelity assessment. The joining process in the field of behavioral mental health, also commonly referred to as the therapeutic alliance or working alliance (Minuchin, 1974), is considered a strong predictor of treatment adherence and outcome (Barber et al, 2006). Minuchin and Fishman (1981) elegantly described “joining” as “the glue that holds the system [family/individual-interventionist relationship] together.” It is of vital importance to effectively join and form a strong therapeutic system in order to affect change in the family or individual. Research on the statistical power of the joining process to engage individuals and families into treatment reflects more than 1000 findings that support this notion (Orlinsky, Ronnestad, & Willutski, 2004).

In the Familias Unidas intervention, joining includes the following components: (a) Facilitator communicates acceptance, respect, and trust to all family members; (b) Facilitator uses humor; (c) Facilitator encourages family to disclose anecdotes; (d) Facilitator addresses individuals’ statements/concerns; (e) Facilitator asks an open-ended question; (f) Facilitator validates family members. Joining is measured on a numerical score (0 to 6) for each 30-minute segment of the Familias Unidas family visits. In an efficacy study, Prado et al. found that joining was directly linked to engagement and retention of families into the intervention (Prado, Pantin, Schwartz, Lupei, & Szapocznik, 2006). Consequently, the Familias Unidas intervention places strong emphasis on joining as a key component of the fidelity rating process, closely monitoring facilitators’ progress in engaging and retaining families into the intervention.

### **Challenges in Fidelity Assessment of Familias Unidas**

While considering the benefits of computational linguistics in fidelity monitoring, we have three motivating principles to guide our computational design: cost, speed, and rating accuracy. In terms of costs, standard fidelity coding of each 45-minute family session of the Familias Unidas intervention takes approximately two hours to complete by the traditional human based method. In the current trial there were 376 families scheduled for 4 family

visits. Thus, the direct cost for coding would exceed \$90,000. Such costs would be prohibitive if implemented in a public school system.

With regard to speed, even when financial resources are available for a system based entirely on human effort, there is considerable time investment required for rating training, calibration, and other tasks. One major recurrent time delay is that between the fidelity assessment of a session and the feedback given to each Familias Unidas facilitator. In the current effectiveness trial, fidelity ratings are typically completed within one week following the intervention session. The feedback is then given to the supervisor, who compiles it with other measures of adherence such as attendance and clinical issues, and delivers the feedback to facilitators during weekly supervision sessions. Several weeks may elapse by the time facilitators receive clinical feedback on a particular intervention session, at which point additional sessions may have occurred that repeat the same clinical delivery concerns. A computational based approach has the potential to reduce this time lag, narrowing the gap closer towards real time, thus allowing session facilitators to receive feedback quickly and potentially thus rectify fidelity concerns.

In terms of rating accuracy, coders can drift in their ratings over time. This rating bias is created because the fidelity raters may not be chosen with equal rigor or may not have the same level of supervision as that within the research teams. In addition, standard methods of measuring fidelity may be biased in the dissemination trials, because the raters may be close colleagues of the facilitators and may or may not feel comfortable with rating their peers. A computational approach removes this potential bias introduced in the intervention dissemination at the school districts or other community settings.

### **What can computational linguistics contribute to Fidelity Monitoring?**

Computational linguistics allows us to recognize spoken words (Holmes and Holmes 2002), ask questions in natural language to search databases (Popescu, Etzioni, & Kautz, 2003), and create approximate translations of text (Lopez 2008) using computer algorithms that can recognize linguistic patterns, a process that previously required human level intelligence. In this paper we present our findings using a system called FARE (Fidelity Automatic RatEr) as a proof of concept for automating measurement of joining in the Familias Unidas intervention. Joining is a complex interactional behavior to assess; raters are trained to pay attention to verbal as well as nonverbal interactions. While nonverbal cues can be detected by computational means (Inoue, Ogihara, Hanada, & Furuyama, 2010), the current approach is limited to transcribed speech of the facilitator only, thus severely limiting the information available for the computational algorithm, but making the computational task sufficiently “simple” to attack even after taking account of the fact that a facilitator may speak in both Spanish and English, sometimes in the same sentence. To the best of our knowledge, this is the first work that uses speech analysis, knowledge engineering, and computational linguistics to measure a component of fidelity.

In this single component of FARE that is described here, the computational system uses a transcribed text to rate facilitators’ utterances (in Spanish and/or English) as input, then applies a decision tree algorithm that categorizes linguistic patterns associated with high or low fidelity on the joining dimension. These linguistic patterns were developed by a process

of knowledge engineering alongside experts on fidelity to the Familias Unidas intervention. While some human based fidelity ratings are coarse, with only a single measure per session, the ratings obtained with FARE are more akin to micro-coding of each spoken contribution made by the facilitator.

### **What does this computational approach have to do with mixed methods?**

Computational approaches are often seen as purely quantitative, i.e., manipulating numbers. But computers also manipulate symbols through well-defined rules (Huth 2004). It is this second use that we describe as we examine fidelity assessment. Symbol manipulation underlies all of natural language processing (NLP), and as the objective is to extract meaning from sentences, this process more closely resembles ethnographic and other types of qualitative analysis that imposes rules so that reliable meaning can be abstracted. Indeed, one could view a computational algorithm that is based on rules to extract information about “fidelity” as a type of qualitative analysis of written text where all human judgment is replaced by clearly stated rules that are followed to the letter. By combining these algorithmic approaches with quantitative analysis of reliability, the methods discussed in this paper represent a mixed methods approach (Palinkas et al., 2011).

Additionally, several methods described in this paper that rely solely on human effort can be considered qualitative methods. The first of these is the elicitation of critical elements behind fidelity, which were obtained through interviews with the Familias Unidas efficacy and effectiveness trials clinical supervisor and the Familias Unidas program developer. While we did not follow formal procedures for taping and extracting these interviews, the method used did closely follow knowledge extraction procedures, an engineering procedure often used to develop a class of artificial intelligence known as expert systems (Bahrammirzaee 2010).

## **Methods**

### **Population and Sampling**

Data for this project are based on an ongoing randomized trial evaluating the relative effectiveness of the Familias Unidas preventive intervention (R01DA025192, NIDA) (Prado et al., 2012; Prado et al., 2013; Prado et al. 2007). In this effectiveness trial, the intervention is delivered by school counselors who serve as facilitators and deliver the program to parents and adolescents, with training and supervision by the research team. School counselors carry out the delivery of this intervention outside of their normal responsibilities in the school system. Seven hundred and forty-six parent-child dyads were recruited and randomized to one of two study arms, Familias Unidas (N=376) or control (N=370). Participants in this study were assessed at baseline, randomized, and reassessed at 6 and 18 months, with the last assessment scheduled to occur at 30 months post baseline. For the intervention group, the Familias Unidas program was delivered through 8 parent group sessions and 4 family visits. A total of 24 groups consisting of 15 parents on average were conducted over the course of two years. All family visits and parent group sessions were led by one of 27 trained Familias Unidas facilitators.

During the family visits, parents had an opportunity to practice the skills learned during group sessions with the help of the facilitator. Family visits were videotaped with participants' consent and lasted approximately 45 minutes to 1 hour. Fidelity scores on several dimensions were obtained by viewing the first half-hour of these tapes and rating the facilitator on 7-point scales, described in more detail below. One of the dimensions scored is joining, and we refer to these scores as traditional joining scores to distinguish them from the micro-level joining scores we developed here. Traditional joining scores, which are described in more detail below, were available on facilitators conducting 111 family visits. For the 88 families assessed in these 111 family sessions, the average total intervention attendance (8 parent group sessions and 4 family visits) was 9, ( $SD= 2.5$ ).

**Familias Unidas Video Selection**—To conduct this automated fidelity study, we selected 33 of the 111 family visit videos for further coding by humans and machine using a new micro level coding system for joining. We selected this subset of family visits to maximize the variation across the following five dimensions: (1) fidelity rating for joining dimension in the first thirty minute segment of video (score range 0 to 6), (2) visit placement in four session sequence (1-4), (3) Familias Unidas facilitator conducting the family visit, (4) number of individuals in the session (between two and five people attending the session plus the facilitator), and (5) the language spoken in session (Spanish only, English only, or Spanish and English). A total of 20 out of 27 facilitators and 31 of 88 parent-child dyads are represented. In this subset of 33 family visits, the average total attendance (8 parent group sessions and 4 family visits) was 8.8 ( $SD= 2.8$ ), and the family visit joining score average was 4 ( $SD=.6$ ). Thus, this sample is representative of the full sample of visits.

## Fidelity Rated by Humans

### Traditional Joining Ratings

**Session-Level Coding:** The traditional system for rating fidelity in Familias Unidas was used to maintain adherence to the Familias Unidas program and for supervision during the effectiveness trial. Four paid, master's level research assistants rated each video in the data set for the prescribed behavior of the facilitator. In order to measure fidelity to the Familias Unidas model, the raters were trained extensively by a senior Familias Unidas expert rater (~5 years rating experience) for a total of 3 days, with several supervised rating sessions. The raters rated this prescribed facilitator behavior (i.e., joining) on an extensiveness/quality rating ranging from “0 = not at all/very poor” to “6 = extensively/excellent.” Videos were rated in 30-minute segments. Raters had to achieve an inter-reliability (intraclass correlation) of .80 or greater with the senior rater before rating sessions on their own. The average family visit joining score was 4, ( $SD=.6$ ). Thus, these traditional joining ratings were nearly all in the good to excellent range.

**Human Utterance-Level Coding of Joining**—In contrast to the more molar, session-level, traditional joining score described above, an utterance-level joining manual for the Familias Unidas intervention was developed with direction from the Familias Unidas program developer and the Familias Unidas clinical supervisor for the effectiveness trials so that it would be similar to and theoretically comparable to the traditional session level joining measure, but applicable to each facilitator utterance. An utterance is defined as a

sequence of contiguous sentences voiced by the facilitator and separated by verbalizations from family members. These utterance scores could then be combined over the session to make a more molar score. We developed coding instructions for humans to assess the quality of these utterances in an instruction manual similar to the traditional manual that assesses the prescribed behavior of the facilitator. Instead of a single score on each fidelity dimension per 30-minute segment in the traditional manual, we instructed coders to assess each utterance of the facilitator and assign a fidelity score. This molecular coding of the intervention's fidelity has the potential to yield a finer detail of behavior and pinpoint correctives during supervision (Busch, Kanter, Callaghan, Baruch 2009).

The first step in utterance level coding was to create written transcripts. While computational approaches can be used to generate written transcripts, and this will ultimately be included in the FARE system, we used human generated transcriptions during this proof of concept. The data set of 33 videos was transcribed by a paid master's level research assistant fluent in both Spanish and English. This person was instructed to transcribe speech from all individuals that appeared in the first thirty-minute segment of the family visit. The transcriber was instructed to ignore repeated words and to write words as intended by the speaker, a standard procedure in human transcriptions so that each word is orthographically correct (Gallo 2010; Gallo 2007). Transcripts were divided into utterances, which are assemblies of contiguous sentences by one speaker. A total of 86,000 words (4,300 utterances) were transcribed. An average session contained 2,618 words and 128 utterances. An average facilitator spoke for 60 utterances; an average parent spoke for 46 utterances; and an average adolescent spoke for 22 utterances. There were a total of 2052 utterances spoken by the facilitator.

We focus in this paper on the joining sub-dimension (e) Facilitator asks an open-ended question; open-endedness invites families to express feelings and thoughts freely, enhances engagement and promotes deeper conversations rather than closed-end questions or statements (Overholser 1995). Clinical knowledge informed systematic rules on how to rate each utterance; i.e., distinguishing what constitutes an open-ended question. Each utterance was rated accordingly as one of the following: *Not Relevant*, *Improbable*, or *Good*. A statement is labeled **Not Relevant** under this dimension if the utterance is a factual statement or a rhetorical question (e.g. “Yo soy nueva aqui” – I am new here; “¿Qué piensas de este clima? Por qué a mi no me gusta” – What do you think of this weather? Because I don't like it). A statement is labeled **Improbable** if it contains one of the following: “Dónde, Cuándo, Cuál, Cuántas, Con qué frecuencia” – Where, When, Which, How many, How often (e.g. “¿Dónde jugaste basket?” – Where did you play basketball?; “¿Qué tan frecuente juegas basket?” – How frequently do you play basketball?; “¿Cuál de tus amigos te gusta mas?” – Which of your friends do you like best?). An utterance is rated as **Good** if it is an open-ended question. It can contain words such as “Qué, Cómo, Por qué” – How, What, Why (e.g. “¿Qué piensas sobre lo que tu hija hizo?” – What do you think about what your daughter did?).

## Machine Utterance-Level Coding of Joining

Automated coding for Good, or open-ended questions, was obtained at the utterance level using the same input, i.e., transcripts of facilitator speech. We developed a system based on expert knowledge that uses linguistic patterns related to joining. We used 193 utterances (9% of 2052 total facilitator utterances) to develop the rules of the system. We developed a decision tree that assigns each utterance a score for open-ended questions. This decision tree was developed based on the patterns uncovered by the expert clinicians involved in the project. Figure 1 displays the decision tree for the joining sub-dimension of open-ended questions. The top diamond is the entry point, and each diamond represents a query to categorize the utterance under review to be assigned one of the three possible labels (*Not Relevant*, *Improvable*, *Good*). The query is based on a regular expression implemented in the Perl programming language (Wall, Christiansen, & Orwant 2000). A regular expression is recognized as a What/Why question, for example, by recognizing punctuation that signifies a question and recognizing question keywords such as “Por qué/Qué” – What/Why. Once we recognized those utterances as What/Why questions, we coded this utterance as Good.

## Analysis

**Ratings**—Our analyses involved reliability comparisons of 1) human-machine coding of human utterance-level assessments of open-ended questions, 2) human-human coding of these same utterance-level assessments; and 3) comparison of these same utterance-level scores that are aggregated at the session level.

**Reliability of Utterance Level Coding**—In order to establish reliability in our coding at the utterance level, we computed kappa scores (Carletta 1996) among the two humans and machine rater using three-levels (*Not Relevant*, *Improvable*, *Good*). We also report reliability at a two-level scale where we merged *Improvable* and *Good* categories together. A kappa score allows us to measure how much the agreement between the labels assigned by two different raters exceeds that of chance agreement.

**Correlations between raters at the session level**—In practice we would aggregate the scores from the micro level (utterance-level) to the molar level (session-level). Thus, we tested reliability of the computed-based rater against human raters by using Pearson correlations to compare two types of aggregate indicators. The first indicator type involved the total number of utterances that were relevant to open-ended questions (i.e. binary, improvable, and good questions and their sums). The second indicator type involves a quality score equal to a weighted average where binary questions were scored 0, improvable questions were scored 1 and good questions were scored 2.

## Results

### Utterance-Level Coding

In the 33 transcripts that were coded by a human, the average session contains 1484 *Not Relevant*, 375 *Improvable*, and 158 *Good* utterance levels. Similarly, there were 1262 *Not*

*Relevant*, 408 *Improvable*, and 304 *Good* utterance labels identified by the machine rater. Thus, the machine rating had a distribution similar to a human coder.

### Reliability of Utterance Level Coding

We computed the agreement among human and machine raters first for the three level coding: *Not Relevant*, *Improvable*, *Good*. The kappa between human and the machine ratings was low, .43. However, on the two-level score where *Improvable* and *Good* were combined, reliability was much higher. The kappa between human rater and machine rater was .83. Regarding the aggregated scores, the correlation between human and machine raters was .84 (total number of utterances were 2052).

We also evaluated our system by finding the correlation in two different ways shown in Table 1, which focuses on different ways to decompose the total counts of the different categories and Table 2, which focuses on overall quality measures. In the first table, we added the label of each utterance to obtain a session level sum for all categories, and compared their relationship between raters using the Pearson correlation. For instance, the correlation between the sum of all questions for the human rater 1 and rater 2 is .95. Correlations between machine and human ratings were acceptably high, above .77, for the sum across all questions or relevant questions, but a few correlations were low when looking at each individual category. Figure 2 shows that the number of questions identified in a session by the machine are almost never less than that for either human rater.

In Table 2, we computed the total quality score which is equal to the sum of all categories as follows: binary questions times zero, plus improvable questions times one, plus good questions times two.

In Table 2 we found that total quality scores between the machine and humans are strongly correlated although not as high as they are between humans (based on a small number of sessions). However, the averaged quality scores for the machine had much low reliability than they did for the total quality scores or the two humans. In Figure 3 (A) we note the strong correlation of total quality score between human raters, and in a similar fashion, Figure 3 (B) shows the strong correlation between each human rater with the machine rater.

### Discussion

In this paper we presented an utterance level measure of open-ended questioning, a major component of joining that can be coded computationally. We tested whether a machine rating of open-ended questioning could compare with similar human ratings on this same utterance level measure. Machine ratings of relevant utterances were reliable when compared to human ratings when this measure was dichotomized, but less reliability was achieved when distinguishing *Improvable* versus *Good* ratings. This lowering of reliability between machine and human with the three category outcome is not surprising given that the inherent challenges in distinguishing improvable versus good utterances. Reliability was much stronger when the utterance level measures were aggregated to the session level. Figure 2 (A) provides evidence that human raters agree with each other with respect to the number of questions labeled at the session level. Figure 2 (B) demonstrates that the machine

rater has the highest recognition of questions than either of the human raters. Figure 3 We found high correlation between aggregate scores at the three category and two category outcomes (highest .98 between the human rater 1 and rater 2, and .79 and .82 against the machine rater). Together, these results suggest that in some instances machine ratings can be similar to human ratings, thus providing initial evidence of our proof of concept. We had only a few sessions where both human raters provided scores that could be compared directly, and in these instances the reliability was high.

In this paper we demonstrated a way to quantify joining using a computational method in a complex behavioral intervention that is delivered to families in the home in Spanish and English. Our work focused on joining as measured by open-ended questions. We chose only to focus on utterances by the facilitator in determining this open-endedness, recognizing that some information would be lost by not attending to the family's response. This focus only on facilitator verbalizations without attending to the content reflects some, but not all of a facilitator's competence on this sub-dimension. In the Familias Unidas intervention, the other dimensions of joining include: the facilitator's ability to validate family members; the facilitators' use of humor; facilitators' encouragement of family members to disclose anecdotes; the facilitators' ability to communicate trust; and the facilitators' ability to address family members' concerns (Prado, Pantin, Schwartz, Lupei, & Szapocznik, 2006). Most, if not all behavioral interventions include a component of joining. Our method has the potential to be applied to other interventions. We have not yet investigated ways that these other dimensions can be rated computationally, and their contribution may improve the reliability and validity of the overall joining score. Furthermore, there remain other proxies for fidelity that need to be tested. For instance, if facilitators actually do facilitate communication between parents and adolescents, such a measure may increase our prediction of participation and attendance. Two proxies to this behavior that can be automated are the ratio of number of words spoken by the facilitator to family members to the number of turns taken between the adolescent, parent, and facilitator. Taken together, these features can be processed automatically and be evidence of engagement based on multiple dimensions of linguistic behavior. Another future avenue of investigation will analyze the valence of the response to the question posed by the facilitator. Future work includes the improvement of each critical step, and a model to integrate the output of these limitations.

Further, there are clearly major technical challenges left to address in developing an automated system, and some of these steps may affect the overall quality. In this first proof of concept project, we note that the input was based on human coded transcripts, a component that would obviously need to be replaced with an automated system. The success of distinguishing different speakers from one another and producing an accurate transcript depends on the quality of the audio signal that is available, and improvements in audio signals recording beyond the one-source videos that were used here will no doubt increase accuracy of these steps.

At the same time, there are reasons to believe that an automated system could ultimately exceed the quality of human ratings of fidelity that are now being used. First, a static computational system has perfect reliability, since the same input processed by the same

program will always produce the same result. Secondly, a computational approach also offers a capacity to learn and improve over time (De Cooman & Zaffalon 2004). The field of machine learning, which updates its own decision-making as additional data are made available, provides an innovative tool to improve the quality of fidelity assessment over time. Third, a molecular level rating system, such as the utterance level approach described here, can focus supervision on specific instances where fidelity can be improved. Also, more work needs to be done to compare the predictive validity of machine generated ratings against that of humans. We anticipate that as the computational methods approach to understanding fidelity becomes more sophisticated through refinement of rules, inclusion of more computational methods and machine learning, this approach will increase our ability to monitor fidelity while a prevention or treatment intervention is implemented in the field.

However, we do not believe that an automated system should completely replace the use of human fidelity ratings. Indeed, we suggest that fidelity monitoring can be improved through a true “mixed method” two-stage approach. An automated system can be used as a first stage; it is not only inexpensive, real-time, and reliable, but it can be used to screen audio transcripts into three broad categories: one where the automated rating of facilitator fidelity is high and we have high confidence of this rating, one where it is low with high confidence, and a third, middle category where our confidence about this automated rating is low. We can then use statistical sampling techniques, coupled with modeling of these ratings over time and client, to select an informative subset of passages, sessions, or facilitators for further human assessment. This information would then be used in a feedback loop to provide selective supervision of facilitators and around topics that are most challenging. In the typology of Palinkas et al. 2011, this computational/human hybrid approach most closely resembles an important mixed model approach, which involves a “Sequential collection and analysis of quantitative and qualitative data (quant --> Qual)” category.

Ultimately, the success or failure of the current “evidence-based approach” to improving mental health and reducing substance abuse and HIV/AIDS will depend in part on our ability to monitor and use high quality fidelity information. Schools and community-based organizations that are challenged with many other responsibilities besides delivery of these prevention programs will require ongoing technical support in order to sustain these programs. Taking an important role in this prevention support system (Chinman et al., 2008) will be the state's public health, social service, and educational systems, which have strategic reasons to partner with researchers in implementation science (Brown et al., 2012). Parent prevention programs such as Familias Unidas and Triple P (Herschell 2010; Prinz, Sanders, Shapiro, Whitaker, & Lutzker, 2009), as well as elementary schools programs to reduce aggressive behavior (Kellam et al., 2011), are already being widely implemented and may be good candidates for statewide support.

Prior research on prevention and treatment programs has included different methods for identifying fidelity to a particular intervention (Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 1998; Hogue et al., 2008). All of these methods are costly, time consuming, and could pose a challenge for a supervisor or rater that does not speak the particular language used in a session. In the innovative mixed method approach to fidelity presented here, we not only address the cost associated with monitoring fidelity, but also attempt to

address the linguistic barrier to implementation fidelity, which could be extremely advantageous towards addressing health disparities through broader implementation of interventions across groups. One of the notable requisites for large-scale implementation is maintaining fidelity (Henggeler, Schoenwald, Liao, Letourneau, & Edwards, 2002; Liddle et al., 2006). The computational linguistic model has the ability to rate sessions during which two languages are spoken, and provide feedback to a supervisor in the language preference of her choice. Thus, this methodology could help conduct fidelity monitoring by allowing either a mono or bi-lingual supervisor to provide feedback to intervention facilitators, regardless of which language was used in delivering the intervention. While addressing health disparities will still require the recruitment of bilingual facilitators, computational linguistic methods could potentially help reduce the number of implementation team members, particularly fidelity raters, who require bilingual skills, a potential barrier to intervention dissemination and multi-site implementation (Suarez-Morales et al., 2007). Lacking this, some research teams may be forced to withdraw implementation efforts due to an inability to devote adequate resources toward bilingual fidelity monitoring. Such flexibility will offer host organization's more autonomy and increased resource capacity to monitor and effectively implement programs, while also providing more information that can be used to monitor outcomes and inform and encourage future intervention dissemination efforts.

The computational method proposed here could enhance the quality of outcomes in implementing evidence-based interventions internationally, when a different language is spoken. Many facilitators have been trained outside of the U.S. to provide empirically validated programs such as Brief Strategic Family Therapy (BSFT; Szapocznik, Hervis, & Schwartz, 2003), functional family therapy (Alexander, Pugh, Parsons, & Sexton, 2000), multidimensional family therapy (Liddle 2002), and multi-systemic therapy (Henggeler & Borduin, 1990). In such cases, the supervision process requires that facilitator's video recordings be translated into English before the supervisor conducts the review of those translated tapes. This process is extremely costly and time consuming. It could potentially create a barrier in the way the process can be lost in translation and interpreted by the supervisor in order to provide good quality feedback of the session and improve clinical outcomes (Rowe et al., 2013).

In this paper we presented a mixed method, proof of concept approach to measure fidelity of an effective behavior intervention. The use of computational linguistics to develop an automated rating system for fidelity presents a viable path for addressing implementation challenges in host organizations (i.e. state level agencies, community organizations). As an implementation tool, an automated fidelity rater may eventually be paired with an effective behavioral intervention, such as Familias Unidas, allowing measurement of fidelity in host organizations.

## Acknowledgments

We acknowledge support from the Center for Computational Science at the University of Miami (CG, MO, CHB) and the National Institute on Drug Abuse (NIDA) for this work through the Center for Prevention Implementation Methodology for Drug Abuse and Sex Risk Behavior, P30DA027828 (HP, JV, GP, MO, CHB) and Familias Unidas Stage III Study: Preventing Substance Abuse in Hispanic Youth, R01DA025192 (HP, GP, MT). This work

was supported by National Institute of Allergy and Infectious Diseases Grant P30AI073961 (GP, HB, MT) and Centers for Disease Control and Prevention Grant U01PS0000671 (GP, HP, MT). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

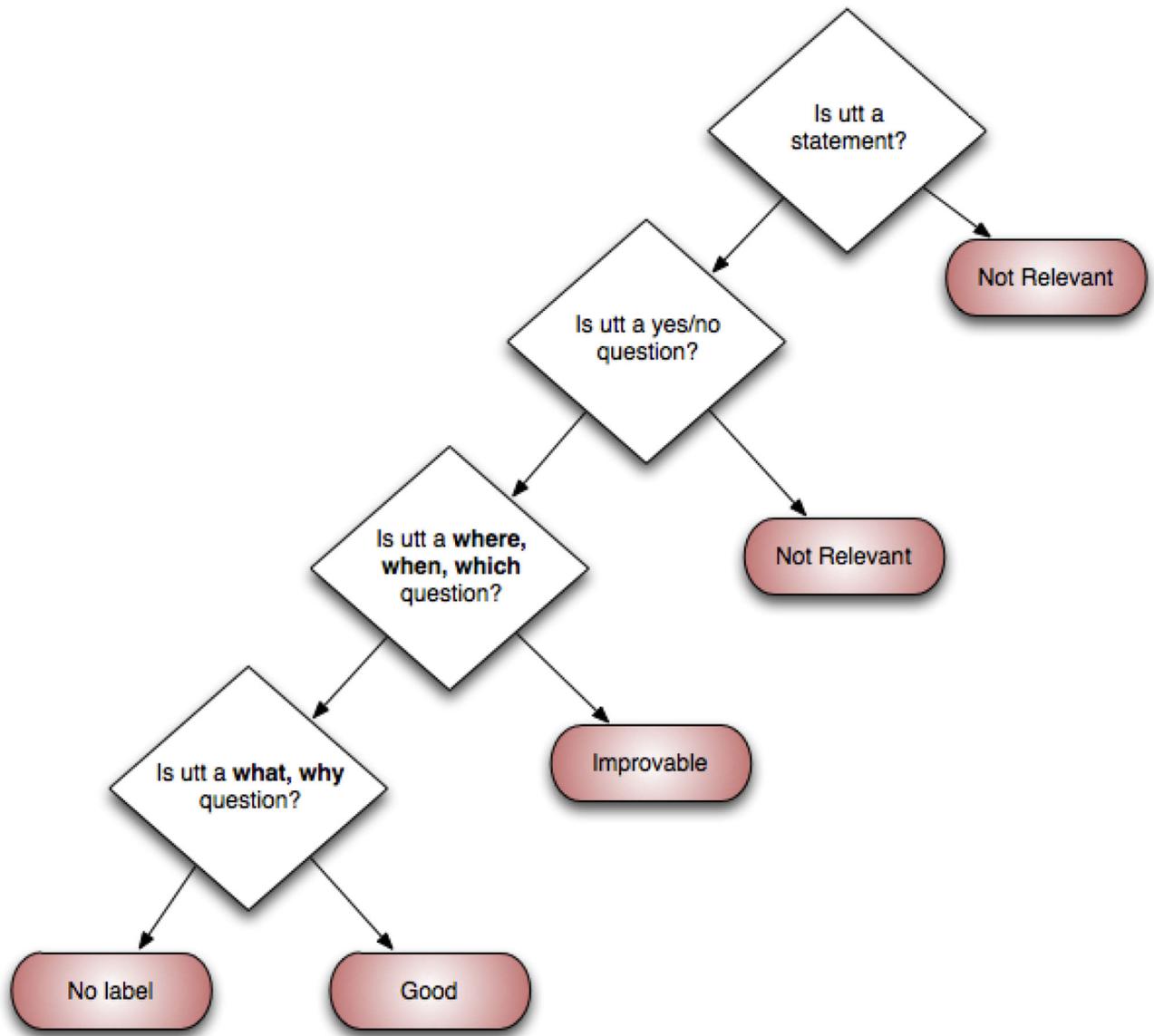
## References

- Aarons GA, Hurlburt M, Horwitz SM. Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health*. 2011; 38(1): 4–23. doi:10.1007/s10488-010-0327-7. [PubMed: 21197565]
- Alexander, J.; Pugh, C.; Parsons, B.; Sexton, T. Family functional therapy.. In: Elliot, DS., editor. *Blueprints for Violence Prevention: Book Three*. Center for the Study and Prevention of Violence; Boulder, CO: 2000. p. 445–453.
- Allen, JD.; Linnan, LA.; Emmons, KM. Fidelity and its relationship to implementation effectiveness, adaptation, and dissemination.. In: Brownson, RC.; Colditz, GA.; Proctor, EK., editors. *Dissemination and implementation in health: Translating science to practice*. Oxford University Press; New York: 2012. p. 281–304.
- Bahrammirzaee A. A comparative survey of artificial intelligence applications in finance: Artificial neural networks, expert system and hybrid intelligent systems. *Neural Computing and Applications*. 2010; 19(8):1165–1195. doi:10.1007/s00521-010-0362-z.
- Barber JP, Gallop R, Crits-Christoph P, Frank A, Thase ME, Weiss RD, Connolly Gibbons MB. The role of therapist adherence, therapist competence, and the alliance in predicting outcome of individual drug counseling: Results from the NIDA collaborative cocaine treatment study. *Psychotherapy Research*. 2006; 16:299–240.
- Becker D, Hogue A, Liddle HA. Methods of engagement in family-based preventive intervention. *Child and Adolescent Social Work Journal*. 2002; 19(2):164–179.
- Brown CH, Mohr DC, Gallo CG, Mader C, Palinkas LA, Wingood G, Prado G, Poduska J, Gibbons RD, Kellam SG, Pantin H, McManus J, Ogihara M, Valente T, Wulczyn F, Czaja S, Sutcliffe G, Villamar J, Jacombs C. 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(Supplement 1):S72–S84. (In press).
- Brown CH, Kellam SG, Kaupert S, Muthén BO, Wang W, Muthén L, McManus J. Partnerships for the design, conduct, and analysis of effectiveness, and implementation research: Experiences of the prevention science and methodology group. *Administration & Policy in Mental Health*. 2012; 39(4): 301–316. doi:10.1007/s10488-011-0387-3. [PubMed: 22160786]
- Busch AM, Kanter JW, Callaghan GM, Baruch DE, Weeks CE, Berlin KS. A micro-process analysis of functional analytic psychotherapy's mechanism of change. *Behavior Therapy*. 2009; 40(3):280–290. doi:10.1016/j.beth.2008.07.003. [PubMed: 19647529]
- Carletta J. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*. 1996; 22(2):249–254.
- Chambers DA. Advancing the science of implementation: A workshop summary. *Administration and Policy in Mental Health and Mental Health Services Research*. 2008; 35(1-2):3–10. doi:10.1007/s10488-007-0146-7. [PubMed: 18040772]
- Chambers D. The interactive systems framework for dissemination and implementation: Enhancing the opportunity for implementation science. *American Journal of Community Psychology*. 2012; 50(3-4):282–284. doi:10.1007/s10464-012-9528-4. [PubMed: 22688847]
- Chinman M, Hunter S, Ebener P, Paddock S, Stillman L, Imm P, Wandersman A. The getting to outcomes demonstration and evaluation: An illustration of the prevention support system. *American Journal of Community Psychology*. 2008; 41(3-4):206–224. doi:10.1007/s10464-008-9163-2. [PubMed: 18278551]
- De Cooman G, Zaffalon M. Updating beliefs with incomplete observations. *Artificial Intelligence*. 2004; 159(1-2):75–125.
- Durlak J, DuPre E. Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*. 2008; 41(3):327–350. doi:10.1007/s10464-008-9165-0. [PubMed: 18322790]

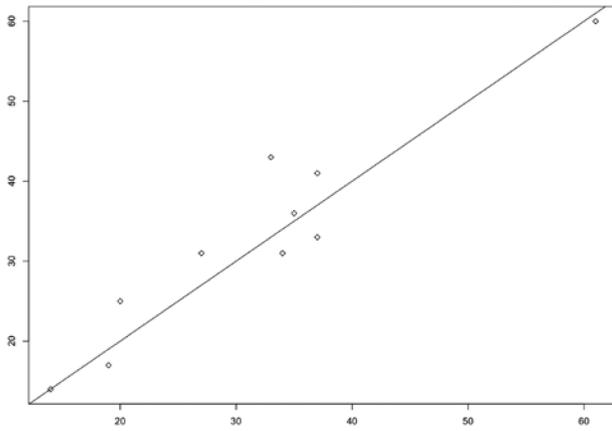
- Feil GE, Sprengelmeyer GP, Davis B, Chamberlain P. Development and testing of a multimedia internet-based system for fidelity and monitoring of multidimensional treatment foster care. *Journal of Medical Internet Research*. 2012; 14(5):e139. [PubMed: 23073495]
- Fixsen, DL.; Naoom, SF.; Blase, KA.; Friedman, RM.; Wallace, F. Implementation research: A synthesis of the literature. University of South Florida, Louis de la Parte Florida Mental Health Institute; Tampa, FL: 2005. (No. FMHI Publication #231).
- Flores G. Language barriers to health care in the United States. *New England Journal of Medicine*. 2006; 355(3):229–231. doi:10.1056/NEJMp058316. [PubMed: 16855260]
- Gallo, CG.; Aist, G.; Allen, JF.; de Beaumont, W.; Coria, S.; Gegg-Harrison, W.; Swift, M. Continuous understanding in a multimodal dialogue corpus.. *Proceedings of DECALOG: The 2007 Workshop on the Semantics and Pragmatics of Dialogue*; Trento, Italy. 2007. p. 75-82.
- Gallo, CG.; Jaeger, TF.; Furth, K. A database for the exploration of Spanish planning.. *The Seventh International Conference on Language Resources and Evaluation (LREC)*.; Valletta, Malta. 2010.
- Gelman, A.; Hill, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press; Cambridge, MA: 2006.
- Hanson R, Gros K, Davidson T, Barr S, Cohen J, Deblinger E, Ruggiero K. 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*, Online First. 2013:1–13. doi:10.1007/s10488-013-0492-6.
- Henggeler, SW.; Borduin, CM. *Family therapy and beyond: A multisystemic approach to treating the behavior problems of children and adolescents*. Brooks/Cole; Pacific Grove, CA.: 1990.
- Henggeler, SW.; Schoenwald, SK.; Borduin, CM.; Rowland, MD.; Cunningham, PB. *Multisystemic treatment of antisocial behavior in children and adolescents. Treatment manuals for practitioners*. Guilford Press; New York, NY: 1998.
- Henggeler SW, Schoenwald SK, Liao JG, Letourneau EJ, Edwards DL. Transporting efficacious treatments to field settings: The link between supervisory practices and therapist fidelity in MST programs. *Journal of Clinical Child and Adolescent Psychology: The Official Journal for the Society of Clinical Child and Adolescent Psychology*, American Psychological Association. Division 53. 2002; 31(2):155–67.
- Herschell AD. Fidelity in the field: Developing infrastructure and fine-tuning measurement. *Clinical Psychology: Science and Practice*. 2010; 17(3):253–257. doi:10.1111/j.1468-2850.2010.01216.x. [PubMed: 21116442]
- Hogue A, Dauber S, Chinchilla P, Fried A, Henderson C, Inclan J, Liddle HA. Assessing fidelity in individual and family therapy for adolescent substance abuse. *Journal of Substance Abuse Treatment*. 2008; 35(2):137–147. [PubMed: 17997268]
- Holmes, JN.; Holmes, WJ. *Speech synthesis and recognition 2e (HBK)*. 2nd ed.. Taylor & Francis; New York, NY: 2002.
- Huth, M.; Ryan, M. *Logic in computer science: Modelling and reasoning about systems*. Cambridge University Press; Cambridge, MA: 2004.
- Inoue M, Ogihara M, Hanada R, Furuyama N. Utility of gestural cues in indexing semantic miscommunication. *Future Information Technology (FutureTech)*, 2010 5th International Conference on Future Information Technology. 2010:1–6.
- Kellam SG, Mackenzie ACL, Brown CH, Poduska JM, Petras H, Wilcox HC. The good behavior game and the future of prevention and treatment. *Addiction Science and Clinical Practice*. 2011; 6(1):73–84. [PubMed: 22003425]
- Landsverk, J.; Brown, CH.; Chamberlain, P.; Palinkas, L.; Rolls Reutz, J.; Horwitz, SM. Design and analysis in dissemination and implementation research.. In: Brownson, RC.; Colditz, GA.; Proctor, EK., editors. *Dissemination and implementation research in health: Translating science to practice*. Oxford University Press; London: 2012. p. 225-260.
- Liddle, H. *Multidimensional family therapy (MDFT) for adolescent cannabis users*. Center for Substance Abuse Treatment, Substance Abuse and Mental Health Services Administration; Rockville, MD: 2002.

- Liddle HA, Dakof GA, Turner RM, Henderson CE, Greenbaum PE. Treating adolescent drug abuse: A randomized trial comparing multidimensional family therapy and cognitive behavior therapy. *Addiction*. 2008; 103(10):1660–70. [PubMed: 18705691]
- Liddle HA, Rowe CL, Gonzalez A, Henderson CE, Dakof GA, Greenbaum PE. Changing provider practices, program environment, and improving outcomes by transporting multidimensional family therapy to an adolescent drug treatment setting. *American Journal on Addictions*. 2006; 15(Suppl): 102–12. [PubMed: 17182425]
- Lopez A. Statistical machine translation. *ACM Computing Surveys (CSUR)*. 2008; 40(3):8.
- Minuchin, S. *Families and family therapy*. Harvard University Press; Cambridge, MA: 1974.
- Minuchin, S.; Fishman, CH. *Family therapy techniques*. Harvard University Press; Cambridge, MA: 1981.
- National Research Council and Institute of Medicine. In Committee on the Prevention of Mental Disorders and Substance Abuse Among Children, Youth, and Young Adults: Research Advances and Promising Interventions. In: O'Connell, ME.; Boat, T.; Warner, KE.; Board on Children, Youth, and Families, Division of Behavioral and Social Sciences and Education. , editors. *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities*. The National Academies Press; Washington, DC: 2009.
- Orlinsky, DE.; Ronnestad, MH.; Willitski, U. Fifty years of psychotherapy process-outcome research: Continuity and change [Handbook of psychotherapy and behaviour change]. 5th ed.. Lambert, MJ., editor. John Wiley & Sons; New York: 2004.
- Overholser JC. Elements of the Socratic Method: IV. Disavowal of knowledge. *Psychotherapy: Theory, Research, Practice, Training*. 1995; 32(2):283–292.
- Palinkas LA, Aarons GA, Horwitz SM, Chamberlain P, Hulburt M, Landsverk J. Mixed method designs in implementation research. *Administration & Policy in Mental Health and Mental Health Services*. 2011; 38(1):44–53.
- Pantin H, Coatsworth JD, Feaster DJ, Newman FL, Briones E, Prado G, Szapocznik J. Familias unidas: The efficacy of an intervention to promote parental investment in Hispanic immigrant families. *Prevention Science*. 2003; 4(3):189–201. [PubMed: 12940469]
- Pantin H, Schwartz SJ, Sullivan S, Prado G, Szapocznik J. Ecodevelopmental HIV prevention programs for Hispanic adolescents. *American Journal of Orthopsychiatry*. 2004; 74(4):545–58. [PubMed: 15554814]
- Pantin H, Prado G, Lopez B, Huang S, Tapia MI, Schwartz SJ, Branchini J. A randomized controlled trial of familias unidas for Hispanic adolescents with behavior problems. *Psychosomatic Medicine*. 2009; 71(9):987–995. doi:10.1097/PSY.0b013e3181bb2913. [PubMed: 19834053]
- Poduska J, Kellam SG, Brown CH, Ford C, Windham A, Keegan N, Wang W. Study protocol for a group randomized controlled trial of a classroom-based intervention aimed at preventing early risk factors for drug abuse: Integrating effectiveness and implementation research. *Implementation Science*. 2009; 4:56. doi:10.1186/1748-5908-4-56. [PubMed: 19725979]
- Popescu, A.; Etzioni, O.; Kautz, H. Towards a theory of natural language interfaces to databases.. *Proceedings of the 8th International Conference on Intelligent User Interfaces*; Miami, FL. 2003. p. 149-157.
- Prado G, Pantin H, Huang S, Cordova D, Tapia MI, Velazquez MR, Estrada Y. Effects of a family intervention in reducing HIV risk behaviors among high-risk Hispanic adolescents: A randomized controlled trial. *Archives of Pediatric and Adolescent Medicine*. 2012; 166(2):127–133. doi: 10.1001/archpediatrics.2011.189.
- Prado G, Huang S, Cordova D, Malcolm S, Estrada Y, Cano N, Brown CH. Ecodevelopmental and intrapersonal moderators of a family based preventive intervention for hispanic youth: A latent profile analysis. *Prevention Science: The Official Journal of the Society for Prevention Research*. 2013 doi:10.1007/s11121-012-0326-x.
- Prado G, Pantin H. Reducing substance use and HIV health disparities among Hispanic youth in the U.S.A.: The Familias Unidas program of research. *Intervencion Psicosocial*. 2011; 20(1):63–73. doi:10.5093/in2011v20n1a6. [PubMed: 21743790]
- Prado G, Pantin H, Briones E, Schwartz SJ, Feaster D, Huang S, Szapocznik J. A randomized controlled trial of a parent-centered intervention in preventing substance use and HIV risk

- behaviors in Hispanic adolescents. *Journal of Consulting and Clinical Psychology*. 2007; 75(6): 914–926. doi:10.1037/0022-006X.75.6.914. [PubMed: 18085908]
- Prado G, Pantin H, Schwartz SJ, Lupei NS, Szapocznik J. Predictors of engagement and retention into a parent-centered, ecodevelopmental HIV preventive intervention for Hispanic adolescents and their families. *Journal of Pediatric Psychology*. 2006; 31(9):874–890. doi:<http://dx.doi.org/10.1093/jpepsy/jsj046>. [PubMed: 16049264]
- Prinz RJ, Sanders MR, Shapiro CJ, Whitaker DJ, Lutzker JR. Population-based prevention of child maltreatment: The U.S. triple p system population trial. *Prevention Science*. 2009; 10(1):1–12. doi: 10.1007/s11121-009-0123-3. [PubMed: 19160053]
- Real, K.; Poole, MS. Innovation implementation: Conceptualization and measurement in organizational research.. In: Woodman, R.; Pasmore, W.; Shani, AB., editors. *Research in organizational change and development*. 15th ed.. Online: Emerald Group Publishing Limited; 2005. p. 63-134. doi:10.1016/S0897-3016(04)15003-9
- Rice E, Holloway IW, Barman-Adhikari A, Fuentes D, Brown CH, Palinkas LA. A mixed methods approach to network data collection. *Field Methods*. (In press).
- Robbins MS, Feaster DJ, Horigian VE, Puccinelli MJ, Henderson C, Szapocznik J. Therapist adherence in brief strategic family therapy for adolescent drug abusers. *Journal of Consulting and Clinical Psychology*. 2011; 79(1):43–53. doi:10.1037/a0022146. [PubMed: 21261433]
- Rowe C, Rigter H, Henderson C, Gantner A, Mos K, Nielsen P, Phan O. Implementation fidelity of multidimensional family therapy in an international trial. *Journal of Substance Abuse Treatment*. 2013; 44(4):391–399. [PubMed: 23085040]
- Sanders MR, Turner KM, Markie-Dadds C. The development and dissemination of the triple P-positive parenting program: A multilevel, evidence-based system of parenting and family support. *Prevention Science*. 2002; 3(3):173–89. [PubMed: 12387553]
- Schoenwald SK, Garland AF. A review of treatment adherence measurement methods. *Psychological Assessment*. 2013; 25:146–156. doi:10.1037/a0029715. [PubMed: 22888981]
- Schoenwald SK, Chapman JE, Kelleher K, Hoagwood KE, Landsverk J, Stevens J, Palinkas L. A survey of the infrastructure for children's mental health services: Implications for the implementation of empirically supported treatments (ESTs). *Administration and Policy in Mental Health and Mental Health Services Research*. 2008; 35(1-2):84–97. [PubMed: 18000750]
- Schoenwald SK, Garland AF, Chapman JE, Frazier SL, Sheidow AJ, Southam-Gerow MA. Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health*. 2011; 38(1):32–43. doi:10.1007/s10488-010-0321-0; 10.1007/s10488-010-0321-0. [PubMed: 20957425]
- Suarez-Morales L, Matthews J, Martino S, Ball SA, Rosa C, Farentinos C, Carroll KM. Issues in designing and implementing a Spanish-language multi-site clinical trial. *American Journal on Addictions*. 2007; 16(3):206–215. [PubMed: 17612825]
- Szapocznik J, Hervis OE, Schwartz S. Brief strategic family therapy for adolescent drug abuse. 2003 National Institute of Drug Abuse Rockville, MD (NIH publication no. 03-4751; NIDA Therapy Manuals for Drug Addiction Series ed.).
- Szapocznik, J.; Coatsworth, JD. *Drug abuse: Origins & interventions*. American Psychological Association; Washington, DC, US: 1999. An ecodevelopmental framework for organizing the influences on drug abuse: A developmental model of risk and protection.; p. 331-366. doi: 10.1037/10341-014
- Wall, L.; Christiansen, T.; Orwant, J. *Programming perl*. 3rd ed.. O'Reilly Media; 2000.

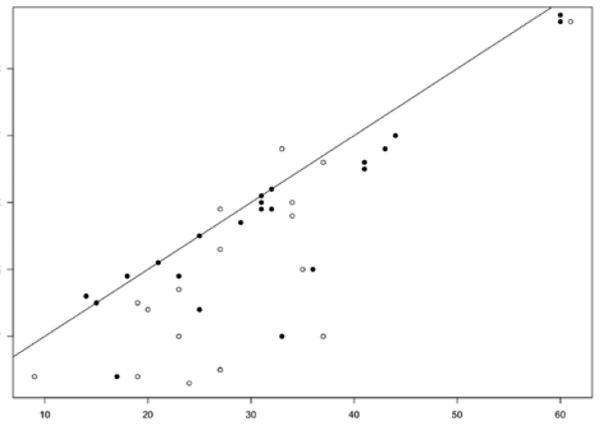


**Figure 1.** Decision tree algorithm for the rating utterances according to joining sub-dimension of open-ended questions  
\*utt = utterance



A) Human rater 1 (x-axis) versus rater 2 (y-axis)

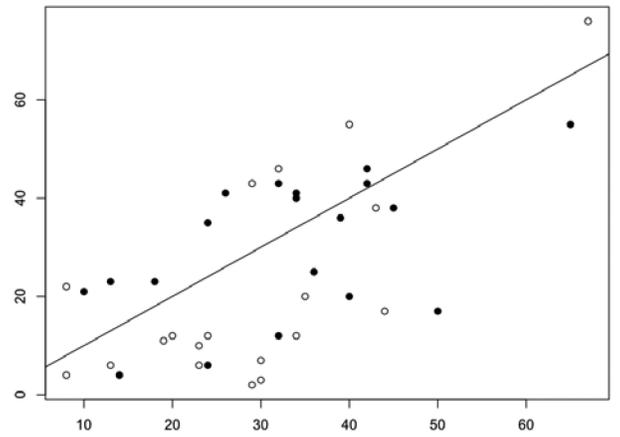
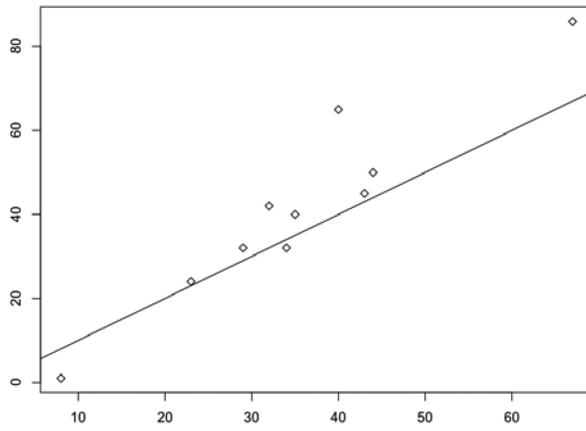
◇



B) Human rater 1 (x-axis) versus Machine rater (y-axis)

○  
●

**Figure 2.**  
Number of questions scored by human and machine raters.



A) Human rater 1 (x-axis) versus rater 2 (y-axis)

B) Human rater 1 (x-axis) versus Machine rater (y-axis)



Human rater 2 (x-axis) versus Machine rater (y-axis)

**Figure 3.**  
Total quality score for human and machine raters

**Table 1**

Pearson correlations based on the total number of utterances in each category at the session level. The categories are binary, improvable, and good questions.

Rater	Binary Questions	Improvable Questions	Good Questions	Sum All Questions (Binary plus Improvable plus Good)	Sum of Relevant Questions (Improvable plus Good)
Rater 1 & 2 (n=10)	.70	.36	.74	.95	.98
Rater 1 & Machine (n=19)	.45	.68	.45	.77	.79
Rater 2 & Machine (n=23)	.68	.27	.75	.91	.82

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 2**

Correlations between quality scores. Total quality score is defined as the sum of utterances that scores improvable 1 and good 2. The average quality score is the total quality score divided by the number of utterances that are questions (33 sessions).

<b>Rater</b>	<b>Total Quality Score</b>	<b>Average Quality Score</b>
Rater 1 & 2 (n=10)	.95	.76
Rater 1 & Machine (n=19)	.69	.35
Rater 2 & Machine (n=23)	.81	.32

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript