Published in final edited form as:

J Couns Psychol. 2019 April ; 66(3): 341–350. doi:10.1037/cou0000322.

Modeling empathy as synchrony in clinician and patient vocally-encoded emotional arousal: A failure to replicate

Jacques Gaume

Lausanne University Hospital

Kevin A. Hallgren

University of Washington

Carole Clair

Lausanne University Hospital

Marianne Schmid Mast

University of Lausanne

Valérie Carrard

Lausanne University Hospital

David C. Atkins

University of Washington

Abstract

Empathy is a well-defined active ingredient in clinical encounters. To measure empathy, the current gold standard is behavioral coding (i.e. trained coders attribute overall ratings of empathy to clinician behaviors within an encounter), which is labor intensive and subject to important reliability challenges. Recently, an alternative measurement has been proposed: capturing empathy as synchrony in vocally encoded arousal, which can be measured as the mean fundamental frequency of the voice (mean F0). This method has received preliminary support by one study (Imel et al. 2014). We aimed to replicate this study by using two large samples of clinical interactions (alcohol brief motivational interventions with young adults, $N = 208$; general practice consultations, $N = 204$). Audio files were segmented to identify respective speakers and mean F0 was measured using speech signal processing software. All sessions were independently rated by behavioral coders using two validated empathy scales. Synchrony between clinician and patient F0 was analyzed using multivariate multilevel models and compared to high and low levels of empathy derived from behavioral coding. Findings showed no support for our hypothesis that mean F0 synchrony between clinicians and patients would be higher in high-empathy sessions. This lack of replication was consistent for both clinical samples, both behavioral coding

Corresponding author: Jacques Gaume, Alcohol Treatment Center, Lausanne University Hospital, Avenue de Beaumont 21 bis, Bâtiment P2, 1011 Lausanne, Switzerland, Tel +41 21 314 41 05, Jacques.Gaume@chuv.ch.

Author Note

Jacques Gaume, Alcohol Treatment Center, Lausanne University Hospital, Lausanne, Switzerland; Kevin A. Hallgren, Department of Psychiatry and Behavioral Sciences, University of Washington; Carole Clair, Department of Ambulatory Care and Community Medicine, Lausanne University Hospital, Lausanne, Switzerland; Marianne Schmid Mast, Faculty of Business and Economics, University of Lausanne, Lausanne, Switzerland; Valérie Carrard, Department of Psychiatry, Lausanne University Hospital, Lausanne, Switzerland; David C. Atkins, Department of Psychiatry and Behavioral Sciences, University of Washington.

instruments, and using measures of F0 synchrony occurring at both the session-level and minute-level. We considered differences in culture and language, patients' characteristics, and setting as explanations for this failure to replicate. Further replication testing and new developments regarding measurement methods and modeling are needed.

Keywords

Empathy; Behavioral coding; Speech signal processing; Mean fundamental frequency of the voice; Replication study

Introduction

Effective communication between clinicians and patients has a range of significant benefits, including alleviating pain, increasing patient adherence to treatment, and improving satisfaction with care (Back, Arnold, & Tulskey, 2009). Empathy, which has been defined as an active interest in and effort to understand another person's internal perspective (e.g. Miller & Rollnick, 2013), is a particularly effective communication skill in the psychotherapeutic process (e.g. Norcross, 2011) and in general medicine (Kim, Kaplowitz, & Johnston, 2004). In general practice, a recent review showed significant effects of physician empathy on patient's satisfaction, patients' anxiety and distress, more accurate diagnosis, and overall better clinical outcomes (Derksen, Bensing, & Lagro-Janssen, 2013). In addition behavioral treatment, a recent review showed that high-empathy counselors had higher success rates and stronger therapeutic alliance, while low-empathy and confrontational counseling was associated with higher drop-out, greater relapse rates, and weaker therapeutic alliance (Moyers & Miller, 2013).

Empathy may be a core active ingredient in clinical encounters; however, its measurement is challenged by several factors. First, different methods may be required to distinguish empathy as a personality trait (i.e., dispositional empathy) versus empathy as a communication behavior that varies across encounters (i.e., situational empathy; e.g. Hallgren et al., 2018). Whereas dispositional empathy is usually evaluated by self-report questionnaires that suffer from poor reliability (Hemmerdinger, Stoddart, & Lilford, 2007), situational empathy is often measured via behavioral coding (the current gold standard) and differences in situational empathy are often substantially larger than differences in dispositional empathy (Hallgren et al., 2018; Imel et al., 2014a). To measure situational empathy, trained coders typically listen to a particular session and attribute overall ratings of empathy to clinician and/or patient behaviors within the session. However, this type of coding is particularly cost-, time-, and labor-intensive and subject to important reliability challenges (Hemmerdinger et al., 2007; Imel et al., 2014b). Behavioral coding requires training a team of raters, establishing inter-rater reliability among the raters, and then performing the time-consuming task of coding. This procedure importantly limits the evaluation of provider fidelity to behavioral interventions, as well as research on interventions processes. In a meta-analysis examining the association of therapist adherence/competence and treatment outcomes, the average sample size across 36 studies was 40 clients per study, which certainly explained part of the noticeable variability observed across

studies (Webb, Derubeis, & Barber, 2010). In a recent meta-analysis of Motivational Interviewing process (Magill et al., 2018), the average sample size was 75 clients. It has thus been advocated that the current method of evaluation using human raters will never scale up to meaningful clinical use and that a technological solution is needed (Atkins, Steyvers, Imel, & Smyth, 2014). Achieving inexpensive, rapid, and reliable evaluations of clinician empathy might help clinical advances, such as providing instant feedback on empathy ratings to clinicians, trainees, or students, as well as research advances on treatment fidelity and therapy processes with a potential to enhance its effect.

Recently, an alternative measurement of situational empathy has been proposed: empathy measured as synchrony in vocally encoded arousal (Imel et al., 2014b; Xiao, Georgiou, Imel, Atkins, & Narayanan, 2013). This approach builds on the conceptualization of empathy as a general process involving mirroring (Preston & de Waal, 2002), imitation, or synchrony wherein humans learn how others feel by experiencing a representation of a similar state (Iacoboni, 2009). It also relies on communication accommodation theory which showed the importance of behavioral synchrony in human dyads (e.g. Giles & Ogay, 2007). Recent research by Imel et al. (2014b) extended this work by modeling empathy in relation to synchrony in vocally encoded emotional arousal in patient and clinician speech during Motivational Interviewing (MI) sessions. Specifically, they used computerized extractions of mean fundamental frequencies (“mean F0”), which represent the lowest harmonic produced during speech and which is a measurable proxy of voice pitch. Mean f0 is hypothesized to encode emotional arousal expressed in the voice (e.g., high = excited, angry, or nervous; low = bored, calm, or content) (Juslin & Scherer, 2005). Their sample included 89 clinicians working in substance abuse treatment facilities in the United States who received up to 15 hours of MI training, and their study used the 89 twenty-minute evaluative MI sessions that the clinicians conducted with one of three standardized patients (i.e. actors portraying generic patients). Imel et al. (2014b) evaluated the degree of synchrony between patients and clinicians by evaluating correlations between clinicians’ and patients’ mean F0 values within a Bayesian multilevel modeling framework. Two types of correlation were investigated: the degree of *session-level*/F0 synchrony (i.e., correlation between clinician and patient mean F0 across sessions) and the degree of *minute-level*/F0 synchrony (i.e., correlation between clinician and patient mean F0 across minutes within individual sessions). They found session-level synchrony was higher ($r = .80$) in sessions rated as high-empathy using the Motivational Interviewing Treatment Integrity Scale (Moyers, Martin, Manuel, Hendrickson, & Miller, 2005) than in low-empathy sessions ($r = .36$) and that minute-level synchrony was present to a modest extent in both high-empathy ($r = .31$) and low-empathy sessions ($r = .20$). These results showed preliminary support for an association of vocal synchrony with empathy ratings; while important and consistent with theory, it is a single sample and relied on the use of standardized patients. Moreover, the generalizability of the work was restricted by the use of therapists who had recently completed MI training and the testing of this relationship only within an American English-speaking sample (Imel et al., 2014b). The findings of Imel et al. could hold promise for psychotherapy research and evaluation by providing a means to automatically evaluate empathy levels, for example through technology that detects and evaluates patients’ and clinicians’ vocal F0 synchrony. These efforts would be especially supported through research demonstrating that these findings

replicated in independent samples and generalized to multiple clinical contexts (e.g., clinical sessions focusing specifically on substance abuse or general medical practice) and linguistic/cultural groups (e.g., Swiss French-speaking individuals).

In the present project, we sought to replicate the study by Imel et al. (2014b) within two large samples ($N = 208$ and $N = 204$) of clinical interactions with real patients from two independent studies of Swiss French-speaking samples. Situational empathy was measured via traditional, labor-intensive behavioral coding, and following Imel et al. (2014b), it was hypothesized that session- and minute-level mean F0 synchrony between clinicians and patients would be significantly higher in high-empathy sessions compared to low-empathy sessions.

Methods

Data source

The present project was a secondary analysis of the audio files of clinical interviews collected for 2 unrelated projects (see Figure 1). Study 1 (Gaume et al., 2014) was a randomized controlled trial of brief motivational interventions (BMI) among hazardous drinkers included within the Army Recruitment Center of Lausanne, Switzerland. Switzerland has a mandatory two-day army conscription process for all males at age 19 to determine eligibility for service in the Swiss military. Women are allowed to join the military service on a voluntary basis but were not included because of their scarcity and resulting non-representativeness. Hazardous drinkers were randomly allocated to the BMI condition ($n = 217$) or to a control condition receiving no intervention ($n = 224$). Within the BMI condition, 208 interventions (95.8%) were completely audio-recorded and used in the present study, while the remaining 9 had technical problems. The intervention was a 20 to 30-minute brief adaptation of MI which combines the person-centered, empathic counseling approach originally developed by Carl Rogers with a behavioral focus on resolving ambivalence in the direction of change (Miller & Rollnick, 2013). Therapists were 9 physicians and 9 psychologists, selected to provide a large range of clinical and MI experience, and gender parity. Study procedures were approved by the Ethics Committee for Clinical Research of Lausanne University Medical School.

Study 2 (Carrard, Schmid Mast, Jaunin-Stalder, Junod Perron, & Sommer, 2018) was an observational study of which the general aim was to investigate physician communication in general practice. The authors observed whether physicians adapting their communication behavior to their patients' preferences leads to greater benefit in patient satisfaction and trust in the physician and how interpersonal accuracy skills would enable this adaptation. The sample consisted of 61 general practitioners (27 female and 34 male) working in private practice in the Cantons of Vaud, Geneva, and Neuchâtel, Switzerland. Each participating physician was videotaped during 4 usual general practice consultations with the first two female patients and the first two male patients agreeing to be included in the study, resulting in 244 videotaped consultations (22.9 minutes long on average, $SD = 10.2$). The procedure of this data collection was approved by the ethic committees of the Cantons of Geneva and Vaud (the latter also serving Canton of Neuchâtel). Due to time and financial constraints that limited the amount our behavioral coding of empathy, we randomly selected 204

consultations, stratified by patient's and clinician's gender, to code for the following analyses.

Clinician and patient speech segmentation and vocal fundamental frequency extraction

Because clinicians and patients were not recorded with separate microphones, audio files had to be manually segmented to identify respective speech (i.e., patient, clinician, or neither/indistinguishable). Speaker segmentation was performed by research assistants trained in using the same procedures, i.e. segmentation and categorization using either Dartfish Team Pro 4.0 software (Dartfish, 2006) or the 'Annotation' module in Praat version 5.3.82 (Boersma & Weenink, 2014). The present analysis only included mean F0 identified within patient and clinician speech segments.

Mean F0 was measured every 0.25 seconds using Praat speech signal processing software (Boersma & Weenink, 2014). In Study 1, sessions were originally audio recorded using a portable MiniDisc recorder; we imported recordings as mp3 files (stereo mpga), with a 44 kHz sampling rate and a 128 kbit/s bit rate. In Study 2, sessions were originally video-recorded using a digital camcorder; we converted recordings to mp3 files (stereo mpga), with a 48 kHz sampling rate and a 192 kbit/s bit rate. We then used a Praat command script automating the extraction of all sound files present in a folder. This script would 1) read the sound file, 2) compute F0 values using the function "To Pitch", 3) generate 0.25 seconds segments, 4) compute mean F0 within each segment, 5) append all information in output text files (1 text file per recorded session) with each line being a 0.25 seconds segment, including the beginning and end time, as well as the mean F0 value in Hz. We used a pitch range setting of 75 Hz to 300 Hz, corresponding to the F0 typically range between 80 and 175 Hz among men, and 160 and 270 Hz among women in adults and post-pubertal adolescents (Fouquet, Pisanski, Mathevon, & Reby, 2016; Kreiman & Sidtis, 2011).

Behavioral coding of empathy

Behavioral coding is the current gold standard to measure empathy during clinical encounters. In the present study, we used two validated empathy scales: the Empathy global rating from the Motivational Interviewing Skill Code (MISC) version 2.1 (Miller, Moyers, Ernst, & Amrhein, 2008), and the Therapist Empathy Scale (TES, Decker, Nich, Carroll, & Martino, 2014). Coders listened to the entire session and then assessed the two scales intended to capture the coders' overall impression of counselor performance during the interview.

The *Empathy global rating from the MISC* is intended to capture the extent to which the counselor understands and/or makes an effort to accurately understand the client's perspective (Miller et al., 2008). It is the exact same scale as that of the Motivational Interviewing Treatment Integrity Scale (Moyers et al., 2005) which was used in Imel et al. (2014b). It is rated via a single item using a 7-point Likert scale. A rating should be made starting at 4, and moving toward either the high (7) or low (1) end of the scale based on anchors of High empathy (e.g. clinicians show active interest in making sure they understand what the client is saying, including the client's perceptions, situation, meaning, and feelings; they accurately follow or perceive a client's complex story) and Low empathy (e.g.

clinicians show little interest in the client's own perspective and experiences; there is little effort to gain a deeper understanding of complex events and emotions; clinicians may probe for factual information or to pursue an agenda).

The *TES* is a 9-item scale that assesses affective, cognitive, attitudinal, and attunement aspects of therapist empathy (Decker et al., 2014). The 9 items assess the following: 1. Concern for the patient; 2. Expressiveness; 3. Resonate or capture patient feelings; 4. Warmth; 5. Attuned to the patient's inner world; 6. Understanding cognitive framework; 7. Understanding feelings/inner experience; 8. Acceptance of patient's feelings and inner experiences; and 9. Responsiveness. It is rated on a 7-point Likert scale with lower values indicating the absence of the targeted empathy component and higher values indicating frequent or extensive demonstration of it (1 = not at all, to 7 = extensively). A summary score (mean of the 9 items) is proposed to capture empathy globally.

Three psychology master-level students and one organizational behavior doctoral student were trained to use the scales and then independently coded encounters from both studies while being blinded to any other data. Coding discrepancies and challenges were addressed in weekly joint trainer-coder meetings, which lasted throughout the entire coding period. Each coder received about 40 total hours of training and supervision. A random subsample of 79 sessions from both studies (19.1%) was double-coded to assess inter-rater reliability. Weighted kappa with quadratic weights was calculated to address inter-rater for ordinal global ratings (Cohen, 1968). Kappa was 0.48 for the MISC empathy rating and 0.49 for the TES summary score, which correspond to "fair" reliability according to the categorization by Cicchetti (1994). While only moderate, this level of inter-rater agreement is common when assessing global ratings at the session level (Moyers et al., 2005), but somewhat lower than in the study by Imel et al. (2014b) where ICC was 0.60 (ICC being equivalent to quadratic weighted kappa, see e.g. Hallgren, 2012).

Data analysis

Statistical analyses aimed to replicate Imel et al. (2014b) by evaluating the degree of minute- and session-level F0 synchrony within high- and low-empathy MI sessions. We first conducted extensive descriptive statistical analysis and graphical exploration of the data to examine distributions, variability across persons and sessions (data not presented in the present report), and normality of residuals after controlling for speaker and gender. Imel et al. (2014b) focused on a categorical comparison of high vs. low empathy sessions, which we followed. We thus examined the distributions of MISC and TES empathy ratings to identify subsets of sessions designated as having high or low empathy by each behavioral coding measure, and corresponding MISC and TES cutoffs were selected by identifying whole-number cutoff values that resulted in approximately one quarter of sessions being in each of the high- and low-empathy categories. We used different MISC cutoffs for high- and low-empathy sessions within each sample, and these differed from those used by Imel et al., in effort to map relatively higher- and lower-empathy cutoffs onto the distributions of our data – for example, there were only 4 sessions meeting Imel et al.'s criterion for "high empathy" (MISC = 6) in Study 2. Imel et al. did not utilize the TES and thus we generated cutoff values based on the observed distribution of scores. (There was some variability in the sizes

of each subsample due to the use of interval, rather than continuous, rating systems, which limited the granularity with which sessions could be subdivided.) Distributions of empathy ratings and selected cutoff values are presented in Table 1. Notably, empathy ratings in Study 1 were distributed around the midpoint of the empathy rating scales, whereas empathy ratings in Study 2 were generally lower and distributed closer to the lower end of the rating scales. MISC and TES cutoff scores for identifying sessions with low and high empathy within each sample were adjusted for each study accordingly. Empathy ratings were also significantly lower for male providers in Study 1 (TES gender difference = -0.43 , $SE=0.11$, $t=-4.12$, $p<.001$; MISC Empathy gender difference = -0.73 , $SE=0.15$, $t=-4.74$, $p<.001$), but not in Study 2 (TES gender difference = -0.11 , $SE=0.11$, $t=-1.04$, $p=.30$; MISC Empathy gender difference = 0.01 , $SE=0.15$, $t=0.05$, $p=.96$; clinician gender was always included as a covariate in the statistical models, see below). In Study 1, the MISC identified 42 low- and 32 high-empathy sessions, and the TES identified 32 low- and 50 high-empathy sessions. In Study 2, the MISC identified 84 low- and 48 high-empathy sessions, and the TES identified 25 and 36.

Analyses of data from both studies were then modeled after those by Imel et al. (2014b) and were carried out separately for both parent studies and for each of the empathy measures separately (i.e., empathy from MISC, overall empathy in TES). This resulted in 4 statistical models in total.

Synchrony between clinician and patient F0 was analyzed using multivariate multilevel models (Baldwin, Imel, Braithwaite, & Atkins, 2014; MacCallum, Kim, & Malarkey, 1997) to account for the repeated measurement of mean F0 sampled every 0.25 seconds that were nested within minutes and within sessions. We used the following model, derived from Imel et al. (2014b):

$$Y_{ijk} = b_0Clinician_{ijk} + b_1Patient_{ijk} + b_2ClinicianMale_k + b_3PatientMale_k + b_4Empathy_k + r_{Cj} + r_{Pj} + u_{Ck} + u_{Pk} + e_{ijk}$$

In the formula above, Y_{ijk} is the measured 0.25-second mean F0 value for observation i within minute j within session k . To fit a multivariate multilevel model of clinician and patient F0 as simultaneous outcomes, the overall intercept is removed, and role-specific intercepts were included. $Clinician_{ijk}$ and $Patient_{ijk}$ were dummy variables indicating who was speaking within the observation period, $ClinicianMale_k$ and $PatientMale_k$ were dummy variables controlling for mean differences in F0 based on clinician and patient gender (0=female, 1=male; the $PatientMale_k$ term was excluded in Study 1 which contained only male patients), and $Empathy_k$ was a dummy variable indicating whether a session was identified as low ($Empathy_k = 0$) or high empathy ($Empathy_k = 1$). The random effects r_{Cjk} and r_{Pjk} modeled between-minute variability in mean F0 for clinicians and patients, respectively, while the u_{Ck} and u_{Pk} modeled between-session variability in mean F0 for clinicians and patients, respectively and e_{ijk} modeled residual error. Random effects were assumed to follow a multivariate normal distribution, with a pre-defined structured covariance matrix:

$$\begin{bmatrix} u_{Ck} \\ u_{Pk} \\ r_{Cjk} \\ r_{Pjk} \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_C}^2 & & & \\ \sigma_{u_P u_C} & \sigma_{u_P}^2 & & \\ & & \sigma_{r_C}^2 & \\ & & \sigma_{r_P r_C} & \sigma_{r_P}^2 \end{bmatrix} \right),$$

where $\sigma_{r_C}^2$, $\sigma_{u_C}^2$, $\sigma_{r_P}^2$, and $\sigma_{u_P}^2$ are the variances of mean F0 across minutes and sessions for clinicians and patients, respectively, and $\sigma_{r_P r_C}$ and $\sigma_{u_P u_C}$ are the covariances between patient and clinician mean F0 across minutes and sessions, respectively. These covariances were transformed into correlations to facilitate interpretation and reflected the degree of similarity between patient and clinician F0 across minutes and sessions on a scale that was analogous to a Pearson correlation coefficient (i.e., 0 indicating almost no synchrony, 1 indicating perfect synchrony). Random effects were stratified based on level of empathy rating (high or low), yielding variance and covariance terms from the model matrix above for each level of empathy and facilitating a comparison of the degree of synchrony for high-versus low-empathy sessions. Residual errors were assumed to be independent and normally distributed with a mean of zero.

Models were fit within a Bayesian approach using the MCMCglmm package in R (Hadfield, 2010; R Core Team, 2016) with an iterative Markov chain Monte Carlo (MCMC) fitting procedure. Minimally informative priors were used for fixed effects (multivariate normal with large variances) and for variance components (inverse Wishart; Gelman & Hill, 2007). MCMC simulations were run with 100,000 iterations, a burn-in of 5,000 iterations, and a thinning interval of 20. Posterior convergence was assessed with autocorrelation and trace plots, and the modes and 95% highest posterior density intervals (HPDs, analogous to a confidence interval; Gelman & Hill, 2007) were used to estimate fixed and random effect values.

Results

Study 1

Across the 74 sessions in Study 1 with high or low empathy as identified by the MISC, there were 190,393 mean F0 observations ($M = 2572$ observations per session) over a total of 1,374 minutes ($M = 18.6$ minutes per session). Similarly, for the 82 sessions in Study 1 with high or low empathy as identified by the TES, there were 216,055 mean F0 observations ($M = 2634$ observations per session) over a total of 1,562 minutes (mean = 19.1 minutes per session). Scatterplot figures illustrating the distributions of F0 values at the minute-level across the two levels of empathy are available in the Supplemental online material.

Fixed effect coefficients (b values in the equation above), which indicate average mean F0 values for patients, providers, and differences in mean F0 based on therapist gender and empathy level are presented in Table 2. Mean F0 was lower for male clinicians compared to females, which is a function of vocal tract anatomical differences between men and women

(i.e., men typically having lower voices). Controlling for speaker and gender, mean F0 was slightly higher in high-empathy sessions.

Correlations of random effects, shown in Figure 2, indicated only small, and often non-significant degrees of synchrony (i.e. correlations) between patient and clinician F0, which is inconsistent with the results of Imel et al. (2014b). At the minute-level, although three of the four correlations were significantly greater than zero (i.e., 95% HPD intervals above zero for MISC Low-empathy sessions and for TES High- and Low-empathy sessions), the magnitudes of these effects were small (range of r s: 0.09 to 0.19). At the session-level, two of the four correlations were significantly greater than zero (MISC Low-empathy sessions and TES High-empathy sessions, range of r s: 0.17 to 0.31). Over all the levels of analysis (minutes and sessions), the differences in low- and high-empathy correlation posterior distributions were not significantly different, indicating a lack of difference in synchrony associated with high versus low empathy (difference of posterior correlation estimates for MISC = 0.12 [CI: -.29, .62] at session-level and -0.04 [CI: -.19, .14] at minute-level; for TES: difference = -0.11 [CI: -.60, .25] at session-level and 0.11 [CI: -.08, .28] at minute-level). In one exception, the correlation coefficient for minute-level synchrony in low-empathy sessions measured by the TES was greater than the HPD interval for high-empathy sessions measured by the TES, which was the opposite pattern of what we hypothesized. In summary, there was no evidence for systematically higher minute- or session-level synchrony in high-empathy sessions compared to low-empathy sessions for either measure.

Study 2

Across the 132 sessions in Study 2 with high or low empathy as identified by the MISC, there were 412,777 mean F0 observations ($M = 3127$ observations per session) over a total of 2,729 minutes ($M = 20.7$ minutes per session). Similarly, for the 61 sessions in Study 2 with high or low empathy as identified by the TES, there were 199,267 mean F0 observations ($M = 3266$ per session) over a total of 1,324 minutes (mean = 21.7 per session). Fixed effect coefficients are presented in Table 2. Male clinicians and patients had lower mean F0 compared to values compared females, but there were no significant differences in mean F0 values between high and low-empathy sessions.

Similar to Study 1, correlations of random effects (Figure 2) in Study 2 were relatively small and inconsistent with those found by Imel et al. (2014b). All of the minute-level correlations were significantly greater than zero with small effect sizes (range of r s: 0.13 to 0.28). At the session-level, one of the four correlations was significantly greater than zero (MISC Low-empathy sessions; range of r s: -0.03 to 0.28). At both levels of analysis (minutes and sessions), correlation coefficients for low-empathy sessions were almost always contained within the 95% HPD intervals of correlation coefficients for high-empathy sessions and vice versa, indicating again a lack of difference in synchrony associated with high versus low empathy. In one exception, the correlation coefficient for minute-level synchrony in low-empathy sessions measured by the TES was lower than the HPD interval for high-empathy sessions measured by the TES, which was consistent with the pattern we hypothesized. Differences of posterior correlation estimates were consistent with these findings (for MISC: difference = 0.23 [CI: -.17, .55] at session-level and -0.08 [CI: -.16, .04] at minute-level;

for TES: difference = 0.01 [CI: -.54, .51] at session-level and -0.13 [CI: -0.30, .00] at minute-level). However, in summary, there was no evidence for systematically higher minute- or session-level synchrony in high-empathy sessions compared to low-empathy sessions for either measure.

Discussion

In the present analysis, we aimed to replicate the study by Imel et al. (2014b) by evaluating potential associations of vocal synchrony with empathy ratings using two large samples ($N = 208$ and $N = 204$) of clinical interactions with real patients. Findings showed no support for our hypothesis that mean F0 synchrony between clinicians and patients would be higher in high-empathy sessions based on empathy ratings measured using traditional, labor intensive, behavioral coding. This lack of replication was consistent for both clinical samples (brief MI with young men and general practice consultation), both behavioral coding instruments (Empathy global rating from the MISC and TES scale), and both levels of analysis (session-level and minute-level synchrony).

There are several potential reasons for this failure to replicate. A first important distinction between our study and the one by Imel et al. (2014b) is related to language and culture. Our study included sessions held in French-speaking Switzerland, whereas Imel et al. included sessions held in English in the Pacific Northwest region of the United States. Paralinguistic features might differ importantly across languages. First, there might be cultural differences, for example, with US Americans talking more loudly and Europeans more quietly, or US Americans seeing silence as embarrassing more than other cultures (Dahl, 2002). Second, there is a growing literature showing that languages can differ in their F0 profile (Keating & Kuo, 2012). Differences in acoustic measures across different languages have been observed at several levels: differences in mean F0, range of F0, variability of F0, and voice intensity (Dolson, 1994; Hanley & Snidecor, 1967; Mennen, Schaeffler, & Docherty, 2012). Some authors also found cross-language variation in acoustic features, particularly F0 range, between French and American English speakers (Pépiot, 2015) and that momentary affect corresponds with different F0 values between the two languages (Boula de Mareüil, Célérier, & Toen, 2002). Other authors showed that English differs from French in prosody (Tremblay, Broersma, Coughlin, & Choi, 2016). Stressed syllables in accented English words have been shown to be related to F0 rise, increased duration, and greater intensity (e.g. Beckman, 1986; Lieberman, 1960).

Also, different “clinical culture” might have been captured. Analyses of F0 evolution over sessions (data not provided) showed relatively flat F0 trajectories in the present study, which might indicate relatively reflective, monotonous speech. On the other hand, Imel et al. (2014b) might have captured more expressive clinicians, which resulted in more variation in mean F0 and thus higher likelihood of finding synchrony. Recent research showed cultural differences in the use of acoustic cues to convey emotion through the human voice (Laukka et al., 2016; Laukka, Neiberg, & Elfenbein, 2014). It is possible that empathy is related to synchrony in vocally-encoded arousal in some languages / cultures (e.g., American English) but not others (e.g., Swiss French). Further research is thus warranted at these different levels.

A second distinction refers to the types of patients used in clinical recordings. In Imel et al. (2014b), clinicians interacted with only one of three standardized patients (i.e. actors portraying generic patients). This almost certainly reduced patient heterogeneity within the sample (i.e., due to the use of 3 standardized patients), which is the intention in standardized patient designs. It is well documented in the literature on therapist effects on patient outcomes (Boswell et al., 2013; Denhag, Gibbons, Barber, Gallop, & Crits-Christoph, 2012; Imel, Baer, Martino, Ball, & Carroll, 2011) that the majority of the variance in patient outcomes is attributable to patient, often 90% or more. Standardized patient designs are used to reduce this variance and focus on clinician variance and some research has suggested that clinician performance with standardized and real patients are consistent (Imel et al., 2014a; Schwartz, Rothpletz-Puglia, Denmark, & Byham-Gray, 2015). Thus, it may be that standardized patients amplify a hypothesized F0 synchrony effect; however, if this is the case, then the present results suggest this effect is very small, requiring large samples to reliably detect.

Finally, there are important setting differences the studies examined here and in Imel et al. (2014b). In the study by Imel et al. (2014b), clinicians were therapists from community substance abuse treatment facilities who were recently exposed to MI training and might have been particularly attentive to empathic reflection of patient emotions. In the present samples, clinicians were alcohol counselors and GPs in private practice. However, counselors providing alcohol brief MI in Study 1 were all trained in MI and their average empathy score compared to this in the study from which data for Imel et al. were drawn (MISC empathy score: mean=4.4, sd=1.2 in Study 1; mean=4.2, sd=1.3 for Imel et al. data, Imel et al., 2014a). Since results from counselors in Study 1 were fully consistent with results from GPs in Study 2, which were not particularly trained to being empathic and consistently showed lower scores on empathy behavioral ratings, the failure to replicate is unlikely to be related to therapists' MI proficiency or empathy skills. Nevertheless, other unmeasured therapist differences might have been at play and could be further investigated.

An additional explanation might be a lack of vocal synchrony overall, which might be attributed to the short acquaintance time between the therapist and the patients. There is indeed evidence that synchrony may develop during the time of acquaintance (Levitan & Hirschberg, 2011). Session duration was similar on average across the three samples (i.e. 20 minutes in Imel et al., 20–30 minutes in Study 1, 23 minutes in Study 2), but the time of acquaintance might have been influenced by the study setting, the type of patient, and/or language and clinical cultures.

Implications for Empathy Research and Evaluation

The present study focused on a single assessment (i.e., vocally encoded emotional arousal) to examine synchrony as a key component of empathy. Other research using alternative physiological, behavioral, and vocal indices has suggested that synchrony is an important aspect of empathy or related constructs, including linguistic style matching (Lord, Sheng, Imel, Baer, & Atkins, 2015), skin conductance (Marci, Ham, Moran, & Orr, 2007), and movement and gestures (Ramseyer & Tschacher, 2011). These results would suggest that synchrony as an indicator of empathy or relationship quality in psychotherapy is a

multivariate process, and it is possible that the current results suggest that there may be culture-specific representations of it. In addition, one of the motivations for the current work is to understand features of empathy that can be quantified from therapy recordings to support automated evaluation of empathy, for example, with natural language processing methods (see e.g. Xiao, Imel, Georgiou, Atkins, & Narayanan, 2015), which in turn could facilitate faster and larger-scale evaluation of provider training and mechanisms of change in MI (Imel, Caperton, Tanana, & Atkins, 2017). Results from the present study may dampen enthusiasm for using vocal F0 synchrony as a means to evaluate empathy, at least within Swiss French speaking clinical samples. Continuing to explore and understand the behavioral, emotional, and vocal features of empathy will assist work toward automated methods of evaluating empathy, which can support research as well as clinical evaluation.

Furthermore, we hypothesized that F0 synchrony would be a measure of empathy, as suggested and tested in Imel et al (2014b). However, the external validity of this assumption has not been further tested. F0 synchrony could reflect other theoretical constructs, such as therapeutic alliance, emotional responsiveness, or interpersonal adjustment. In particular, therapeutic alliance might be a relevant candidate as it has been proposed that alliance is related to interpersonal synchrony in psychotherapy (Koole & Tschacher, 2016). Higher levels of bodily synchrony have been shown to be correlated with more positive ratings of the therapeutic alliance (e.g. Ramseyer & Tschacher, 2011). We found one study examining whether synchrony in vocal mean F0 was related to therapeutic alliance (Reich, Berman, Dale, & Levitt, 2014). Results indicated that F0 synchrony did occur in the sessions but higher levels of synchrony were related to poorer therapeutic alliance. We believe that further analyses should be conducted to explore these research avenues.

On the other hand, it could also be that acoustic features other than mean F0 are related to empathy. For example, speech rate, percentage of silences, intensity, or jitter might be candidates for serving as synchronization measures and related to empathy. These measures could also be tested in future studies as potential indicators for automatically detecting empathy.

Limitations, Future Directions, and Conclusion

In addition to these distinctions, a limitation of the present study should be borne in mind. Interrater reliability in empathy behavioral coding was only moderate. While this level of interrater agreement is common when assessing global ratings at the session level (see e.g. Moyers et al., 2005), it was higher in Imel et al. (2014b) where intra-class correlation was 0.60 indicating “good” reliability.

Taking into account limitations and specificities of the present study and the one by Imel et al. (2014b), clearly call for further research. This research should include replication with different type of patients, different clinical settings, and different languages and clinical cultures. Also, new developments regarding measurement methods and modeling might be considered. Nevertheless, we believe that understanding and improving communication between clinicians and patients is crucial and effective measures of these communication features would thus be essential.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was funded by the Research Fund of the Department of Community Health and Medicine of Lausanne University Hospital, Lausanne, Switzerland. Dr. Atkins is a co-founder with equity stake in a technology company, Lyssn.io, focused on tools to support training, supervision, and quality assurance of psychotherapy and counseling.

References

- Atkins DC, Steyvers M, Imel ZE, & Smyth P (2014). Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9, 49. doi:10.1186/1748-5908-9-49 [PubMed: 24758152]
- Back AL, Arnold RM, & Tulskey JA (2009). *Mastering Communication with Seriously Ill Patients: Balancing Honesty with Empathy and Hope*. New York: Cambridge University Press.
- Baldwin SA, Imel ZE, Braithwaite SR, & Atkins DC (2014). Analyzing multiple outcomes in clinical research using multivariate multilevel models. *Journal of Consulting and Clinical Psychology*, 82(5), 920–930. doi:10.1037/a0035628 [PubMed: 24491071]
- Beckman ME (1986). *Stress and Non-Stress Accent*. Dordrecht: Foris.
- Boersma P, & Weenink D (2014). *Praat: Doing phonetics by computer* [Computer program]. Version 5.3.82. Amsterdam, NL: University of Amsterdam.
- Boswell JF, Gallagher MW, Sauer-Zavala SE, Bullis J, Gorman JM, Shear MK, ... Barlow DH (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. doi:10.1037/a0031437 [PubMed: 23339537]
- Boula de Mareüil P, Célériér P, & Toen J (2002). Generation of Emotions by a Morphing Technique in English, French and Spanish. *Proceedings of Speech Prosody*, 187–190. Retrieved from https://www.isca-speech.org/archive/sp2002/papers/sp02_187.pdf
- Carrard V, Schmid Mast M, Jaunin-Stalder N, Junod Perron N, & Sommer J (2018). Patient-Centeredness as Physician Behavioral Adaptability to Patient Preferences. *Health Communication*, 33(5), 593–600. doi:10.1080/10410236.2017.1286282 [PubMed: 28278605]
- Cicchetti DV (1994). Guidelines, criteria and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6, 284–290. doi:10.1037/1040-3590.6.4.284
- Cohen J (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220. doi:10.1037/h0026256 [PubMed: 19673146]
- Dahl S (2002). *Communication and Culture Transformation: Cultural Diversity, Globalization and Cultural Convergence*. London, UK: ECE Europacom Retrieved from <http://dahl.at/docs/capstone.pdf>.
- Dartfish. (2006). *Dartfish Software TeamPro 4.0* [Computer program]. Fribourg, Switzerland: Dartfish Ltd.
- Decker SE, Nich C, Carroll KM, & Martino S (2014). Development of the Therapist Empathy Scale. *Behavioural and Cognitive Psychotherapy*, 42(3), 339–354. doi:10.1017/S1352465813000039 [PubMed: 23481373]
- Dennhag I, Gibbons MB, Barber JP, Gallop R, & Crits-Christoph P (2012). How many treatment sessions and patients are needed to create a stable score of adherence and competence in the treatment of cocaine dependence? *Psychotherapy Research*, 22(4), 475–488. doi:10.1080/10503307.2012.674790 [PubMed: 22449079]
- Derksen F, Bensing J, & Lagro-Janssen A (2013). Effectiveness of empathy in general practice: a systematic review. *British Journal of General Practice*, 63(606), e76–84. doi:10.3399/bjgp13X660814 [PubMed: 23336477]

- Dolson M (1994). The pitch of speech as a function of linguistic community. *Music Perception*, 11, 321–331. Retrieved from <https://www.jstor.org/stable/40285626>
- Fouquet M, Pisanski K, Mathevon N, & Reby D (2016). Seven and up: individual differences in male voice fundamental frequency emerge before puberty and remain stable throughout adulthood. *Royal Society Open Science*, 3(10), 160395. doi:10.1098/rsos.160395 [PubMed: 27853555]
- Gaume J, Magill M, Longabaugh R, Bertholet N, Gmel G, & Daeppen JB (2014). Influence of counselor characteristics and behaviors on the efficacy of a brief motivational intervention for heavy drinking in young men--a randomized controlled trial. *Alcoholism, Clinical and Experimental Research*, 38(7), 2138–2147. doi:10.1111/acer.12469
- Gelman A, & Hill J (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press.
- Giles H, & Ogay T (2007). Communication accommodation theory In Whaley BB & Samter W (Eds.), *Explaining communication: Contemporary theories and exemplars* (pp. 293–310). Mahwah, NJ: Erlbaum.
- Hadfield JD (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33, 1–22. doi:10.18637/jss.v033.i02 [PubMed: 20808728]
- Hallgren KA (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/> [PubMed: 22833776]
- Hallgren KA, Dembe A, Pace BT, Imel ZE, Lee CM, & Atkins DC (2018). Variability in motivational interviewing adherence across sessions, providers, sites, and research contexts. *Journal of Substance Abuse Treatment*, 84, 30–41. doi:10.1016/j.jsat.2017.10.011 [PubMed: 29195591]
- Hanley TD, & Snidecor JC (1967). Some Acoustic Similarities among Languages. *Phonetica* 17, 141–148 Retrieved from <https://www.karger.com/Article/Pdf/258584>
- Hemmerdinger JM, Stoddart SD, & Lilford RJ (2007). A systematic review of tests of empathy in medicine. *BMC Medical Education*, 7, 24. doi:10.1186/1472-6920-7-24 [PubMed: 17651477]
- Iacoboni M (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology*, 60, 653–670. doi:10.1146/annurev.psych.60.110707.163604
- Imel ZE, Baer JS, Martino S, Ball SA, & Carroll KM (2011). Mutual influence in therapist competence and adherence to motivational enhancement therapy. *Drug and Alcohol Dependence*, 115(3), 229–236. doi:10.1016/j.drugalcdep.2010.11.010 [PubMed: 21227601]
- Imel ZE, Baldwin SA, Baer JS, Hartzler B, Dunn C, Rosengren DB, & Atkins DC (2014a). Evaluating therapist adherence in motivational interviewing by comparing performance with standardized and real patients. *Journal of Consulting and Clinical Psychology*, 82(3), 472–481. doi:10.1037/a0036158 [PubMed: 24588405]
- Imel ZE, Barco JS, Brown HJ, Baucom BR, Baer JS, Kircher JC, & Atkins DC (2014b). The association of therapist empathy and synchrony in vocally encoded arousal. *Journal of Counseling Psychology*, 61(1), 146–153. doi:10.1037/a0034943 [PubMed: 24274679]
- Imel ZE, Caperton DD, Tanana M, & Atkins DC (2017). Technology-enhanced human interaction in psychotherapy. *Journal of Counseling Psychology*, 64(4), 385–393. doi:10.1037/cou0000213 [PubMed: 28318277]
- Juslin PN, & Scherer K (2005). Vocal expression of affect In Harrigan J, Rosenthal R & Scherer KR (Eds.), *The new Handbook of Methods in Nonverbal Behavioral Research* (pp. 65–136). New York: Oxford University Press.
- Keating P, & Kuo G (2012). Comparison of speaking fundamental frequency in English and Mandarin. *Journal of the Acoustical Society of America*, 132(2), 1050–1060. doi:10.1121/1.4730893 [PubMed: 22894225]
- Kim SS, Kaplowitz S, & Johnston MV (2004). The effects of physician empathy on patient satisfaction and compliance. *Evaluation and the Health Professions*, 27(3), 237–251. doi:10.1177/0163278704267037 [PubMed: 15312283]
- Koole SL, & Tschacher W (2016). Synchrony in Psychotherapy: A Review and an Integrative Framework for the Therapeutic Alliance. *Frontiers in Psychology*, 7, 862. doi:10.3389/fpsyg.2016.00862 [PubMed: 27378968]

- Kreiman J, & Sidtis D (2011). Foundations of voice studies: an interdisciplinary approach to voice production and perception. New York, NY: Wiley-Blackwell.
- Laukka P, Elfenbein HA, Thingujam NS, Rockstuhl T, Iraki FK, Chui W, & Althoff J (2016). The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *Journal of Personality and Social Psychology*, 111(5), 686–705. doi:10.1037/pspi0000066 [PubMed: 27537275]
- Laukka P, Neiberg D, & Elfenbein HA (2014). Evidence for cultural dialects in vocal emotion expression: acoustic classification within and across five nations. *Emotion*, 14(3), 445–449. doi:10.1037/a0036048 [PubMed: 24749633]
- Levitan R, & Hirschberg JB (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proceedings of INTERSPEECH*, 3081–3084. Retrieved from https://www.isca-speech.org/archive/archive_papers/interspeech_2011/i11_3081.pdf
- Lieberman P (1960). Some acoustic correlates of word stress in American English. *J Acoust Soc Am*, 32 451–454. doi:10.1121/1.1908095
- Lord SP, Sheng E, Imel ZE, Baer J, & Atkins DC (2015). More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior Therapy*, 46(3), 296–303. doi:10.1016/j.beth.2014.11.002 [PubMed: 25892166]
- MacCallum RC, Kim C, & Malarkey WB (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, 32, 215–253. doi:10.1207/s15327906mbr3203_1 [PubMed: 26761610]
- Magill M, Apodaca TR, Borsari B, Gaume J, Hoadley A, Gordon REF, ... 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. doi:10.1037/ccp0000250 [PubMed: 29265832]
- Marci CD, Ham J, Moran E, & Orr SP (2007). Physiologic correlates of perceived therapist empathy and social-emotional process during psychotherapy. *The Journal of Nervous and Mental Disease*, 195(2), 103–111. doi:10.1097/01.nmd.0000253731.71025.fc [PubMed: 17299296]
- Mennen I, Schaeffler F, & Docherty G (2012). Cross-language differences in fundamental frequency range: a comparison of English and German. *Journal of the Acoustical Society of America*, 131(3), 2249–2260. doi:10.1121/1.3681950 [PubMed: 22423720]
- Miller WR, Moyers TB, Ernst D, & Amrhein PC (2008). *Manual for the Motivational Interviewing Skill Code (MISC)*. Albuquerque, NM: University of New Mexico Retrieved from <http://casaa.unm.edu/download/misc.pdf>.
- Miller WR, & Rollnick S (2013). *Motivational Interviewing: Helping People Change* (3rd ed.). New York, NY: Guilford Press.
- Moyers TB, Martin T, Manuel JK, Hendrickson SM, & Miller WR (2005). Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment*, 28(1), 19–26. doi:10.1016/j.jsat.2004.11.001 [PubMed: 15723728]
- Moyers TB, & Miller WR (2013). Is low therapist empathy toxic? *Psychology of Addictive Behaviors*, 27(3), 878–884. doi:10.1037/a0030274 [PubMed: 23025709]
- Norcross JC (2011). *Psychotherapy Relationships that Work: Evidence-Based Responsiveness*. New York: Oxford University Press.
- Pépiot E (2015). Voice, speech and gender: male-female acoustic differences and cross-language variation in English and French speakers. *Corela*, 268, 1–13. doi:10.4000/corela.3783
- Preston SD, & de Waal FB (2002). Empathy: Its ultimate and proximate bases. *The Behavioral and Brain Sciences*, 25(1), 1–20; discussion 20–71. doi:10.1017/S0140525X02000018 [PubMed: 12625087]
- R Core Team. (2016). *R: A language and environment for statistical computing* [Computer program]. Vienna, Austria: R Foundation for Statistical Computing Retrieved from <https://www.R-project.org/>.
- Ramseyer F, & Tschacher W (2011). Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *Journal of Consulting and Clinical Psychology*, 79(3), 284–295. doi:10.1037/a0023419 [PubMed: 21639608]

- Reich CM, Berman JS, Dale R, & Levitt HM (2014). Vocal Synchrony in Psychotherapy. *Journal of Social and Clinical Psychology*, 33(5), 481–494. doi:DOI 10.1521/jscp.2014.33.5.481
- Schwartz VS, Rothpletz-Puglia P, Denmark R, & Byham-Gray L (2015). Comparison of standardized patients and real patients as an experiential teaching strategy in a nutrition counseling course for dietetic students. *Patient Education and Counseling*, 98(2), 168–173. doi:10.1016/j.pec.2014.11.009 [PubMed: 25433966]
- Tremblay A, Broersma M, Coughlin CE, & Choi J (2016). Effects of the Native Language on the Learning of Fundamental Frequency in Second-Language Speech Segmentation. *Frontiers in Psychology*, 7, 985. doi:10.3389/fpsyg.2016.00985 [PubMed: 27445943]
- Webb CA, Derubeis RJ, & Barber JP (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, 78(2), 200–211. doi:10.1037/a0018912 [PubMed: 20350031]
- Xiao B, Georgiou PG, Imel ZE, Atkins DC, & Narayanan SS (2013). Modeling therapist empathy and vocal entrainment in drug addiction counseling. *Proceedings of INTERSPEECH*, 2861–2865. Retrieved from https://www.isca-speech.org/archive/archive_papers/interspeech_2013/i13_2861.pdf
- Xiao B, Imel ZE, Georgiou PG, Atkins DC, & Narayanan SS (2015). “Rate My Therapist”: Automated Detection of Empathy in Drug and Alcohol Counseling via Speech and Language Processing. *PloS One*, 10(12), e0143055. doi:10.1371/journal.pone.0143055 [PubMed: 26630392]

Public Significance Statement

In this study, we replicated the methods of a first analysis which tried to use new technology (automated coding of speech features) to facilitate coding of empathy in clinical interactions. Results were not replicated and showed no support for this new method. The present findings call for new developments to measure empathy, an important active ingredient of clinical encounters and therapy.

Figure 1.
Studies inclusion procedures and sample sizes. GP = General practitioner.

Figure 2.

Posterior modes (circles) and 95% credible intervals for the correlation of patient and clinician mean F0 from random effects at minute and session level for high- and low-empathy sessions. MISC = Empathy global rating from the Motivational Interviewing Skill Code; TES = Therapist Empathy Scale.

Table 1

Empathy Ratings and Cutoffs

Empathy measure		Full Sample			Low Empathy Sessions		High Empathy Sessions	
		<i>n</i> sessions	Mean	(SD)	Cutoff	n	Cutoff	n
Study 1	MISC	208	4.41	(116)	3	42 (20.2%)	6	32 (15.4%)
	TES	208	3.92	(0.79)	3	32 (15.4%)	4.5	50(24.0%)
Study 2	MISC	204	2.90	(105)	2	84 (41.2%)	4	48 (23.5%)
	TES	204	3.28	(0.77)	2.5	25 (12.3%)	4	36 (17.6%)

Note. MISC = Empathy global rating from the Motivational Interviewing Skill Code; TES = Therapist Empathy Scale.

Table 2

Fixed Effect Predictors of Mean F0

Study	Empathy measure	Fixed Effect Coefficient	Estimate	Lower 95%	Upper 95%	MCMC <i>p</i> -value
Study 1	MISC	Patient speaker (intercept)	113.34	109.91	116.58	.0002
		Clinician speaker (intercept)	188.70	181.90	195.67	.0002
		Male clinician (difference)	-71.47	-79.12	-63.88	.0002
		High empathy session (difference)	8.73	3.92	13.42	.0004
	TES	Patient speaker (intercept)	114.06	110.70	117.42	.0002
		Clinician speaker (intercept)	187.66	182.07	193.65	.0002
		Male clinician (difference)	-70.53	-77.34	-63.87	.0002
		High empathy session (difference)	8.12	3.43	12.65	.0013
Study 2	MISC	Patient speaker (intercept)	185.81	181.59	190.04	.0002
		Clinician speaker (intercept)	200.16	195.72	204.62	.0002
		Male clinician (difference)	-64.10	-69.66	-58.51	.0002
		Male patient (difference)	-58.16	-63.25	-52.28	.0002
	TES	High empathy session (difference)	-1.93	-6.23	2.86	.3996
		Patient speaker (intercept)	190.91	184.13	196.98	.0002
		Clinician speaker (intercept)	202.55	194.39	211.84	.0002
		Male clinician (difference)	-63.02	-72.44	-54.02	.0002
		Male patient (difference)	-62.26	-70.70	-54.05	.0002
		High empathy session (difference)	-5.66	-11.83	0.51	.0783

Note. Interval ranges reflect 95% highest posterior density intervals. Patient's gender was not entered in Study 1 models since all patients were male. MISC = Empathy global rating from the Motivational Interviewing Skill Code; TES = Therapist Empathy Scale.