Appendix

Full Inclusion and Exclusion Criteria

For the broader clinical trial, patients were required to meet the following inclusion criteria: (1) age 18–65; (2) DSM 5-defined MDD, current episode (3) Hamilton Rating Scale for Depression-17 items (HRSD-17) score ≥ 14; (4) if currently on antidepressant medication, on a stable dose for at least one month at the time of study entry and agreeable to remaining on that dose for study duration; (5) ability to read and speak English fluently, (6) capacity to give informed consent. For *Phase 2* (remote teletherapy), the following inclusion criteria were added: (7) a computer or smartphone with a camera and microphone; (8) headphones or earbuds that connect to device; (9) a private area to complete psychotherapy sessions without interruption; (10) broadband internet connection that meets the bandwidth criteria of minimum of 10 Mbps for upload and download.

Patients were excluded from the broader clinical trial if they met any of the following criteria: (1) high risk for suicide, that, in the clinical opinion of the investigator, would warrant a higher level of care such as hospitalization or intensive outpatient programs; (2) current depressive episode has psychotic features; (3) current depressive episode has been present for >104 weeks; (4) substance use disorders, as defined by DSM 5, in the past 3 months, except for caffeine or nicotine; (5) prior manic or hypomanic episode (Bipolar I or II disorder) or a psychotic disorder including schizoaffective disorder or schizophrenia; (6) antisocial personality disorder; (7) significant, unstable, psychiatric co-morbidity that, in the opinion of the investigators, requires an alternative treatment approach (e.g., unstable eating disorder, unstable borderline personality disorder); (8) significant unstable medical illness that may explain depressive symptoms such as epilepsy, autoimmune disorder, chronic pain, or unstable endocrine disorder; (9) cognitive deficits that would preclude completion of study questionnaires or participation in psychotherapy; (10) unable or unwilling to comply with study requirements; (11) neurologic or medical condition that would interfere with nonverbal communication.

In the broader clinical trial, therapists were also research participants and consented to participation. Inclusion criteria for therapist-participants were: (1) currently functioning as a therapist or supervisor for the [redacted] clinic; (2) capacity to give informed consent. There were no exclusion criteria for therapists.

To be included in the current study, sessions needed to have both video recordings and working alliance ratings, and participants needed to have completed at least one session before exiting the study. Figure S1 depicts the flow and count of participants and sessions from recruitment to inclusion in the current study.

Recording Information

In Phase 1, separate tripod-mounted cameras captured each participant's full body at 59.94 frames per second. Patients' and therapists' videos were synchronized using hardware-provided timecodes and the OBS software. In Phase 2, the same Phase 1 cameras were used to capture therapists' upper body and head but patients' upper body and head were captured using their own person webcam. Transfer and synchronization was handled using HIPAA-compliant Zoom, and the OBS software was used to record the Zoom screen showing both participants at 30.00 frames per second. All videos of each participant were rescaled to a height of 720 pixels prior to analysis with OpenFace.

Modeling Details

Bayesian estimation and priors

We estimated our models using the *brms* R package (Bürkner, 2017) as a high-level interface to the Stan platform for statistical computing (Gelman, Lee, & Guo, 2015). Model estimation was performed using Markov chain Monte Carlo via the No-U-Turn Sampler algorithm (NUTS; Hoffman & Gelman, 2014), which converges quickly in high-dimensional models and eliminates the need for hand-tuning of hyperparameters. Each model used three chains of 10000 iterations each (with the first 8000 of each chain used for warm-up and the last 2000 used for inference). To aid convergence, we increased the `adapt_delta` and `max_treedepth` arguments to the more conservative values of 0.99

and 15, respectively. To assess convergence, we used the Rhat and ESS metrics as well as the presence or absence of divergent transition warnings from NUTS.

Bayesian estimation requires the analyst to define prior probability distributions for each model parameter. Our selection of priors was guided by the goal of ruling out unreasonable values (e.g., negative variances) without ruling out reasonable values. In the case of the main hypothesis tests, the most conservative (i.e., null) value was assigned the highest prior probability. For the slope parameters, we used standard normal priors to apply light regularization. For all other model parameters, we used the default weakly informative priors supplied by *brms*, which are presented in the model equations below.

Our point estimates were posterior medians and our interval estimates were posterior equal-tailed intervals. To calculate Bayesian p-values reported in the main paper, we calculated $1 - pd$ where pd is the "probability of direction" (i.e., the proportion of the posterior distribution that has the same sign as the posterior median). For details, see Makowski, Ben-Shachar, Chen, and Lüdecke (2019).

Transparency and Openness

Our sample size was determined by including all participants available in the dataset and then removing sessions that were missing video recordings or working alliance ratings (assumed to be missing completely at random). No other data exclusions were applied and no manipulations were used other than randomization to therapy type and assignment to a therapist. The most important instrument-measured constructs in the current study are the working alliance and depression severity. The former was measured by the patient-reported WAI-SR and the therapist-reported WAI-SR-T, both of which were used in the study. The latter was measured by the HRSD-17 and MADRS at screening/baseline, midpoint, and endpoint as well as by the QIDS-SR-16 before each session. We used the screening HRSD-17 score to control for each patient's baseline depression severity because it was listed as the primary outcome measure in the clinical trial. The complete list of measures is provided in the supplemental materials (i.e., the Assessment Grid file).

References

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10/gddxwp

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543. doi: 10/gfzk2m

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, *10*. doi: 10/ggfw2j

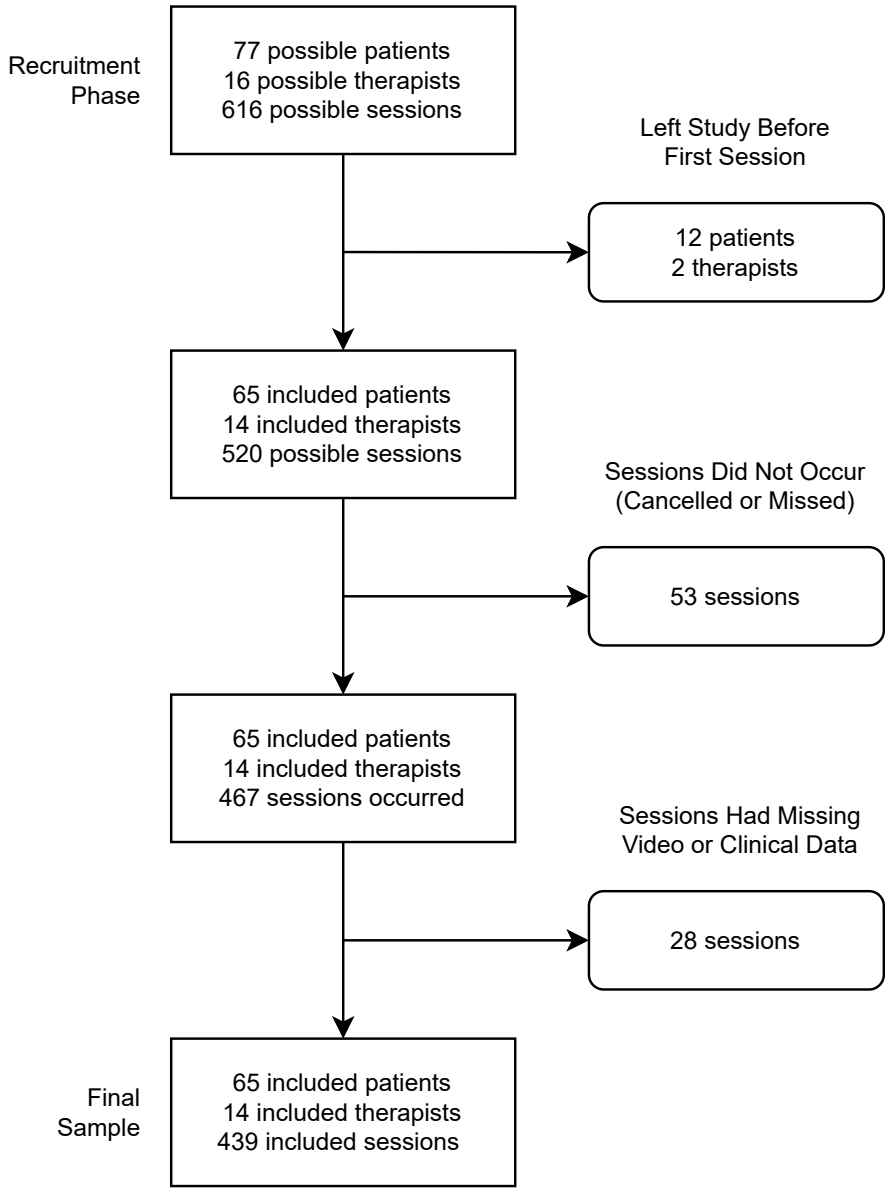

Figure S1

Participant Recruitment and Data Collection Flowchart