Association of Machine Learning Rated Supportive Counseling Skills with Psychotherapy Outcome

Xinyao Zhang[1], Simon B. Goldberg[2], Scott A. Baldwin[3], Michael Tanana[4], Lauren Weitzman[1], Shrikanth Narayanan[5], David C. Atkins[4, 6], Zac E. Imel[1, 4]

[1] Department of Educational Psychology, University of Utah

[2] Department of Counseling Psychology, University of Wisconsin – Madison

[3] Department of Psychology, Brigham Young University

[4] Lyssn.io

[5] Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California

[6] Department of Psychiatry and Behavioral Sciences, University of Washington

Author Note

MT, SN, DCA, and ZEI each have minority equity interests in Lyssn.io, a technology company focused on developing technologies to evaluate the contents of psychotherapy. SN is also the Chief Scientist and Co-founder with equity stake of Behavioral Signals, a technology company focused on creating technologies for emotional and behavioral machine intelligence. SBG was supported by the National Center for Complementary and Integrative Health (K23AT010879). This study was funded by the National Institute on Alcohol Abuse and Alcoholism (AA018673).

Correspondence concerning this article should be addressed to Xinyao Zhang, University of Utah, 1721 Campus Center Dr, Salt Lake City, UT 84112. Email: xinyao.zhang@utah.edu.

Abstract

Objective: This study applied a machine-learning-based skill assessment system to investigate the association between supportive counseling skills (empathy, open questions, and reflections) and treatment outcomes. We hypothesized that higher empathy and higher use of open questions and reflections would be associated with greater symptom reduction. **Method**: We used a dataset with 2974 sessions, 610 clients, and 48 therapists collected from a university counseling center, which included 845,953 rated therapist statements. Client outcome was routinely monitored by the Counseling Center Assessment of Psychological Symptoms Instruments. Therapists' skills were measured via computer by a bidirectional-long-short-term-memory-based system that rated use of supportive counseling skills. We used multilevel modeling to separate the between-therapist and the within-therapist associations of the skills and outcome. **Results**: Use of open questions and reflections was associated with client symptom reduction between therapists but not within therapists. We did not find significant associations between therapist empathy and client symptom reduction, but found that empathy was negatively associated with clients' baseline symptom level within therapists. **Conclusions**: Therapist exploration of clients' experience and expression of understanding may be important skills that are associated with clients' better outcomes. This study highlights the importance of support counseling skills, as well as the potential of machine-learning-based measures in psychotherapy research. We discuss the limitations of the study, including the limitations related to the speaker recognition system and potential reasons for the lack of association between empathy and client outcome.

Keywords: Supportive counseling skills, Open question, Reflection, Large-scale study, Multilevel model, Machine learning.

Association of Machine Learning Rated Supportive Counseling Skills with Psychotherapy Outcome

Supportive counseling is typically composed of therapist skills such as active listening, encouraging, reflecting, and helping clients explore their own experiences and feelings (Areán et al., 2010). Some of these skills are also prevalent in other specific treatments, such as reflection of feelings in psychodynamic therapy, emotion focused therapy, etc., and open questions in nearly all treatments that explore clients' personal experience in depth. There is a long history of research indicating that listening attentively and empathically is therapeutic (Conte, 1994; Winston et al., 1986). A meta-analysis of 31 randomized trials showed that nonspecific supportive counseling alone can achieve approximately 75% of the effect of active evidence-based treatments (Cuijpers et al., 2012). Despite the considerable effect of supportive counseling as a whole, results from studies on the associations of individual supportive counseling skills and client outcomes are mixed (Elliott et al., 2023; Hill, 1992; Kadur et al., 2020; Williams et al. 2023). However, most of the prior studies are limited in size and focused on the "total" relationship between these skills and outcomes, which may not capture more nuanced associations between and within different therapists. The present study assessed these more nuanced associations on a large scale. We used a previously trained automatic skill assessment system to assess therapist skills at scale based on verbal content. We then applied multilevel modeling to a collection of 2974 sessions to assess 1) whether a therapist's use of a supportive skill on average was associated with client outcomes (between-therapist effect), and 2) whether therapists varying their skill use based on their clients was associated with client outcomes (within-therapist effect).

Prior Process-Outcome Studies on Supportive Counseling Skills

Early studies on supportive counseling skills suffered from idiosyncratic coding systems, lack of standardization, and minimal validity testing (Hill, 1978). However, even with the standardization of coding systems such as verbal response modes (e.g., Elliott, 1985; Hill, 1978; Stiles, 1978), findings on supportive counseling skills and therapy outcomes remained mixed (Elliott et al., 1985; Hill et al., 1988, 1992; Kadur et al., 2020). For example, a recent review indicated mixed associations between open questions and outcomes, except that they may facilitate emotional processing (Williams, 2023). A meta-analysis (client $n = 2710$) found no association between reflection and outcomes (Elliott et al., 2023).

Looking only at "total" linear associations of skill use and outcomes is often not sufficient to capture the full picture of the relationship (Elliott et al., 2023; Hill, 1992). One limitation is that these analyses preclude examination of how therapists' differences in using these skills are associated with client outcomes. Clients' treatment responses depend on the therapists they work with (Baldwin & Imel, 2013; Crits-Christoph et al., 1991), and therapists' adherence to particular treatment often varies (Boswell et al., 2013; Imel et al., 2011). A therapist can also vary their skill use when working with different clients (Uhl et al., 2022). A recent study showed that, like other treatment-specific skills, these supportive skills also varied significantly between-therapist and within-therapist (Zhang et al., 2022). Most of the prior studies on supportive skills and outcomes did not take these therapist variations into consideration.

Knowing how supportive counseling skills relate to outcomes between and within therapists has important clinical implications. For example, while causality cannot be established from a correlational design, if therapists who use more reflections in general have better client outcomes (between-therapist effect), it could be useful to consider training strategies that increase therapists ability to perform these skills overall. Instead, if client outcomes covary with

a therapist's use of reflection relative to their other clients (within-therapist effect), it may imply the importance of client-therapist interaction, suggesting further studies on dyadic processes such as mutual influence (Imel et al., 2011) or therapist responsiveness (Constantino et al., 2020). Between- and within-therapist associations of process factors (including skills) and outcomes are not always consistent with each other. For example, one study showed that alliance and outcome were associated between therapists not within (Baldwin et al., 2007), suggesting that it is a therapist's overall alliance building skill that relates to client outcomes. On the other hand, studies on CBT skills indicated that the association with client outcomes occurred within therapists, not between therapists (Uhl et al., 2022). Therefore, it is important to separate the between-therapist and within-therapist associations of individual supportive skills and outcomes.

Another limitation of the prior studies is that the sample sizes tend to be small. For example, the average number of clients per study in a meta-analysis on reflections was 70, with earlier studies having even smaller samples (Elliott et al., 2023). This could make it difficult to detect small effects. In addition, a skill can be operationalized differently with different coding schemes. Reflection could be defined as rephrasing client's statements with reference to feelings (Hill, 1978). In other systems, this may be called clarification of feelings (Frank & Sweetland, 1962). Taken together, these issues could dampen the estimation of the effects of supportive counseling skills. A large-scale naturalistic study with adequate sample size and a homogeneous skill coding system may be helpful for the above limitations.

Machine Learning in Psychotherapy Research

Large-scale evaluation of therapist skills relying on observer rating tends to demand massive time and resources (Imel et al., 2017). Studies of this kind could take years to complete (see Creed et al., 2016; therapist $n = 321$). Machine learning has shown promise in automating

observer-rated psychotherapy evaluations addressing time and resource demand (Aafjes-van Doorn et al., 2020). For example, support vector machines (SVM) are a group of machine learning methods that find a classifying hyperplane or a regressing curve using a special group of data points called support vectors (Cortes & Vapnik, 1995). More recently, long short-term memory (LSTM) is a type of deep neural network method that is specialized to process sequential data such as natural language (Hochreiter & Schmidhuber, 1997). These methods have been used to build automatic assessors for key supportive counseling skills such as empathy (Gibson et al., 2016; Xiao et al., 2015), open questions (Flemotomos et al., 2021), and reflections (Can et al., 2016), as well as other treatment-specific skills (e.g., CBT skills; Ewbank et al., 2019). For example, Gibson et al. (2016) obtained a sensitivity of .79 in distinguishing high vs. low therapist empathy using LSTM. Xiao et al. (2015) also found that a simple method like SVM can achieve a correlation of .65 to .71 with human coders on empathy evaluation. More importantly, using these models has significantly reduced the time needed to code therapist skills in large-scale studies (e.g., session n = 90,934; Ewbank et al., 2019).

Present Study

In this study, we assessed the between-therapist and within-therapist associations of supportive counseling skills and treatment outcomes in a naturalistic setting based on 2974 session recordings. We used an automatic skill assessment system, previously trained with a standard therapist skill coding protocol (Houck et al., 2013). The system can automatically transcribe and code thousands of sessions based on verbal content (Flemotomos et al., 2021). We discuss the details of the system in the Method section.

We examined the between-therapist and the within-therapist effects of three supportive skills: empathy, open questions, and reflections. We made the following hypotheses: **Hypothesis**

1a: Higher between-therapist empathy would be associated with larger symptom reduction; **Hypothesis 1b**: Higher within-therapist empathy would be associated with larger symptom reduction. **Hypothesis 2a:** Higher between-therapist use of open questions would be associated with larger symptom reduction; **Hypothesis 2b**: Higher within-therapist use of open questions would be associated with larger symptom reduction. **Hypothesis 3a:** Higher between-therapist use of reflection would be associated with larger symptom reduction; **Hypothesis 3b**: Higher within-therapist use of reflection would be associated with larger symptom reduction.

Method

Participants

The dataset was collected from a counseling center at a large university in the western United States from 2017 to 2020 (Flemotomos et al., 2021). Sessions with recording or processing errors were excluded (Flemotomos et al., 2021). Clients with only one session (intake) were excluded.[1] The final dataset contained 610 clients, 48 therapists, and 2974 sessions. The average number of sessions per client was 4.9 (*Mdn* = 4), ranging from 2 to 17 sessions. The average number of clients per therapist was 12.7 (*Mdn* = 11.5), ranging from 1 to 43 clients. Tables S1 and S2 contain the demographic information of clients and therapists, respectively.

Measures

Three supportive counseling skills were selected for this study: empathy, open questions, and reflections. These skills were automatically assessed by a previously trained skill assessing system (Flemotomos et al., 2021). Clients' pre- and post-treatment symptom levels were

[1] In addition, we conducted a sensitivity analysis to test the robustness of associations when therapists with only one client were removed. Including therapists with only one client could introduce noise in the estimation of between-therapist effects. However, there were no substantial differences in the full dataset and the restricted (Table S4), suggesting that results were robust to therapist caseload size.

assessed by the Distress Index (DI) from the Counseling Center Assessment of Psychological Symptoms Instruments (CCAPS; Locke et al., 2011, 2012).

Skill Assessment System

We used a previously trained skill assessing system to automatically code therapist skills (Flemotomos et al., 2021). The system is composed of an automatic pipeline that transcribes session audio, predicts a single skill label (e.g., open questions) for each therapist utterance, and evaluates session-level scores (e.g., empathy) for an entire session.

As described in Flemotomos et al., 2021, the transcription system follows five steps: voice activity detection, speaker recognition, transcription, speaker role assignment (therapist vs. client), and utterance segmentation. Voice activity detection (VAD) is used to distinguish human voice from background noise. The VAD module included a pre-trained feedforward neural network, fine-tuned on 26 psychotherapy recordings from a university counseling center (same setting as the dataset in this study). The accuracy of the fine-tuned model on two test sets (20 and 92 sessions) from the same counseling center was .85 and .82, respectively. The speaker recognition module first calculated the similarity of each pair of human voice segments with a probabilistic linear discriminant analysis. The segments were then clustered into two groups (speakers) using agglomerative hierarchical clustering, a bottom-up approach to clustering similar segments. The error rates of clustering on the two test sets were .07 and .08. The transcribing module consists of two components: an acoustic model (i.e., predict phonetic units by acoustic information) and a language model (i.e., predict word sequence in language). The acoustic model used a time-delay neural network. The language model used a weighted ensemble of two 3-gram language models (i.e., language models in the unit of 3-word sequences) - one trained with 300k psychotherapy utterances from a subscribed archive of therapy sessions, the

other with open-sourced telephone conversations that served as a background model. The word error rate (similar to 1 - accuracy but can be greater than 1) of the transcribing module was about .30, which is typical for machine-learning models on spontaneous medical conversations (Kodish-Wachs et al., 2018). Two ensembles of language models (therapist and client) were individually trained. Speaker roles were assigned by matching a voice segment to one of the two ensembles with the lowest perplexity, an entropy-based measure of discrepancy. The speaker role assignment module achieved perfect recognition on the two test sets. Finally, the utterance segmentation module aggregated adjacent segments with the same speaker role, and then segmented them into talk turns using DeepSegment[2].

Quality assurance of included sessions was based on the length of a session (1 minute to 5 hours), the ratio of voice vs. silence (> 1:3), segment lengths (< 20s), minimum speaker time (10%), and a comparison with manual transcriptions. The percentage of the transcripts satisfying these criteria was 83.7%. These transcripts were included in the final dataset (see Flemotomos et al., 2021).

The skill label system used a type of neural network called a bidirectional long short-term memory (BiLSTM; Singla et al., 2018) with attention layers (Vaswani et al., 2017). The attention layers helped focus on salient words with useful information to predict labels. The rating of session-level scores was trained on a support vector regressor, a type of SVM that outputs continuous variables. The system was first trained on 242 sessions of psychotherapy transcripts from six clinical trials (Baer et al., 2009; Krupski et al., 2012; Lee et al., 2013, 2014; Neighbors et al., 2012; Tollison et al., 2008) coded with Motivational Interviewing Skill Code 2.5 (Houck et al., 2013), and then fine-tuned on 50 sessions from the university counseling center coded with

[2] github.com/notAI-tech/deepsegment

the same coding scheme. We selected three supportive counseling skills: empathy, open questions, and reflections. The reliability for each skill is described in the following section.

Empathy

Empathy is a session-level score, defined as the extent to which therapists understand or attempt to understand clients' perspective (Houck et al., 2013), ranging from 1 (low) to 5 (high), with a low score indicating therapists showing little interest in clients' experiences, and a high score indicating therapists actively attending to clients. Session-level scores like empathy are evaluated for an entire session. Manual rating requires raters to read through the entire session and offer a holistic judgment. The system generates a numeric score for each session that mimics this manual rating process. The accuracy within one Likert-scale discrepancy was .85 (Flemotomos et al., 2021).

Open Questions

Open question is an utterance-level label, and is a subcategory of the 'question' label. An utterance is labeled as an open question if 1) it is intended to gather information, or understand or elicit clients' stories, and 2) it leaves latitude for response rather than being answered by yes or no (Houck et al., 2013). Examples of open questions can be "How might you do that?" or "Tell me more." Human raters label the presence or absence of this skill for each utterance. The system was trained to mimic this process by labeling an utterance as an open question or not. The system has an F_1 score of .83 relative to human labeling (Flemotomos et al., 2021). F_1 score is a standard metric in machine learning that represents the agreement between models and humans, calculated as the harmonic mean of sensitivity and positive predictive value (1 - false positive rate). An F_1 score of .83 means that by averaging the proportion of correctly identified labels out

of all the true labels, and the proportion of correctly identified labels out of all the predicted labels, the system identified 83% of them correctly.

Reflection

The system provides scores for two types of reflections, complex and simple. In addition to reflecting back what a client said, complex reflections add significant meaning or emphasis to the client statement, conveying a deeper picture of the client statement. Simple reflection simply reflects or paraphrases what the client said without adding significant meaning or emphasis. For the current study, we combined both reflection scores into a single 'reflection' code.[3] The F_1 score of the combined reflect code was .61 (Flemotomos et al., 2021).

Counseling Center Assessments of Psychological Symptoms

Counseling Center Assessment of Psychological Symptoms-62 (CCAPS-62) is a comprehensive outcome assessment for college counseling centers (Locke et al., 2011). The assessment includes eight subscales of common mental health issues among college students. CCAPS-34 is a short form of CCAPS-62 that assesses seven key aspects of mental health of college students (Locke et al., 2012). We used the Distress Index (DI), which is an aggregate score of all the subscales (Locke et al., 2011, 2012), indicating a client's overall symptom level. CCAPS-62 was administered at intake and CCAPS-34 was administered at each subsequent session, including the final session. We took the DI of the intake CCAPS-62 as a client's baseline symptom level, and the DI from the CCAPS-34 of the last session as the posttreatment symptom level. The two DIs were calculated using the same CCAPS subscales for comparability

[3] We initially analyzed complex and simple reflections separately. As both simple and complex reflections index a similar construct (i.e., reflection; see Houck et al. 2015), we combined the two codes to increase reliability. Inspection of inter-rater sets showed that confusion of the two labels was one of the major sources of lower reliability (Flemotomos et al., 2021). Regression results with complex reflections aligned with the combined reflection code, between-therapist β = -.10, $CI_{95\%}$ [-.19, -.01], within-therapist β = .01, $CI_{95\%}$ [-.06, .08]. The results of simple reflections were not significant, suggesting that the association of combined reflections and client outcome (Table 1) may mainly come from complex reflections.

(Youn et al., 2015) and were highly correlated in the current dataset, $r = .84$, $CI_{95\%}$ [.81, .87]. The DI variables in the dataset ranged from 0 to 4, with 0 indicating no symptoms at all and 4 very high-level distress. The two DI variables were used to calculate clients' symptom change over the course of the treatment.

Data Analysis

We averaged empathy scores over all the sessions for each client. Likewise, we calculated the proportions of open questions and reflections in therapist's utterances for each session, respectively, and then averaged the proportions of each skill over all the sessions for a client. The mean empathy scores, and the mean proportions of open questions, and reflection, were entered into the analysis.

We used multilevel modeling (MLM) to assess the associations between counseling skills and treatment outcomes (Raudenbush & Bryk, 2002). Empathy, open questions, and reflections were group-mean centered and grand-mean centered to represent the within-therapist and between-therapist effects, respectively. We also partitioned the within-therapist and between-therapist effect of pretreatment (baseline) DI because clients were not randomly assigned (Baldwin et al., 2007).[4]

The two-level model was specified with clients nested within therapists:

$$Y_{ij} = \gamma_{00} + \gamma_{10}(Z_{ij} - \bar{Z}_j) + \gamma_{01}(\bar{Z}_j - \bar{Z}) + \gamma_{20}(X_{1ij} - \bar{X}_{1j}) + \gamma_{02}(\bar{X}_{1j} - \bar{X}_1) + \gamma_{30}(X_{2ij} - \bar{X}_{2j}) + \gamma_{03}(\bar{X}_{2j} - \bar{X}_2) + \gamma_{40}(X_{3ij} - \bar{X}_{3j}) + \gamma_{04}(\bar{X}_{3j} - \bar{X}_3) + u_{0j} + \epsilon_{ij}$$

[4] We also tested the nonlinear effects of each skill up to cubic terms (Kivlighan & Shaughnessy, 2000), and compared the models using likelihood ratio test and BIC. The likelihood ratio test indicated no significant difference between the linear, quadratic, and cubic models (*p*s = .20 and .08). Therefore, we report the results from the most parsimonious model - linear model with the lowest BIC = 1149.3 (Quadratic model BIC = 1179.2; cubic model BIC = 1206.3).

Y_{ij} is the posttreatment DI of client *i* treated by therapist *j*. *Z* represents the pretreatment DI to be controlled as a covariate. X_{1ij} is the empathy score of therapist *j* with client *i*. $\bar{X}_{1j}$ is the mean empathy score of therapist *j* across all their clients. $\bar{X}_{1}$ is the grand-mean empathy of all therapists. X_2 and X_3 represent open questions and reflections, respectively. γ_{n0} are linear fixed effects of within-therapist variables (n = 1, 2, 3, 4). γ_{0m} are linear fixed effects of between-therapist variables (m = 1, 2, 3, 4). u_{0j} is the between-therapist random effect in the intercept that is unexplained by the predictors. ε_{ij} is the within-therapist random effect (residual).

We examined statistical significance and effect sizes (standardized β) of the fixed effects to determine the associations between the predictors and clients' outcome. Because empathy (1-5) and the other three measures (0-1) are inherently on different scales, we used standardized coefficients (β) to compare effect sizes across scales. We also examined the associations between the predictors and the baseline DI to explore unassessed confounds.

Transparency and Openness

Because the clients did not consent to make their clinical data public, we are unable to provide open access to the dataset. All the analyses were performed in *R*, version 4.1.0 (R Core Team, 2021). Multilevel models were built using *lmerTest*, version 3.1-3 (Kuznetsova et al., 2017). Adjusted *ICC*s were calculated using *performance*, version 0.10.1 (Lüdecke et al., 2021). Standardized coefficients, effect sizes, and their confidence intervals were calculated using *effectsize*, version 0.8.2 (Ben-Shachar et al., 2020). Code for the current study is included in the supplemental material. This study's design and analysis were not pre-registered.

Results

The average baseline DI was 1.88, *SD* = 0.70. The average posttreatment DI was 1.26, *SD* = 0.73. Clients' pre-post symptom reduction was large, Cohen's *d* = -0.93, $CI_{95\%}$ [-1.02, -

0.83]. The average of empathy was 3.86 out of 5. On average, open questions made up 3.9% of a therapist's statements, and reflections 14.7%. For process-outcome correlations, empathy (Spearman's ρ = -.08), open questions (-.06), and reflection (-.09) were negatively correlated with posttreatment DI. For process-process correlations, empathy was positively correlated with open questions (.09) and reflections (.16). Open questions were positively correlated with reflections (.50). See Table S3 for additional details.

Multilevel Model

See Table 1 for specific model estimates. After controlling for baseline DI, higher between-therapist open question ratio was associated with lower post-treatment DI, β = -.08, $CI_{95\%}$ [-.16, -.004]. Higher between-therapist reflection ratio was associated with lower post-treatment DI, β = -.09, $CI_{95\%}$ [-.17, -.01]. The finding indicates that therapists who used more open questions or reflections by ratio had lower clients' posttreatment symptom levels than other therapists. Within-therapist open question ratio (β = .03 [-.05, .11]) and reflection ratio (β = -.03 [-.11, .05]) were not associated with client outcome. Figures 1 and 2 illustrate the relationships among between-therapist open question ratio, between-therapist reflection ratio, and post-treatment DI after controlling for baseline DI.

INSERT TABLE 1 ABOUT HERE

Empathy score was unrelated to posttreatment DI, between-therapist β = .03, $CI_{95\%}$ [-.04, .10], within-therapist -.03 [-.10, .03].

INSERT FIGURE 1 ABOUT HERE

INSERT FIGURE 2 ABOUT HERE

To explore potential confounds, we examined the relationship between baseline DI and empathy, reflections, and open questions (Table S5). We found a negative association between

baseline DI and within-therapist empathy, b = -0.27, $CI_{95\%}$ [-0.47, -0.07]. This is consistent with the previous findings that client baseline factors such as baseline severity (Imel et al., 2011) or client aggressiveness (Boswell et al., 2015) were associated with therapist use of some skills. We did not find associations with baseline DI for open question and reflection ratios, implying that their associations with client improvement were unlikely to be biased by unassessed confounds that would impact both the skills and the client symptom levels. No multicollinearity issue was found, *VIF*s = [1.01, 1.49]. The normality of random effects and residuals were checked by *qq*-plot. Our findings supported Hypotheses 2a and 3a. The rest of the hypotheses were unsupported.

Discussion

Although supportive counseling accounts for a considerable proportion of therapeutic change (Cuijpers et al., 2012), the relationship between individual supportive counseling skills and outcomes are unclear. The current study complements previous work on therapists' supportive skill use and client outcomes, including studies on empathy (Elliott et al., 2011, 2018), open questions (Williams, 2023), and reflections (Elliott et al., 2023). Limitations of prior studies include not considering therapist differences in using these skills, small sample sizes, and heterogeneous skill coding schemes. The present study addressed these limitations by leveraging machine-learning models trained to evaluate counseling skills. We applied a multilevel model to a large naturalistic dataset that was rated by a previously trained machine-learning-based skill assessment system.

Hypotheses on the association of empathy and outcome were not supported (Hypotheses 1a and 1b). Although empathy ratings are consistently associated with client outcomes (Elliott et al., 2011, 2018), we did not observe this association in our data. However, these meta-analyses primarily included studies that relied on client perception of therapist empathy, which often does

not correspond to observational ratings (Gurman, 1977). In prior meta-analyses with observational rating systems similar to the one used in this study, there was not a significant association between empathy and symptom reduction (Pace et al., 2017; Magill et al., 2018). Further, a previous meta-analysis found an inverse relationship between client sample size and empathy-outcome correlation (Elliott et al., 2018). Some prior studies with large sample sizes also failed to find significant relationships (e.g., Dormaar et al., 1989; Gillispie et al., 2005). The present study with a large sample (client n = 610) and a small effect size appeared to be consistent with these studies. One possible reason for the lack of association may be the low variability in the empathy measure. As shown in Table S3, the standard deviation of empathy was 0.33 - only about 8.6% of the mean score. The lack of variability could have made it difficult to detect an association with client outcome, especially in large-scale studies where variability is less likely to be only the result of sampling error. Another possible reason for the lack of association may be heterogeneity in the client sample. Elliott et al. (2018) found significant heterogeneity in empathy-outcome associations across clients' symptom types. Clients in this sample presented with various diagnoses, which might be moderators of the relationship between empathy and outcomes. It may be beneficial to assess empathy-outcome associations in a therapeutic context where there is a greater range of empathy (e.g., substance use treatments that use a confrontational style; see Moyers & Miller, 2013) or add moderators to future analyses.

Our findings suggested positive associations for between-therapist open question ratio, between-therapist reflection ratio, and client improvement (Hypotheses 2a and 3a). No within-therapist associations were found. In other words, if a therapist used a higher rate of open questions than other therapists, on average, their clients had better outcomes than other

therapists' clients. But therapist differences in the rate of open questions within their caseload was not associated with differences in client outcome. This pattern of findings presents a nuanced interpretation of how these specific skills might be related to improvement that is similar to previous therapist effects for the working alliance (e.g., Baldwin et al., 2007). Specifically, if a client was seen by a therapist who generally asked more open questions or reflections, their outcome was more likely to be positive. However, the absolute level of open questions or reflections experienced by that client was not a predictor of outcome. It is possible that therapists who use open questions or reflections at a high rate will have better outcomes than those who do not use open questions as often. However, some clients may require more or less of these skills depending on their interpersonal or symptomatic presentation - resulting in the lack of within-therapist associations. It may be helpful to conduct further studies to investigate the causal relationships between the two skills and client outcome. Further, this pattern of results suggest that research that focuses on therapist use of specific skills may miss important information if they focus solely on the total correlation between outcome and the skill.

Our findings about reflections did not align with a recent meta-analysis (Elliott et al., 2023) that indicated reflections were not associated with outcomes. A potential reason may be the different clientele between the studies included in the meta-analysis and the present study. The majority of the prior studies on reflections focused on substance use or unprotected sex behaviors in the context of motivational interviewing (Elliott et al., 2023). However, the primary concerns of the clients in our dataset were anxiety, depression, and academic distress. Substance use (4.08%) and sexual concerns (4.29%) made up only a very small portion. It may be possible that reflections as a supportive and exploratory skill are more helpful for clients with particular

disorders or who may benefit from facilitated self-exploration. It may be interesting to assess reflections with different client populations and concerns.

The present study suggests that such automatic measures can help clarify how process is related to outcome. By leveraging a previously trained machine-learning-based skill assessment system, we overcame the time and labor barriers of large-scale observer-rated research on therapists' interventions. In the past two decades, machine learning research related to psychotherapy has been mainly focused on developing automated measures for therapy process and outcome variables (Aafjes-van Doorn et al., 2020). The current study and other recent work in text-based psychotherapy (e.g., Ewbank et al., 2019) highlights that these measures can be utilized to understand how the treatment process is related to outcome on a scale and level of specificity that was previously challenging. The application of machine learning techniques that can immediately evaluate the content of a conversation may eventually include using feedback tools in real time to nudge the therapist to use (or not use) specific interventions, allowing experimental manipulation of therapist behavior within sessions. Such designs will rely on an infrastructure of reliable and clinically validated measures.

Limitations and Future Directions

Besides the limitations mentioned previously, there are several other notable limitations. First, some of our machine-learning based measures may benefit from continuing to improve F_1 scores for the specific process measures. For example, complex and simple reflections tend to be more difficult to differentiate and thus tend to have lower inter-rater reliability (Tollison et al., 2008). To address the low F_1 score of complex and simple reflections, we combined the two codes into a single reflection code, which increased the F_1 score to .61. Although the F_1 score of this combined code is still lower than empathy and open questions, the significance of its

association with client outcome suggest that this measure remains clinically meaningful. However, one drawback of combining the two codes is that it prevents the understanding of the differential effects of complex reflection and simple reflection. In addition, the word error rate of the transcription system was about 30%, which could partly explain some errors in the subsequent code prediction. We expect the associations to be stronger and more robust once more advanced machine learning methods are used to update our skill assessment system.

Second, our model only predicted a single skill code per therapist statement (Flemotomos et al., 2021), which could miss important skills if a therapist statement includes multiple skills. For example, if a therapist says "I heard anger in your tone when you said to your partner 'this is nonsense,' which reminds me of how you mentioned your mom talked to you," the system may only predict reflection but miss the important psychodynamic interpretation of this statement.

To improve the performance of machine-learning models that assess psychotherapy process, researchers might explore new state-of-art machine learning methods as they emerge (Wolf et al., 2020). Transformer models are being updated at a rapid pace, and may offer improvements in accuracy compared to previous methods. For example, a recent study using transformer-based methods to assess the association of session transcripts and session outcomes (Kuo et al., 2023). To increase granularity or allow multi-labeling, researchers may also adopt hierarchical algorithms (see Flemotomos et al., 2021). Researchers may also use multimodal machine learning that incorporates paraverbal features such as prosodies and tones (Tavabi et al., 2019), or consider immediate contexts that include both client and therapist statements in label prediction (Broadbent et al., 2023). In addition, it may be beneficial to choose highly reliable psychological measures to train machine-learning models, to reduce the impacts of the inherent

uncertainty (i.e., interrater discrepancy) in a measure on the model being trained (see Flemotomos et al., 2021).

Third, this study focused only on the frequency of the skills based on the verbal content. Our automatic system could miss key information about the quality of an intervention. What is said, when to say it, and how the intervention is delivered can all be important to explore (Stiles, 1988; Williams, 2023). However, studies at this level of granularity may be limited by low occurrence rates of labels and excessive resource demand in rating. For example, an appropriately timed therapist self-disclosure may be very rare (perhaps occuring once in 5-10 sessions), though it can be transformative to clients. Our study is an example of leveraging previously trained machine-learning models to reduce resource demand and scale up study so that rare labels may also have adequate statistical power.

Fourth, this study did not include other nonspecific factors that are known to be associated with outcomes, such as alliance (Alldredge et al., 2021; Flückiger et al., 2018; Horvath & Symonds, 1991; Probst et al., 2019). It may be possible that some null findings were a result of a moderating effect by alliance. For example, clients could at times feel “challenged” and “negative” about a therapist’s open questions (Hill et al., 1988), suggesting the importance of considering bond and task/goal agreement when a question is asked. Future studies may benefit from considering the moderating or mediating effects of process factors that are associated with outcomes.

Finally, this study did not include session-by-session temporal associations of supportive counseling skills and outcomes, which could reveal useful immediate association of skills with the treatment progress. Statistical models such as cross-lagged panel models or latent growth curve models may help reveal such patterns. However, these models often need sufficient

consecutive data. Despite the scale of observational ratings in the current dataset, the number of consecutive recordings within clients is limited due to the naturalistic setting from which the data were collected. Some therapists recorded regularly (e.g., required for counseling trainees) and others recorded sporadically. Some clients entered treatment late in the semester with only a few sessions left with their training therapists. As a result, the number of clients with at least two recorded sessions was 610 (72.6% of the total number of clients in the original data), and dropped drastically to 205 with at least six sessions (24.4%). To obtain sufficient consecutive recordings, researchers may look for protocol-based treatments where there is often a set number of sessions per treatment, or clinical settings where recording and routine outcome monitoring is required for most of the clinicians.

Conclusion

This study illustrated the feasibility of scaling up observer-rated studies on supportive counseling skills using a previously trained machine-learning skill assessing system. We found a positive association between both open question and reflection ratios and client improvement. This result offered more nuanced evidence regarding the association of supportive counseling skills with outcome beyond previous research on the total correlation, which revealed mixed or no correlations. We could not replicate the previous findings on empathy-outcome associations, possibly due to the type of measure (observer-rated vs. self-reported), low variability of the measure, and sample heterogeneity. This line of research may benefit from further improvement on the reliability of machine-learning models.

References

Aafjes-van Doorn, K., Kamsteeg, C., Bate, J., & Aafjes, M. (2020). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research*, *31*(1), 1–33. https://doi.org/10.1080/10503307.2020.1808729

Alldredge, C. T., Burlingame, G. M., Yang, C., & Rosendahl, J. (2021). Alliance in group therapy: A meta-analysis. *Group Dynamics: Theory, Research, and Practice*, *25*(1), 13–28. https://doi.org/10.1037/gdn0000135

Areán, P. A., Raue, P., Mackin, R. S., Kanellopoulos, D., McCulloch, C., & Alexopoulos, G. S. (2010). Problem-solving therapy and supportive therapy in older adults with major depression and executive dysfunction. *American Journal of Psychiatry*, *167*(11), 1391–1398. https://doi.org/10.1176/appi.ajp.2010.09091327

Baer, J. S., Wells, E. A., Rosengren, D. B., Hartzler, B., Beadnell, B., & Dunn, C. (2009). Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors. *Journal of substance abuse treatment*, *37*(2), 191-202

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 258–297). Wiley.

Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the alliance-outcome correlation: Exploring the relative importance of therapist and patient variability in the alliance. *Journal of Consulting and Clinical Psychology*, *75*(6), 842–852. https://doi.org/10.1037/0022-006X.75.6.842

Boswell, J. F., Gallagher, M. W., Sauer-Zavala, S. E., Bullis, J., Gorman, J. M., Shear, M. K., Woods, S., & Barlow, D. H. (2013). Patient characteristics and variability in adherence and competence in cognitive-behavioral therapy for panic disorder. *Journal of Consulting and Clinical Psychology*, *81*(3), 443–454. https://doi.org/10.1037/a0031437

Broadbent, M., Medina Grespan, M., Axford, K., Zhang, X., Srikumar, V., Kious, B., & Imel, Z. (2023). A machine learning approach to identifying suicide risk among text-based crisis counseling encounters. *Frontiers in Psychiatry*, *14*, 1110527. https://doi.org/10.3389/fpsyt.2023.1110527

Bromberg, P. M. (1998). *Standing in the spaces: Essays on clinical process trauma and dissociation*. Routledge.

Can, D., Marín, R. A., Georgiou, P. G., Imel, Z. E., Atkins, D. C., & Narayanan, S. S. (2016). "It sounds like...": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology*, *63*(3), 343–350. https://doi.org/10.1037/cou0000111

Constantino, M. J., Coyne, A. E., & Muir, H. J. (2020). Evidence-based therapist responsivity to disruptive clinical process. *Cognitive and Behavioral Practice*, *27*(4), 405–416. https://doi.org/10.1016/j.cbpra.2020.01.003

Conte, H. R. (1994). Review of research in supportive psychotherapy: An update. *American Journal of Psychotherapy*, *48*(4), 494–504. https://doi.org/10.1176/appi.psychotherapy.1994.48.4.494

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Creed, T. A., Wolk, C. B., Feinberg, B., Evans, A. C., & Beck, A. T. (2016). Beyond the label: Relationship between community therapists' self-report of a cognitive behavioral therapy orientation and observed skills. *Administration and Policy in Mental Health and Mental Health Services Research*, *43*(1), 36–43. https://doi.org/10.1007/s10488-014-0618-5

Crits-Christoph, P., Baranackie, K., Kurcias, J., Beck, A., Carroll, K., Perry, K., Luborsky, L., McLellan, A., Woody, G., Thompson, L., Gallagher, D., & Zitrin, C. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research*, *1*(2), 81–91. https://doi.org/10.1080/10503309112331335511

Cuijpers, P., Driessen, E., Hollon, S. D., van Oppen, P., Barth, J., & Andersson, G. (2012). The efficacy of non-directive supportive therapy for adult depression: A meta-analysis. *Clinical Psychology Review*, *32*(4), 280–291. https://doi.org/10.1016/j.cpr.2012.01.003

Dormaar, M., Dijkman, C. I. M., & De Vries, M. W. (1989). Consensus in patient-therapist interactions. *Psychotherapy and Psychosomatics*, *51*(2), 69–76. https://doi.org/10.1159/000288138

Elliott, R. (1985). Helpful and nonhelpful events in brief counseling interviews: An empirical taxonomy. *Journal of Counseling Psychology*, *32*(3), 307–322. https://doi.org/10.1037/0022-0167.32.3.307

Elliott, R., Bohart, A. C., Watson, J. C., & Greenberg, L. S. (2011). Empathy. *Psychotherapy*, *48*(1), 43.

Elliott, R., Bohart, A. C., Watson, J. C., & Murphy, D. (2018). Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, *55*(4), 399.

Elliott, R., Bohart, A., Larson, D., Muntigl, P., & Smoliak, O. (2023). Empathic reflections by themselves are not effective: Meta-analysis and qualitative synthesis. *Psychotherapy Research*, 1–17. https://doi.org/10.1080/10503307.2023.2218981

Ewbank, M. P., Cummins, R., Tablan, V., Bateup, S., Catarino, A., Martin, A. J., & Blackwell, A. D. (2019). Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiatry*, *77*(1), 35. https://doi.org/10.1001/jamapsychiatry.2019.2664

Flemotomos, N., Martinez, V. R., Chen, Z., Singla, K., Ardulov, V., Peri, R., Caperton, D. D., Gibson, J., Tanana, M. J., Georgiou, P., Van Epps, J., Lord, S. P., Hirsch, T., Imel, Z. E., Atkins, D. C., & Narayanan, S. (2021). Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01623-4

Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, *55*(4), 316–340. https://doi.org/10.1037/pst0000172

Frank, G. H., & Sweetland, A. (1962). A study of the process of psychotherapy: The verbal interaction. *Journal of Consulting Psychology*, *26*(2), 135–138. https://doi.org/10.1037/h0047799

Gibson, J., Can, D., Xiao, B., Imel, Z. E., Atkins, D. C., Georgiou, P., & Narayanan, S. S. (2016). A deep learning approach to modeling empathy in addiction counseling. *INTERSPEECH*, 1447–1451. https://doi.org/10.21437/Interspeech.2016-554

Gillispie, R., Williams, E., & Gillispie, C. (2005). Hospitalized African American mental health consumers: Some antecedents to service satisfaction and intent to comply with aftercare.

American Journal of Orthopsychiatry, *75*(2), 254–261. https://doi.org/10.1037/0002-9432.75.2.254

Gurman, A. S. (1977). Therapist and patient factors influencing the patient's perception of facilitative therapeutic conditions. *Psychiatry*, *40*(3), 218–231.

Hill, C. E. (1978). Development of a counselor verbal response category. *Journal of Counseling Psychology*, *25*(5), 461–468. https://doi.org/10.1037/0022-0167.25.5.461

Hill, C. E. (1992). Research on therapist techniques in brief individual therapy: Implications for practitioners. *The Counseling Psychologist*, *20*(4), 689–711. https://doi.org/10.1177/0011000092204012

Hill, C. E., Helms, J. E., Tichenor, V., Spiegel, S. B., O'Grady, K. E., & Perry, E. S. (1988). Effects of therapist response modes in brief psychotherapy. *Journal of Counseling Psychology*, *35*(3), 222–233. https://doi.org/10.1037/0022-0167.35.3.222

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Horvath, A. O., & Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology*, *38*(2), 139–149. https://doi.org/10.1037/0022-0167.38.2.139

Houck, J., Moyers, T., Miller, W., Glynn, L., & Hallgren, K. (2013). Motivational interviewing skill code (MISC) 2.5. *Painamaton Julkaisu. Haettu*, *26*, 2015.

Imel, Z. E., Baer, J. S., Martino, S., Ball, S. A., & Carroll, K. M. (2011). Mutual influence in therapist competence and adherence to motivational enhancement therapy. *Drug and Alcohol Dependence*, *115*(3), 229–236. https://doi.org/10.1016/j.drugalcdep.2010.11.010

Imel, Z. E., Caperton, D. D., Tanana, M., & Atkins, D. C. (2017). Technology-enhanced human interaction in psychotherapy. *Journal of Counseling Psychology*, *64*(4), 385–393. https://doi.org/10.1037/cou0000213

Kivlighan, D. M., Jr., & Shaughnessy, P. (2000). Patterns of working alliance development: A typology of client's working alliance ratings. *Journal of Counseling Psychology*, *47*(3), 362–371. https://doi.org/10.1037/0022-0167.47.3.362

Kodish-Wachs, J., Agassi, E., Kenny III, P., & Overhage, J. M. (2018). A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. *AMIA Annual Symposium Proceedings*, *2018*, 683.

Kuo, P. B., Tanana, M. J., Goldberg, S. B., Caperton, D. D., Narayanan, S., Atkins, D. C., & Imel, Z. E. (2023). Machine-learning-based prediction of client distress from session recordings. *Clinical Psychological Science*, *0*(0). https://doi.org/10.1177/21677026231172694

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). *lmerTest* package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Kadur, J., Lüdemann, J., & Andreas, S. (2020). Effects of the therapist's statements on the patient's outcome and the therapeutic alliance: A systematic review. *Clinical Psychology & Psychotherapy*, *27*(2), 168–178. https://doi.org/10.1002/cpp.2416

Krupski, A., Joesch, J. M., Dunn, C., Donovan, D., Bumgardner, K., Lord, S. P., ... & Roy-Byrne, P. (2012). Testing the effects of brief intervention in primary care for problem drug use in a randomized controlled trial: rationale, design, and methods. *Addiction science & clinical practice*, *7*, 1-10

Lee, C. M., Neighbors, C., Lewis, M. A., Kaysen, D., Mittmann, A., Geisner, I. M., ... & Larimer, M. E. (2014). Randomized controlled trial of a Spring Break intervention to reduce high-risk drinking. *Journal of consulting and clinical psychology*, *82*(2), 189.

Lee, C. M., Kilmer, J. R., Neighbors, C., Atkins, D. C., Zheng, C., Walker, D. D., & Larimer, M. E. (2013). Indicated prevention for college student marijuana use: A randomized controlled trial. *Journal of consulting and clinical psychology*, *81*(4), 702

Locke, B. D., Buzolitz, J. S., Lei, P.-W., Boswell, J. F., McAleavey, A. A., Sevig, T. D., Dowis, J. D., & Hayes, J. A. (2011). Development of the counseling center assessment of psychological symptoms-62 (CCAPS-62). *Journal of Counseling Psychology*, *58*(1), 97–109. https://doi.org/10.1037/a0021282

Locke, B. D., McAleavey, A. A., Zhao, Y., Lei, P.-W., Hayes, J. A., Castonguay, L. G., Li, H., Tate, R., & Lin, Y.-C. (2012). Development and initial validation of the counseling center assessment of psychological symptoms–34. *Measurement and Evaluation in Counseling and Development*, *45*(3), 151–169. https://doi.org/10.1177/0748175611432642

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An *R* package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(59), 3112. https://doi.org/10.31234/osf.io/vtq8f

Magill, M., Apodaca, T. R., Borsari, B., Gaume, J., Hoadley, A., Gordon, R. E. F., Tonigan, J. S., & Moyers, T. (2018). A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of Consulting and Clinical Psychology*, *86*(2), 140–157. https://doi.org/10.1037/ccp0000250

Moyers, T. B., & Miller, W. R. (2013). Is low therapist empathy toxic? *Psychology of Addictive Behaviors*, *27*(3), 878–884. https://doi.org/10.1037/a0030274

Neighbors, C., Lee, C. M., Atkins, D. C., Lewis, M. A., Kaysen, D., Mittmann, A., ... & Larimer, M. E. (2012). A randomized controlled trial of event-specific prevention strategies for reducing problematic drinking associated with 21st birthday celebrations. *Journal of consulting and clinical psychology*, *80*(5), 850

Pace, B. T., Dembe, A., Soma, C. S., Baldwin, S. A., Atkins, D. C., & Imel, Z. E. (2017). A multivariate meta-analysis of motivational interviewing process and outcome. *Psychology of Addictive Behaviors*, *31*(5), 524–533. https://doi.org/10.1037/adb0000280

Probst, G. H., Berger, T., & Flückiger, C. (2019). The alliance-outcome relation in internet-based interventions for psychological disorders: A correlational meta-analysis. *Verhaltenstherapie*, 1–12. https://doi.org/10.1159/000503432

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed). Sage Publications.

Ribeiro, A. P., Gonçalves, M. M., Silva, J. R., Brás, A., & Sousa, I. (2016). Ambivalence in narrative therapy: A comparison between recovered and unchanged cases. *Clinical Psychology & Psychotherapy*, *23*(2), 166–175. https://doi.org/10.1002/cpp.1945

Singla, K., Chen, Z., Flemotomos, N., Gibson, J., Can, D., Atkins, D. C., & Narayanan, S. (2018). Using prosodic and lexical information for learning utterance-level behaviors in psychotherapy. *Interspeech*, *2018*, 3413.

Stiles, W. B. (1978). Verbal response modes and dimensions of interpersonal roles: A method of discourse analysis. *Journal of Personality and Social Psychology*, *36*(7), 693.

Stiles, W. B. (1988). Psychotherapy process-outcome correlations may be misleading. *Psychotherapy: Theory, Research, Practice, Training*, *25*(1), 27–35. https://doi.org/10.1037/h0085320

Tavabi, L., Stefanov, K., Nasihati Gilani, S., Traum, D., & Soleymani, M. (2019). Multimodal learning for identifying opportunities for empathetic responses. *2019 International Conference on Multimodal Interaction*, 95–104. https://doi.org/10.1145/3340555.3353750

Tollison, S. J., Lee, C. M., Neighbors, C., Neil, T. A., Olson, N. D., & Larimer, M. E. (2008). Questions and reflections: The use of motivational interviewing microskills in a peer-led brief alcohol intervention for college students. *Behavior Therapy*, *39*(2), 183–194. https://doi.org/10.1016/j.beth.2007.07.001

Uhl, J., Schaffrath, J., Schwartz, B., Poster, K., & Lutz, W. (2022). Within and between associations of clinical microskills and correct application of techniques/strategies: A longitudinal multilevel approach. *Journal of Consulting and Clinical Psychology*, *90*(6), 478–490. https://doi.org/10.1037/ccp0000738

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Williams, E. N. (2023). The use of questions in psychotherapy: A review of research on immediate outcomes. *Psychotherapy*. https://doi.org/10.1037/pst0000471

Winston, A., Pinsker, H., & McCullough, L. (1986). A review of supportive psychotherapy. *Psychiatric Services*, *37*(11), 1105–1114. https://doi.org/10.1176/ps.37.11.1105

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2015). "Rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLOS ONE*, *10*(12), e0143055. https://doi.org/10.1371/journal.pone.0143055

Youn, S. J., Castonguay, L. G., Xiao, H., Janis, R., McAleavey, A. A., Lockard, A. J., Locke, B. D., & Hayes, J. A. (2015). The counseling center assessment of psychological symptoms (CCAPS): Merging clinical practice, training, and research. *Psychotherapy*, *52*(4), 432–441. https://doi.org/10.1037/pst0000029

Zhang, X., Tanana, M., Weitzman, L., Narayanan, S., Atkins, D., & Imel, Z. (2022). You never know what you are going to get: Large-scale assessment of therapists' supportive counseling skill use. *Psychotherapy*. https://doi.org/10.1037/pst0000460

Table 1

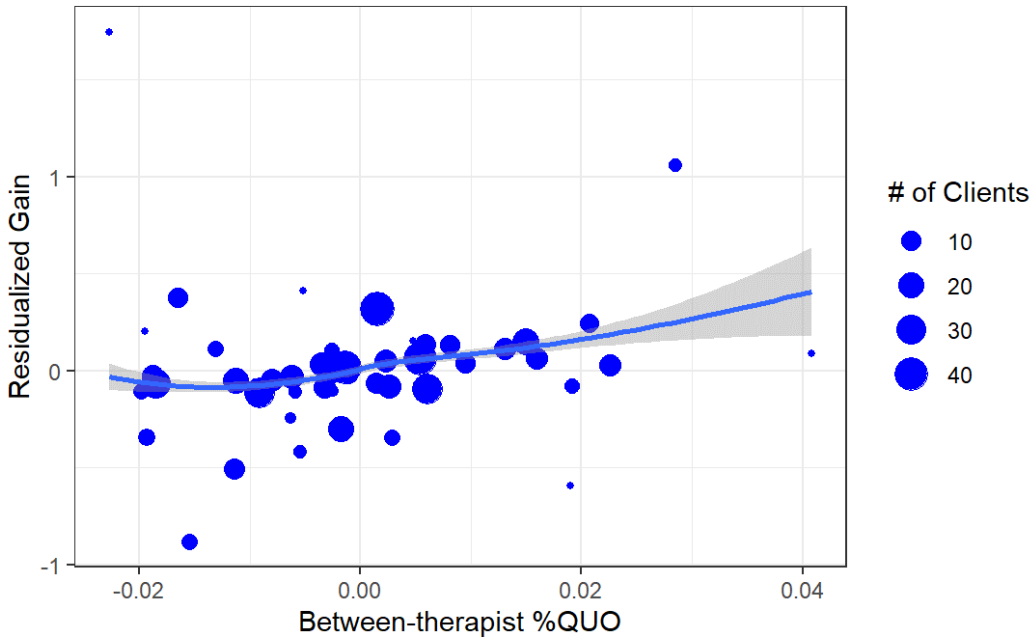
Multilevel Model of Client Outcome and Supportive Counseling Skills

Variable	Estimate [95% CI]	β [95% CI]	p
Fixed effects			
Intercept (γ_{00})	1.25 [1.18, 1.31]	.01 [-.07, .09]	
Pretreatment DI			
Within therapist (γ_{10})	0.58 [0.51, 0.65]	.53 [.47, .60]	<.001***
Between therapist (γ_{01})	0.60 [0.35, 0.85]	.18 [.10, .25]	<.001***
Empathy			
Within therapist (γ_{20})	-0.09 [-0.26, 0.08]	-.03 [-.10, .03]	.30
Between therapist (γ_{02})	0.13 [-0.17, 0.43]	.03 [-.04, .10]	.41
QUO			
Within therapist (γ_{30})	1.47 [-2.35, 5.29]	.03 [-.05, .11]	.45
Between therapist (γ_{03})	-5.76 [-11.1, -0.43]	-.08 [-.16, -.004]	.04*
REF			
Within therapist (γ_{40})	-0.79 [-2.75, 1.16]	-.03 [-.11, .05]	.43
Between therapist (γ_{04})	-3.07 [-5.75, -0.39]	-.09 [-.17, -.01]	.04*
Random effects			
Within therapist (σ_{ij}^2)	0.34 [0.30, 0.38]		
Between therapist (τ_j^2)	0.01 [0.00, 0.02]		
Adjusted *ICC*	.03		
Marginal R^2	.35		

Note. DI = distress index; QUO = open question; REF = reflection. * $p < .05$, ** $p < .01$, *** $p < .001$.

Figure 1

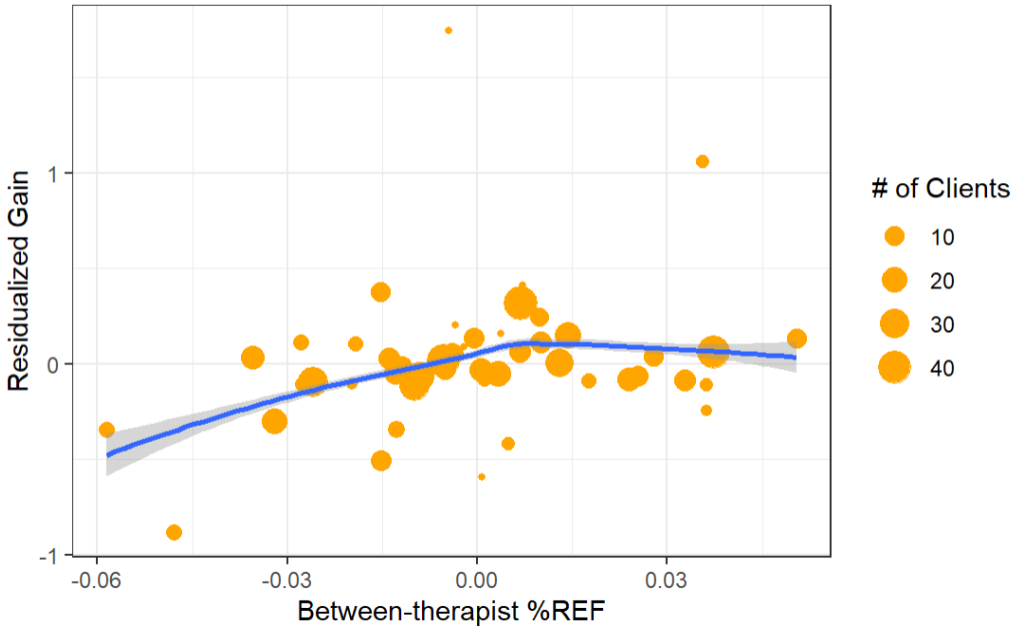
Between-Therapist Association of Residualized Gain and Proportion of Open Questions



Note. QUO = open question. Blue curve was fitted by locally estimated scatterplot smoothing (loess). One datapoint represents one therapist. The size of a datapoint represents the number of clients within a therapist. Therapists who used more open questions on average were associated with better clients' outcomes.

Figure 2

Between-Therapist Association of Residualized Gain and Proportion of Reflections



Note. REF = reflection. Blue curve was fitted by locally estimated scatterplot smoothing (loess). One datapoint represents one therapist. The size of a datapoint represents the number of clients within a therapist. For therapists who used relatively less reflection than other therapists, using more reflection was associated with better clients' outcomes.

Table S1

Demographic Information of Clients

Variable	*n*	%	*M* /*Mdn*	*SD*
Age (*n* = 596)			23.5 / 22	4.8
Gender (*n* = 607)				
Male	263	43.3		
Female	315	51.9		
Transgender	7	1.2		
Nonbinary	10	1.6		
Others	12	2.0		
Race/Ethnicity (*n* = 607)				
Asian/Asian American	48	7.9		
Black/African American	8	1.3		
Latino/Latina	53	8.4		
Native American/Alaskan	3	0.5		
Pacific Islander	1	0.2		
White/European American	455	75.0		
Multi-racial	32	5.3		
Others	7	1.5		
Sexual orientation (*n* = 597)				
Straight	427	71.5		
Lesbian	15	2.5		
Gay	34	5.7		
Bisexual	73	12.2		
Questioning	27	4.5		
Others	21	3.5		
Five most identified religions (*n* = 421)				
Latter-Day Saints	116	27.6		
Atheist	62	14.7		
Catholic	54	12.8		
Agnostic	45	10.7		
Others	27	6.4		
Five most endorsed concerns (*n* = 608) [a]				
Anxiety	422	69.4		
Depression	390	64.1		
Academic distress	266	43.8		
Self-esteem	255	41.9		
Loneliness	215	35.4		

Note. *N* = 610. Latter-Day Saints stands for the Church of Jesus Christ of Latter-Day Saints.

[a] Clients may endorse multiple concerns. Therefore, the percentages do not add up to 100%.

Table S2

Demographic Information of Therapists

Variable	*n*	%	*M* /*Mdn*	*SD*
Age (*n* = 38)			32.4 / 29	12.1
Gender (*n* = 37)				
Male	13	35.1		
Female	23	62.2		
Nonbinary	1	2.7		
Transgender	0	0.0		
Race/Ethnicity (*n* = 38)				
Asian/Asian American	5	13.2		
Black/African American	2	5.3		
Latino/Latina	3	7.9		
Native American/Alaskan	0	0.0		
Pacific Islander	0	0.0		
White/European American	24	63.2		
Multi-racial	3	7.9		
Others	1	2.6		
Sexual orientation (*n* = 36)				
Straight	24	66.7		
Lesbian or Gay	4	11.1		
Bisexual	4	11.1		
Questioning	1	2.8		
Others	3	8.3		
Five most identified theoretical orientations (*n* = 38) [a]				
Cognitive-behavioral	21	55.3		
Feminist-multicultural	20	52.6		
Interpersonal	16	42.1		
Humanistic	14	36.8		
Integrative	12	31.6		

Note. *N* = 48. Some therapists did not report demographic information. These therapists were still included in statistical analyses.

[a] Therapists may endorse multiple theoretical orientations. Therefore, the percentages do not add up to 100%.

Table S3

Means, Standard Deviations, and Zero-Order Correlations of Supportive Counseling Skills and Distress Index

Variable	*M*	*SD*	1	2	3	4
1. Empathy	3.85	0.33				
2. QUO	.039	.018	.09*			
3. REF	.147	.036	.16***	.50***		
4. Pretreatment DI	1.88	0.70	-.09*	.01	-.01	
5. Posttreatment DI	1.26	0.73	-.08*	-.06	-.09*	.56***

Note. QUO = proportion of open question; REF = proportion of reflection; DI = distress index. QUO, REF range from 0 to 1. Correlations are Spearman's ρ. *p*-values are Holm-Bonferroni corrected. * $p < .05$, ** $p < .01$, *** $p < .001$

Table S4

Multilevel Model of Supportive Counseling Skills, with Single-Client Therapists Removed

Variable	Estimate [95% CI]	β [95% CI]	p
Fixed effects			
Intercept (γ_{00})	1.24 [1.17, 1.31]	.01 [-.07, .09]	
Pretreatment DI			
Within therapist (γ_{10})	0.58 [0.51, 0.65]	.54 [.47, .60]	<.001***
Between therapist (γ_{01})	0.58 [0.31, 0.84]	.16 [.09, .24]	<.001***
Empathy			
Within therapist (γ_{20})	-0.09 [-0.26, 0.08]	-.03 [-.10, .03]	.30
Between therapist (γ_{02})	0.21 [-0.13, 0.54]	.05 [-.03, .12]	.25
QUO			
Within therapist (γ_{30})	1.47 [-2.36, 5.30]	.03 [-.05, .11]	.45
Between therapist (γ_{03})	-6.95 [-12.6, -1.25]	-.10 [-.18, -.02]	.02*
REF			
Within therapist (γ_{40})	-0.79 [-2.75, 1.16]	-.03 [-.11, .05]	.43
Between therapist (γ_{04})	-2.92 [-5.63, -0.20]	-.09 [-.17, -.003]	.04*
Random effects			
Within therapist (σ_{ij}^2)	0.35 [0.30, 0.38]		
Between therapist (τ_j^2)	0.01 [0.00, 0.02]		
Adjusted *ICC*	.03		
Marginal R^2	.35		

Note. DI = distress index; QUO = open question; REC = complex reflection; RES = simple reflection. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table S5

Associations of Supportive Counseling Skills and Baseline Distress Index

Variable	Estimate [95% CI]	*p*
Fixed effects		
Intercept	1.85 [1.77, 1.92]	<.001***
Empathy		
Within therapist	-0.27 [-0.47, -0.07]	.008**
Between therapist	0.20 [-0.16, 0.56]	.29
QUO		
Within therapist	0.57 [-3.95, 5.10]	.80
Between therapist	-1.49 [-7.94, 4.93]	.66
REF		
Within therapist	-0.23 [-2.09, 2.54]	.85
Between therapist	-0.16 [-3.42, 3.10]	.92

Note. DI = distress index; QUO = open question; REF = reflection. * $p < .05$, ** $p < .01$, *** $p < .001$.

R Codes for Multilevel Modeling

```
require(data.table)     # data processing
require(lmerTest)     # multilevel modeling, leveraging lmer4
require(effectsize)   # report effect sizes
require(performance)    # report R2 & icc
source('dataloader.R')    # call dataset named "dt"


# disaggregate pretreatment DI
## calculate group-mean centered part
dt[, pt_pretest_group := scale(pt_pretest, scale=F), by=therapistid]

## calculate grand-mean centered part
temp <- dt[, .(pt_pretest_grand=mean(pt_pretest, na.rm=T)),
by=therapistid][, pt_pretest_grand := scale(pt_pretest_grand,
scale=F)]
dt <- temp[dt, on='therapistid']
rm(temp)


# disaggregate reflection, open questions, and empathy
dt[, ave_REF_group := scale(ave_REF, scale=F), by=therapistid]
dt[, ave_QUO_group := scale(ave_QUO, scale=F), by=therapistid]
dt[, ave_empathy_group := scale(ave_empathy, scale=F), by=therapistid]

temp <- dt[, .(ave_REF_grand=mean(ave_REF, na.rm=T)),
by=therapistid][, ave_REF_grand := scale(ave_REF_grand, scale=F)]
dt <- temp[dt, on='therapistid']

temp <- dt[, .(ave_QUO_grand=mean(ave_QUO, na.rm=T)),
by=therapistid][, ave_QUO_grand := scale(ave_QUO_grand, scale=F)]
dt <- temp[dt, on='therapistid']

temp <- dt[, .(ave_empathy_grand=mean(empathy, na.rm=T)),
by=therapistid][, ave_empathy_grand := scale(ave_empathy_grand,
scale=F)]
dt <- temp[dt, on='therapistid']

rm(temp)

# build the model
eq <- '
```

```
pt_posttest ~ pt_pretest_group + pt_pretest_grand + ave_empathy_group
+ ave_empathy_grand + ave_QUO_group + ave_QUO_grand + ave_REF_group +
ave_REF_grand + (1 | therapistid)
'
model_mlm_intv <- lmer(eq, dt)


# check assumptions of the model
qqnorm(resid(model_mlm_intv))
qqnorm(ranef(model_mlm_intv))
plot(model_mlm_intv)


# report all the estimates
summary(model_mlm_intv)     # report model estimates
icc(model_mlm_intv)    # calculate adjusted ICC
r2(model_mlm_intv)     # calculate marginal R2
effectsize(model_mlm_intv)     # calculate beta
confint(model_mlm_intv)     # calculate CI of the estimates
```

Published Manuscripts that Involved the Same Dataset as in the Present Study

The data reported in this manuscript have been previously published. Findings from the data collection have been reported in separate manuscripts. Goldberg et al. (2020) focuses on building a natural language processing model to predict working alliance. Caperton (2021) focuses on building a statement-level adaptation of the Multitheoretical List of Therapeutic Interventions-30 (MULTI-30; Solomonov et al., 2019). Flemotomos et al. (2021) focuses on building a natural language processing model based on the Motivational Interviewing Skill Code 2.5 (MISC 2.5; Houck et al., 2015). Trevino et al. (2021) focuses on a qualitative analysis of cultural conversation in psychotherapy. Goldberg et al. (2022) focuses on the influence of outliers in alliance-outcome correlations. Zhang et al. (2022) focuses on the between-therapist and within-therapist variabilities of supportive counseling skills. Mehta et al. (2022) focuses on building a natural language processing model based on Caperton's (2021) statement-level adaptation of MULTI-30. Kuo et al. (2023) focuses on the prediction of the next session's outcome based solely on the transcript of a psychotherapy session using natural language processing. The current manuscript focuses on the between-therapist and within-therapist associations of supportive counseling skills and outcomes.