American Psychological Association
Society for the Advancement of Psychotherapy

Psychotherapy

© 2019 American Psychological Association
0033-3204/19/$12.00

2019, Vol. 56, No. 2, 318–328
http://dx.doi.org/10.1037/pst0000221

Design Feasibility of an Automated, Machine-Learning Based Feedback System for Motivational Interviewing

Zac E. Imel, Brian T. Pace, Christina S. Soma, and Michael Tanana
University of Utah

Tad Hirsch
Northeastern University

James Gibson
University of Southern California

Panayiotis Georgiou and Shrikanth Narayanan
University of Southern California

David C. Atkins
University of Washington

Direct observation of psychotherapy and providing performance-based feedback is the gold-standard approach for training psychotherapists. At present, this requires experts and training human coding teams, which is slow, expensive, and labor intensive. Machine learning and speech signal processing technologies provide a way to scale up feedback in psychotherapy. We evaluated an initial proof of concept automated feedback system that generates motivational interviewing quality metrics and provides easy access to other session data (e.g., transcripts). The system automatically provides a report of session-level metrics (e.g., therapist empathy) and therapist behavior codes at the talk-turn level (e.g., reflections). We assessed usability, therapist satisfaction, perceived accuracy, and intentions to adopt. A sample of 21 novice ($n = 10$) or experienced ($n = 11$) therapists each completed a 10-min session with a standardized patient. The system received the audio from the session as input and then automatically generated feedback that therapists accessed via a web portal. All participants found the system easy to use and were satisfied with their feedback, 83% found the feedback consistent with their own perceptions of their clinical performance, and 90% reported they were likely to use the feedback in their practice. We discuss the implications of applying new technologies to evaluation of psychotherapy.

Clinical Impact Statement
Question: How do therapists experience automated evaluations of their sessions? ***Findings:*** Therapists endorsed strong satisfaction, usability, and perceived accuracy of the automated feedback. ***Meaning:*** Machine-learning technologies have the potential to dramatically scale up the amount of feedback therapists receive after their sessions. ***Next Steps:*** Building on this pilot study, both usability and accuracy should be tested in larger and different types of therapist samples. Additional work should focus on the potential impact of automated feedback on therapist behavior in session.

Keywords: motivational interviewing, adherence, dissemination, machine learning, feedback

This article was published Online First April 8, 2019.

Zac E. Imel, Brian T. Pace, Christina S. Soma, and Michael Tanana, Department of Educational Psychology, University of Utah; Tad Hirsch, Department of Art and Design, Northeastern University; James Gibson, Department of Electrical Engineering, University of Southern California; Panayiotis Georgiou and Shrikanth Narayanan, Department of Electrical Engineering, University of Southern California; David C. Atkins, Department of Psychiatry and Public Health, University of Washington.

Funding for the preparation of this article was provided by the National Institutes of Health / National Institute on Alcohol Abuse and Alcoholism (NIAAA) under award R01/AA018673 and National Institute on Drug Abuse under award (R34/DA034860). David C. Atkins effort was also supported by K02/AA023814. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Some findings presented in this article were presented at the Annual Convention of the Association of Behavioral and Cognitive Therapies in New York City, and the Annual Convention of the American Psychological Association in Denver Colorado. Zac E. Imel, David C. Atkins, Michael Tanana, Tad Hirsch, and Shrikanth Narayanan are cofounders and have equity stakes in a technology company—lyssn.io—focused on developing computational models that quantify aspects of patient–provider interactions. Shrikanth Narayanan is also Chief Scientist and co-founder with equity stake in Behavioral Signals, a technology company focused on creating technologies for emotional and behavioral machine intelligence.

Correspondence concerning this article should be addressed to Zac E. Imel, Department of Educational Psychology, University of Utah, 1705 Campus Center Drive, Salt Lake City, UT 84112-9225. E-mail: zac.imel@utah.edu

> The acquisition of skills requires a regular environment, an adequate opportunity to practice, and rapid and unequivocal feedback about the correctness of thoughts and actions. When these conditions are fulfilled, skill eventually develops, and the intuitive judgments and choices that quickly come to mind will mostly be accurate.
>
> —Daniel Kahneman

Psychotherapy research focuses on the development and evaluation of interventions that address mental health problems. There are now thousands of clinical trials demonstrating the efficacy of different specific interventions for a range of problems (see, e.g., American Psychological Association Division 12 website on Empirically Supported Treatments; https://www.div12.org/treatments). Recent research and quality improvement efforts have focused on how to best disseminate and implement these treatments in community settings. In the past decade, health systems have spent at least $2 billion on various efforts to train providers in the provision of specific evidence-based treatments (McHugh & Barlow, 2010). However, these and other efforts to provide psychotherapy training, supervision, and quality assurance are hampered by the impracticality—if not near impossibility—of offering providers rapid, objective, performance-based feedback. According to McHugh and Barlow (2010), trainings would ideally assess ". . . objective assessment of fidelity including clinician competence and number and percentage of clinicians who complete training, achieve competence, and sustain competence . . ." (p. 83). Given current technology, this is not a realistic goal.

The current gold standard for monitoring provider fidelity to treatment relies on human raters for assessment, which is slow and unsustainable in the vast majority of clinical settings. Human evaluation of psychotherapy sessions is simply a nonstarter in community settings, and thus, virtually all of the more than 80 million psychotherapy sessions in a given year (Olfson & Marcus, 2010) are not evaluated. However, recent developments in machine learning and speech signal processing now offer a route to rapid, performance-based feedback for counseling and psychotherapy (Imel, Steyvers, & Atkins, 2015). The current research describes a pilot feasibility study of a proof-of-concept system that provides rapid, performance-based feedback developed via user-centered design.

The Effects of Skill-Based Feedback in Psychotherapy Training

There is a well-developed literature on the necessary conditions for skill development across domains (Kulik & Kulik, 1988). One crucial component is regular feedback on whether the skill has been performed correctly. Feedback, in general, has been shown to increase skill performance (Kluger & DeNisi, 1996), though certain types of feedback have been found to be more important than others. For example, feedback on whether or not a task was performed correctly has been shown to be an extremely effective means for teaching skills (Hattie & Timperley, 2007). In contrast, feedback about the characteristics of the person being evaluated has negative effects on learning skills (Hattie & Timperley, 2007). More immediate feedback has a stronger impact on skill development compared with delayed feedback (Epstein, Epstein, & Brosvic, 2001; Kulik & Kulik, 1988).

In psychotherapy, there have been successful examples of feedback interventions, though most have significant drawbacks. For example, informing clinicians whether their clients are improving can bolster outcomes for at-risk clients (Kendrick et al., 2016). Feedback on client's symptoms has several advantages; most notably it is an efficient way of providing feedback and does not require significant time from trained staff. However, client outcome feedback also has limitations. For example, the changes in therapist performance are not durable after the feedback is removed (Lambert, Harmon, Slade, Whipple, & Hawkins, 2005). Second, this type of feedback does not target any specific therapist behaviors (e.g., what did the therapist do to lead to the problem or what can they do to solve it), and there is no yet evidence yet that symptom-based feedback has a measurable impact on therapist behavior.

Motivational interviewing (MI) researchers have been at the forefront of studying how to train therapists to use specific skills (e.g., increase in empathy and use of open questions and reflections; Baer et al., 2009). Ordinarily, trainings for community providers involve an in-person workshop. The training itself typically includes lectures on theory and research of MI, demonstrations, and practice with role-played clients, where providers receive some feedback on their use of MI. Less commonly, trainee therapists submit tapes demonstrating their utilization of MI skills (Miller, Yahne, Moyers, Martinez, & Pirritano, 2004). In a recent meta-analysis of training studies, workshop-based training increased provider utilization of MI skills, but without ongoing performance-based feedback or coaching, the training effect deteriorated over time (Schwalbe, Oh, & Zweben, 2014). Unfortunately, ongoing performance-based feedback of provider behavior is rare in community settings (see Creed et al., 2016 for an exception). A major barrier to high-quality implementation of behavioral interventions is the need for observer-rated fidelity (Proctor et al., 2009). As noted previously, the time-consuming nature of behavioral coding of provider fidelity prohibits its use in community settings, with a few rare exceptions.

Without ongoing feedback, providers are not likely to maintain newly learned skills that are present when they are initially trained. The necessary posttraining support to maintain new skills is clear (i.e., performance-based feedback), but the standard process of generating this feedback (i.e., using humans as the assessment tool via behavioral coding) is too slow and expensive to support wide scale adoption. Fortunately, there is initial research that suggests modern methods from computer science have the potential to speed up the feedback process.

Technology-Based Evaluation of Psychotherapy

Around the time that Carl Rogers first recorded psychotherapy sessions in the 1940s (Rogers, 1951), natural language processing (NLP) began to emerge as a subfield of computer science (Jones, 1994). Currently, NLP is a subfield of machine learning with the primary aim of training computers to learn, understand, and analyze language (Hirschberg & Manning, 2015). A psychotherapy session is essentially a conversation between two individuals, rich with semantic data that often goes unanalyzed due to the laborious nature of human coding. NLP techniques allow researchers to take this conversational data, often in the form of large collections of unstructured text, and help answer important process questions, such as "What did the client and therapist talk about during this session?" or "How empathic was this therapist?". Recent examples of machine learning/NLP methods applied to psychotherapy include (a) identifying therapist reflections (Can et al., 2016), (b) deriving the primary content of psychotherapy sessions and at what point in the session (Gaut, Steyvers, Imel, Atkins, & Smyth, 2017; Howes, Purver, & McCabe, 2013, 2014; Imel et al., 2015), (c) classifying different treatment approaches such as cognitive-behavioral therapy or psychodynamic therapy (Imel et al., 2015), and (d) evaluating how empathic a therapist was solely based on audio recordings of a session (Hasan et al., 2016; Pace et al., 2016; Pérez-Rosas et al., 2017; Xiao, Imel, Georgiou, Atkins, & Narayanan, 2015). Finally, using a sample of more than 300 MI sessions, NLP models were able to estimate a full range of MI fidelity metrics based on therapist statements (e.g., open questions and reflections) with accuracy that approaches human performance (Tanana, Hallgren, Imel, Atkins, & Srikumar, 2016).

Current Investigation

There are now several studies demonstrating that machine-learning algorithms can use session audio or transcripts to generate ratings of psychotherapy sessions that are consistent with traditional human-derived observer ratings. However, no study has attempted to use these methods to provide feedback to providers, and no research has explored how such feedback should be delivered to clinicians when it is generated rapidly by machine-learning models. MI training research incorporating provider feedback has typically presented therapist fidelity scores in a paper or electronic document format or in the context of a supervisory or consultative interaction with a trainer (Miller et al., 2004). The use of computer-generated feedback introduces both complications and opportunities: Immediate, objective feedback offered in a visually appealing manner that offers interactivity could enhance provider engagement and learning; however providers may be skeptical of their work being evaluated by a computer.

In this research, we present an initial evaluation of a web-based interactive tool that automatically provides feedback to providers on specific psychotherapy skills.[1] To do so, we developed a tool that rapidly generated machine-learning-based feedback on MI to a sample of 21 therapists who recorded 10-min sessions with standardized patients describing problems with substance use. The primary aim of the study was to assess the usability of the tool itself, and thus the sample was not powered to conduct a reliable test of consistency with human codes (see Tanana et al., 2016; Xiao et al., 2015 for prior large-scale evaluations). However, we provide an exploration of the correspondence of machine generated MI ratings with human ratings. We expected providers would be generally satisfied with the feedback and its presentation, find the feedback easy to use and interpret, perceive the quantitative feedback as reasonably accurate of their actual performance, and be inclined to adopt the technology if it were available.

Method

Participants

Participants were 11 experienced (i.e., licensed) and 10 novice (i.e., trainee) therapists (total $N = 21$), recruited to record a brief substance abuse counseling session with a standardized patient (SP). Note that we systematically recruited providers with different levels of experience to ensure meaningful variability in therapy performance and supervision, but we did not hypothesize or test specific differences between these groups.

Experienced therapists were local practicing and licensed professionals, recruited via a snowballing sampling procedure. The authors utilized their professional connections in the local community to recruit experienced therapists from local university-based and community mental health sites. These experienced therapists were also asked to identify other empathic and licensed community therapists, who were then contacted via e-mail by a research assistant. We specifically recruited experienced providers who had prior knowledge or training with MI, but this was not required because we were also in interested in feedback from therapist with no prior training in MI. The majority of experienced therapists had received their doctorate and had been licensed more than 2 years (64%, $n = 7$). Experienced therapists reported a variety of MI training from frequent use of MI in their practice ($n = 2$, 18%), membership in an MI organization (e.g., Motivational Interviewing Network of Trainers; $n = 1$, 9%), familiarity with and formal training in MI ($n = 7$, 64%), and familiarity but no formal training ($n = 1$, 9%).

Beginning therapists were recruited from a local university's master's- and doctoral-level clinical training program. Clinicians were 86% White/Caucasian ($n = 18$), 10% Hispanic/Latin ($n = 2$), and 5% Asian American/Pacific Islander ($n = 1$), with a mean age of 42.1 ($SD = 11.7$). Trainees were all enrolled in introductory training courses (e.g., counseling micro skills and counseling theories) and were within the first year of their training program. Of the novice therapists ($n = 10$), 80% were master's students ($n = 8$) and reported having either no training or experience with MI ($n = 4$, 40%), or some familiarity with MI but no training ($n = 5$, 60%; 1 missing response).

Procedure. After recruitment, participants completed study consent and demographic information via an online survey platform. A research assistant then contacted them to schedule and complete a 10-min substance abuse session with an SP. The session was recorded with two lavalier microphones that were clipped to both the therapist and SP, allowing high-quality audio data capture. Following the session, participants received computerized feedback via web portal (described in the following text). Participants were then asked to complete a web-based survey regarding their satisfaction with the tool.

Standardized patients. Doctoral students ($n = 4$; two female and two male) functioned as SPs. Three SP profiles were created, where all focused on presenting concerns related to substance abuse. SPs are commonly utilized in the psychotherapy and medical training literature when the target of investigation is provider behavior. SPs also avoid issues with missing data and audio quality that occur when samples are requested from community-based therapists (Baer et al., 2004). One profile described an individual struggling with methamphetamine usage who was required to attend therapy. Two profiles described college age students who were experiencing negative consequences related to drinking. SPs

[1] By automatic, we mean that no human evaluation was necessary to generate the specific feedback scores or the report presenting them. However, the technology tested does not provide feedback in "real time" during the session. It is available shortly after the session is completed, the amount of time dependent on the processing speed of the computers used to run the models.

were trained to respond in ways consistent with their profiles but maintained flexibility to individual therapist responses (i.e., there was no set script).

Feedback system. In previous publications, we provided descriptions of the visual design (Gibson et al., 2016), as well as development and validation of the speech signal and machine-learning components of the feedback system (Atkins, Steyvers, Imel, & Smyth, 2014; Tanana et al., 2016; Xiao et al., 2015). Briefly, the system first separates audio segments of speech from nonspeech using a process called voice activity detection; then, all the speech segments are separated into two groups, one for each speaker, in a process called speaker diarization. Person-specific speech segments are transcribed using an automatic speech recognition (ASR) pipeline developed with the *kaldi* software library (Povey et al., 2011). Using the automatically transcribed words, we then used a role-matching model to identify which speech segments belong to the counselor and which belong to the client (details on each of these steps can be found in the study Xiao et al., 2016). The system next used the lexical content from the ASR session transcripts to predict specific MI fidelity codes for each session (see the Measures section in the following for details on MI fidelity codes).[2] The ASR transcript results were used as inputs to a support vector regression model (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997) based on information fused from a maxent language model (Berger, Pietra, & Pietra, 1996), a maximum likelihood language model (Jurafsky & Martin, 2008), and ASR lattice rescoring. In total, the combination of the speech processing and machine-learning pipeline yields predicted scores for each MI fidelity code, which are either single values for the entire session (i.e., global scores) or utterance-specific (i.e., behavioral codes). Descriptions for each of the MI fidelity codes that was included in the feedback report are listed in Table 1. The prediction model was trained with a fully transcribed and behaviorally coded set of 345 MI substance abuse treatment sessions (Lord et al., 2015; Xiao et al., 2015). In addition, ASR language models were also trained by transcripts from a larger, general psychotherapy corpus (Imel et al., 2015).

The therapist directed web-based tool (see example report in Figure 1 and interactive examples online here: http://sri.utah.edu/psychtest/misc/demoinfo.html)[3] provided visualizations of MI fidelity scores, and session content was developed through an iterative, user-centered design process (Gibson et al., 2016; Norman, 2013). Following their SP session, therapists could access their automated feedback report via a password-protected web portal. The report included visual summaries of MI fidelity and also included a session timeline visualization, indicating when either the therapist or client was talking at each point throughout the session. In addition, the session timeline linked to the ASR-based transcript of the session as well as predicted MI fidelity codes for each utterance.

Measures

MI fidelity. All SP sessions were rated by machine-learning models that had been trained to evaluate fidelity to MI. In the feedback portal, every therapist utterance in the ASR transcript received a specific MI code. In addition, the system provided session-level ratings that were compared with standard benchmarks for MI spirit, empathy, reflection-to-question ratio, and percentage of open questions. Both MI spirit and empathy are Likert ratings scored from 1 to 5, and the other metrics are aggregated from the utterance-level labels.

The machine-learning models of MI fidelity were trained by two of the most common observer-rated measures of MI fidelity, the Motivational Interviewing Skills Code (MISC) and Motivational Interviewing Treatment Integrity (MITI) Scale. The MISC v2.1 (Miller, Moyers, Ernst, & Amrhein, 2008) is an utterance-level system that we used to train machine-learning models for behavior codes (e.g., simple and complex reflections, open and closed questions, and giving information). Human reliability (intraclass correlations [ICCs]) for the data used to train these models ranged from .92 (giving information) to .61 (MI-nonadherent), with six of the seven utterance-level codes above .7, $M = .80$. Human reliability reflects the amount of measurement error in the data used to train the machine-learning models and as such provides an upper bound for the correlations of machine-generated predictions. As with human raters, the models classify each utterance, and therapist behavior codes are tallied and used to calculate specific fidelity metrics presented to the counselor in the interactive report.

To train machine-learning models of session level ratings of empathy and MI spirit, we used the MITI v3.1 (Moyers, Martin, Manuel, Miller, & Ernst, 2010) a less intensive version of the MISC (Xiao et al., 2015). The MI Spirit composite score was calculated by aggregating ratings of evocation, collaboration, and autonomy/support. Human agreement on the data used to train these models was adequate (MI spirit: ICC = .68; empathy: ICC = .75). The accuracy of both utterance-level and session-level models in a sample of over 300 sessions has been previously reported (Tanana et al., 2016; Xiao et al., 2016).

Finally, a single human rater coded all sessions using the MITI 3.1. The coder is a member of the Motivational Interviewing Network of Trainers and was the trainer of the two previous coding teams that generated the data that trained the machine-learning models described earlier. Each session was coded in its entirety in a single pass, as recommended by the MITI manual. Agreement (ICCs) between the human rating and the machine generated codes varied from .23 (empathy) to .80 (closed questions), $M = .48$, which was on average 62% of human agreement ($SD = 23$).

Therapist evaluation of system. After receiving feedback on their session, we provided therapists a survey that assessed usability, satisfaction, perceived accuracy, and intentions to adopt the technology. The survey was designed for this study, as many of the items were designed to assess features unique to the specific feedback tool. The present study focuses on two sets of clinically relevant items. These include four items that evaluated (a) ease of use of the interface, (b) how representative the feedback was of their performance, (c) overall satisfaction, and (d) whether they would use the tool in their clinical practice. In addition, we asked participants five total questions on how easy it was to understand their feedback on a several key metrics including (a) empathy, (b) MI spirit, (c) reflection-to-question ratio, (d) percentage of open

[2] It is likely that acoustic features of human speech are also predictive of MI fidelity codes. However, the models currently in the feedback system do not yet incorporate this information. Addition of acoustic features is a focus of ongoing research.

[3] The session available online is not a research study session, but a fully roleplayed session.

Table 1
Descriptions of Each Motivational Interviewing Metric Provided in the Feedback Tool[a]

MI Metric	Description
Overall MI fidelity	The overall MI fidelity score ranges from 0 to 12, where 12 represents excellent fidelity to MI. You receive 0, 1, or 2 points for your performance on each of the six MI fidelity metrics (MI spirit, empathy, reflection-to-question ratio, percent open questions, percent complex reflections, and percent MI adherence). You receive 1 point for scores that meet (basic) proficiency benchmarks, and 2 points for scores that meet benchmarks for (advanced) competence.
MI adherence	MI counselors who are "adherent" are those who ask open questions, make complex reflections, support and affirm their clients, and emphasize their client's autonomy. Nonadherent counselors are confrontational, directing, warning, and advice-giving. This measure is the total number of MI adherent behaviors divided by the sum of adherent and nonadherent behaviors. Higher is better, and anything less than 100% indicates some nonadherent behaviors.
MI spirit	MI spirit captures the general counseling style of MI, which (a) is a collaborative approach, (b) shows interest in and evokes the client's perspective, and (c) does not impose views on the client but rather supports the client's ability to make their own choices for their life.
Empathy	Empathy is a rating of how well the therapist understands the client's perspective and makes efforts to see the world as their client sees it.
Reflections-to-questions ratio	An effective MI counselor uses more reflective statements (i.e., summaries of what the client has said) versus questions. This measure is a ratio of the total number of reflections divided by the total number of questions, and, thus, higher is better.
Percent open question	When MI counselors ask questions, they strive to ask "open" questions that invite a range of possible answers and may invite the client's perspective or encourage self-exploration. Closed questions are ones that can be answered in a single or a few words. This measure is the total number of open questions divided by the sum of open and closed questions. Higher is better.
Percent complex reflections	Reflections are summaries of what the client has expressed and said. A simple reflection is an almost verbatim restatement of what the client said. A complex reflection is a summary that adds meaning or emphasis or might integrate additional information. A complex reflection is one way that a therapist conveys they are trying to understand their client and his or her worldview. This measure is the total number of complex reflections divided by the sum of complex and simple reflections. Higher is better.

Note. MI = motivational interviewing. Descriptions of MI fidelity metrics were adapted from the Motivational Interviewing Skills Code Version 2.1 (Miller, Moyers, Ernst, & Amrhein, 2008).
[a] Participants could access this information by clicking on the "i" button next to score.

questions, and (e) percentage of complex reflections. Across items, users responded to questions on a qualitative Likert scale with answers ranging from "*strongly disagree*" to "*strongly agree*" or "*very unhelpful*" to "*very helpful*." Users were asked to make qualitative comments at the end of the survey, and we include a representative selection of responses (both supportive and critical) to illustrate the quantitative results.

Results

Figure 2 reports participant feedback on clinical feasibility. All 21 participants rated the tool as easy to use (*strongly agree*, n = 13, 62%; *slightly agree*, n = 8, 38%). In all, 18 of 21 participants (86%) either strongly (n = 8; 38%) or slightly (n = 10; 48%) agreed that the feedback was representative of their clinical performance (one slightly disagreed, and two were unsure). All 21 participants indicated that they were satisfied with the computer-generated feedback report (*strongly agree*, n = 14, 67%; *slightly agree*, n = 7, 33%). In all, 19 of 21 participants (90%) participants agreed that if the tool was available, they would use it in their clinical practice (*strongly agree*, n = 15, 71%; *slightly agree*, n = 4, 19%; and *unsure*, n = 2, 10%).

We also assessed how easy participants found it to understand the scores they received in different MI domains (Figure 3). Across all scores, no fewer than 17 participants (81%) and up to all 21 participants found their scores either very or somewhat easy to understand. A total of 17 participants (81%) found the MI spirit rating very (n = 11) or somewhat (n = 7) easy to understand (three found it somewhat hard). A total of 19 participants (90%) found the empathy rating very (n = 16) or somewhat (n = 3) easy to understand (one found it somewhat hard, and one, very hard). A total of 19 participants (90%) found the reflection-to-question ratio metric very (n = 14) or somewhat (n = 5) easy to understand (one found it somewhat hard). All 21 participants (100%) found the percent open questions measure very (n = 17) or somewhat (n = 5) easy to understand. All 21 participants (100%) found the percent complex reflection rating very (n = 17) or somewhat (n = 5) easy to understand (Figure 3). A sample of participant free text comments are listed in Table 2. We included a selection of satisfied responses, as well as a sample of those more confused with their individual feedback.

Discussion

To our knowledge, this is the first evaluation of machine–learning-based technology to provide feedback to counselors. Previous research demonstrated the accuracy of the machine-learning models that generated the feedback (Tanana et al., 2016; Xiao et al., 2015), and the technical details of the processing pipeline (Xiao et al., 2016), but this is the first evaluation of an integrated system that provides machine generated feedback directly to therapists based on a session audio recording.

At present, specific performance-based feedback such as those provided by fidelity codes are rare during training and often completely absent for post licensure therapists. There is a substantial literature on the positive effects of feedback based on client