The final paper is available at Springer via http://dx.doi.org/10.3758/s13428-017-0915-5

DARMA: Software for Dual Axis Rating and Media Annotation

Jeffrey M. Girard and Aidan G. C. Wright
University of Pittsburgh

Continuous measurement systems provide a means of measuring dynamic behavioral and experiential processes as they play out over time. DARMA is a modernized continuous measurement system that synchronizes media playback and the continuous recording of two-dimensional measurements. These measurements can be observational or self-reported and are provided in real-time through the manipulation of a computer joystick. DARMA also provides tools for reviewing and comparing collected measurements and for customizing various settings. DARMA is a domain-independent software tool that was designed to aid researchers who are interested in gaining a deeper understanding of behavior and experience. It is especially well-suited to the study of affective and interpersonal processes, such as the perception and expression of emotional states and the communication of social signals. DARMA is open-source using the GNU General Public License (GPL) and is available for free download from `http://darma.jmgirard.com`.

Keywords: research software, continuous measurement, media annotation, observational measurement, inter-rater reliability

Full understanding of behavior and experience requires an appreciation of time-dependent patterns. However, traditional methods of observational measurement and self-reporting are ill-suited to this task. These methods tend to polarize into either macro-level (*gist*) analyses of large swaths of time or micro-level (*atomic*) analyses of discrete segments. Unfortunately, both of these approaches miss the continuous and dynamic flow of many psychological processes. Specialized methods are needed that can capture such processes as they unfold over time and across multiple dimensions.

Continuous measurement systems were developed to meet this need (Girard & Cohn, 2016). Such systems rely on raters to provide high frequency reports of their perceptions or experiences (e.g., what they see, hear, or feel). The burden of such repetition is reduced through clever engineering. Rather than repeatedly stopping raters to collect their reports, measurements are unobtrusively sampled from an apparatus that can be continuously manipulated by the rater. For example, raters may be trained to rotate a dial or move a joystick to indicate changes in their emotional state. Data about the current position of the apparatus can be collected at a high frequency and saved for later use. As a result, dimensional measurements can be captured in real time.

Continuous measurement systems were first pioneered by Gottman and Levenson (Gottman & Levenson, 1985) for the study of communication during marital interactions. Their "affect rating dial" was a circular plastic knob mounted to the arm of a chair that could be rotated 180 degrees to indicate the range of very negative to very positive feelings. This method and its derivatives have been used to study dyadic interactions, empathic accuracy, the impact of alcohol on anxiety, and emotional responses to multimedia (for a review, see Ruef & Levenson, 2007).

Later continuous measurement systems transitioned to software packages, as these implementations were more accessible to many users than custom-built electronic devices. Early software packages included ESL (Schubert, 1999) and FEELtrace (Cowie et al., 2000), while later packages included RTCRR (Schubert, 2007), EMuJoy (Nagel, Kopiez, Grewe, & Altenmuller, 2007), CMS (Messinger, Cassel, Acosta, Ambadar, & Cohn, 2008), JoyMon (Sadler, Ethier, Gunn, Duong, & Woody, 2009), Gtrace (Cowie, McKeown, & Douglas-Cowie, 2012), and CARMA (Girard, 2014). Table 1 lists these packages and provides information about their features and dependencies.

One continuous measurement software that bears highlighting is CARMA, which included a novel suite of features for reviewing collected measurements. These features increase the efficiency of training and quality control. CARMA requires measurements to be made on one dimension at a time, however, and many research areas are primarily interested in collecting two-dimensional measurements.

Two-dimensional measurements often derive from *circumplex* models, which arrange variables into a circular space defined by two continuous (and theoretically bipolar)

Correspondence concerning this article should be addressed to Jeffrey Girard, 210 S. Bouquet St, Department of Psychology, Pittsburgh, PA 15260. Email: j.girard@pitt.edu

dimensions (Gurtman, 1994). The intersection of two dimensions allows for much richer description than either dimension could provide alone, as they provide coverage of the interstitial space defined by blends of the two dimensions. Two popular circumplex models are the affective circumplex and the interpersonal circumplex. The affective circumplex (Russell, 1980) describes emotion using the dimensions of valence (unpleasant to pleasant) and arousal (low energy to high energy), while the interpersonal circumplex (Fournier, Moskowitz, & Zuroff, 2010; Horowitz et al., 2006) describes social interaction using the dimensions of agency (submission to dominance) and communion (coldness to warmth). For example, the affective experience of intense fear could be characterized as negative in valence (i.e., unpleasant) and positive in arousal (i.e., high energy), whereas the interpersonal behavior of withdrawing and sulking could be characterized as negative in agency (i.e., submissive) and negative in communion (i.e., cold or unfriendly).

Researchers in affective computing (Cowie et al., 2012; Gunes & Schuller, 2013) and music psychology (Geringer, Madsen, & Gregory, 2004; Juslin & Sloboda, 2011) routinely collect ratings of perceived and experienced emotion using the affective circumplex. Similarly, it is becoming increasingly common in clinical and personality psychology (Lizdek, Sadler, Woody, Ethier, & Malet, 2012; Markey, Lowmaster, & Eichler, 2010; Tracey, Bludworth, & Glidden-Tracey, 2012) to collect ratings of social communication using the interpersonal circumplex. As such, a feature-rich software package for collecting two-dimensional/circumplex ratings is needed.

The current paper describes DARMA, an open-source software package for collecting continuous two-dimensional measurements that builds upon CARMA. It can be used to collect observational measurements of behavior or self-reports of various aspects of experience. While it lends itself most readily to the study of affective and interpersonal processes, DARMA can be customized to collect two-dimensional measurements of any kind.

Table 1
Continuous measurement software packages with feature information

Software	1D Annotation	2D Annotation	Synchronization	Customization	Review Tools	Open-Source	Platform
ESL (Schubert, 1999)	–	•	•	–	–	–	MacOS
FEELtrace (Cowie et al., 2000)	–	•	•	–	–	–	Windows
RTCRR (Schubert, 2007)	•	•	•	•	–	–	MacOS
EMuJoy (Nagel et al., 2007)	–	•	•	•	–	•	Java
CMS (Messinger et al., 2008)	•	–	•	•	–	•	Windows
JoyMon (Sadler et al., 2009)	–	•	–	•	–	–	Windows
Gtrace (Cowie et al., 2012)	•	–	•	•	–	•	Windows
CARMA (Girard, 2014)	•	–	•	•	•	•	Windows
DARMA	–	•	•	•	•	•	Windows

Case Study

To illustrate a potential use of DARMA, we provide a case study analyzing the interpersonal dynamics of romantic couples with varying levels of personality pathology. This work was first presented by Girard, Wright, Stepp, and Pilkonis (2016, May).

Personality pathology is characterized by maladaptive patterns of behavior, cognition, and experience that are pervasive, inflexible, and distressing. In particular, personality pathology is marked by difficulty in one or more of the following domains: establishing a stable and positive identity, setting and achieving realistic goals, establishing and maintaining intimate relationships, and empathizing with others' feelings and motivations (American Psychiatric Association,

2013). Given the interpersonal nature of the latter two domains, we hypothesized that personality pathology would influence patterns of interpersonal behavior between romantic partners, especially during moments of conflict.

A total of 74 romantic couples were sampled from a larger study on the influence of personality pathology on romantic functioning. Roughly half of these participants met diagnostic criteria for at least one personality disorder. Couples engaged in a 10 min conflict discussion, during which they discussed problems in their relationship; discussions were videotaped for later analysis. Six undergraduate researchers were trained to use DARMA for the observational measurement of interpersonal behavior (Lizdek et al., 2012). After watching the video of each discussion once to gain an appreciation of its broader context, raters provided continuous ratings of each couple member's agency and communion during the conflict discussion. Raters met once per week to review and compare their ratings using DARMA. The average of all six raters was used for analysis; mean inter-rater reliability was excellent at ICC = .85 for agency and ICC = .75 for communion.

Multilevel structural equation modeling (Heck, 1999) was used to evaluate the influence of personality pathology on general patterns of behavior, as well as the moment-to-moment association between couple members' behavior. Higher levels of personality pathology predicted significantly lower average levels of communion (i.e., a more cold interpersonal stance). This finding may explain the interpersonal problems associated with personality pathology and provide a target for therapeutic intervention. Regardless of personality pathology, moment-to-moment, couple members' communion ratings were significantly and positively associated and their agency ratings were significantly and negatively associated. This finding provides support for the theory of interpersonal complementarity (Kiesler, 1983)—that people tend to respond to warmth with warmth, to coldness with coldness, to dominance with submission, and to submission with dominance—and suggests that the behavior of one's partner is an important contextual factor that should not be ignored.

Software Architecture

DARMA was written in the MATLAB programming language and compiled into a standalone application for Microsoft Windows (i.e., 64-bit Windows 7 or newer) using the MATLAB Compiler. MATLAB is a commercially-available software package and programming language from MathWorks, Inc. that is commonly used in psychology, computer science, and related fields. Note that the MATLAB software package is not required to use DARMA but is required to modify DARMA's source code.

DARMA has two software dependencies. First, the open-source VLC Media Player is required. This software is embedded into DARMA using an ActiveX plugin and is used to control playback of a wide variety of media files. Second, the freely-available MATLAB Runtime is required. This software provides a set of shared libraries that enables the execution of compiled MATLAB applications. The appropriate version of the MATLAB Runtime is automatically downloaded and installed alongside DARMA.

Additionally, in order to collect ratings on two dimensions simultaneously, DARMA requires users to manipulate a computer joystick. Any device that is recognized by Windows as a joystick will work, although full-sized joysticks are recommended due to their greater precision of control. Such devices can usually be purchased for $15 to $50. We have used the Logitech Extreme 3D Pro for several years now and would recommend it.

Software Functionalities

DARMA has two windows that each correspond to one of its primary functions. The *Collect Ratings* window allows users to collect new continuous measurements (i.e., ratings) and save them to 'annotation files,' while the *Review Ratings* window allows users to examine previously saved annotation files alongside the media files that they describe.

Collecting new two-dimensional annotations

The Collect Ratings window enables users to synchronize the playback of media files and the collection of continuous measurements. Upon first opening, this window displays two large rectangular panels. On the left is a panel that displays media files as they play, and on the right is a panel that contains a visual representation of the two-dimensional measurement space. Labels can be configured to appear in different regions of this space to remind users what measurements in that region correspond to (e.g., Figure 1 shows that behavior that is high in communion and moderate in agency is often considered "friendly"). While any alphanumeric labels can be specified, labels for the popular interpersonal (Horowitz et al., 2006) and affective (Russell, 2003) circumplex models are built into the program. This panel also displays the *rating indicator:* a moving circle that depicts the current position of the joystick.

After opening a media file, information about that file is displayed along the bottom of the window. Users can then click the 'Begin Rating' button to begin media playback and measurement collection. Playback and collection can be paused and resumed at any time by clicking the 'Pause/Resume Rating' toggle button or by pressing the space bar on the keyboard.

The position of the joystick is sampled at a rate of 20 Hz. This rate was chosen to strike a balance between data redundancy and computational efficiency. After each sample is written to memory, the position of the rating indicator is updated to match it. Samples are ultimately combined into

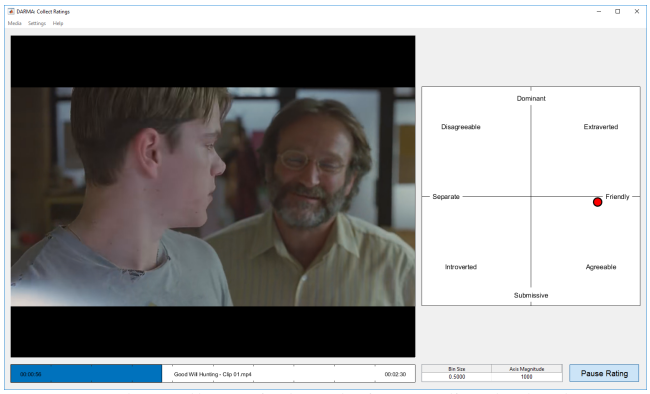

Figure 1. The Collect window during media playback

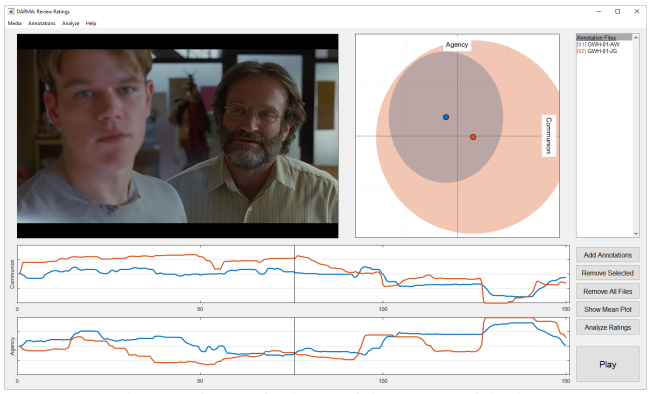

Figure 2. The Review window with a centroid plot

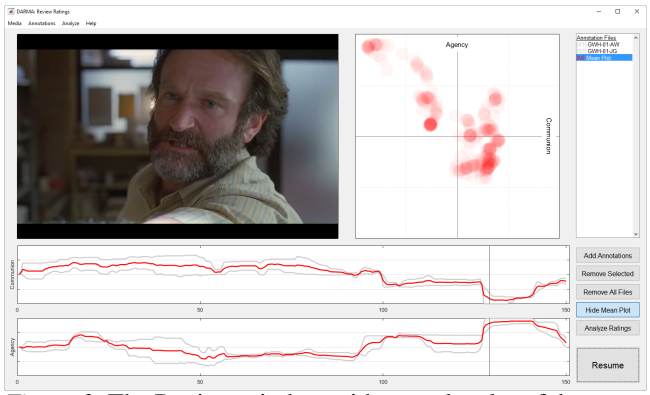

Figure 3. The Review window with a smoke plot of the mean

temporal bins by averaging the values of all samples that occurred between specified start and stop points. The temporal location of these start and stop points can be customized, which enables users to record measurements on a time scale that matches the constructs being measured.

At the conclusion of the media file, the user is prompted to save their measurements to a comma-separated values (.csv) file. These annotation files contain each bin's average value and corresponding timestamp (i.e., stop point); they also contain metadata about DARMA's configuration.

Reviewing annotations and reliability

The Review window (Figures 2 and 3) enables users to view previously saved annotation files alongside the original media file. Annotation files from any number of raters can be loaded simultaneously. Options are provided for visualizing each rater's measurements and for assessing inter-rater agreement and reliability. The mean of multiple raters' measurements can also be calculated, visualized, and exported. There are two types of visualization.

Toward the bottom of the window, ratings are represented as time-series (i.e., line graphs of ratings against time). Clicking on any segment of these time-series will seek to that moment in the media file, which facilitates the training of raters and the review of ratings. Toward the top of the window, the distribution of measurements can be displayed. 'Smoke' plots provide detailed information about a single rater (or the mean series), while 'centroid' plots enable comparison between multiple raters.

Tools for analyzing and interpreting ratings are also available (and are in the process of being expanded). For instance, the *Analyze Ratings* option calculates and displays descriptive statistics for each annotation file, as well as estimates of their inter-rater agreement and reliability (Girard & Cohn, 2016; LeBreton, Burgess, Kaiser, Atchley, & James, 2003; McGraw & Wong, 1996). More options are planned for time-series analysis.

Impact

DARMA is a powerful new tool for capturing the streams of behavior and experience as they unfold over time. It is the first software implementation of the continuous measurement system that combines playback-collection synchronization, qualitative and quantitative review tools, and the ability to collect two-dimensional measurements (Table 1). These features increase the efficiency and quality of important research tasks that are currently being tackled with older technology such as JoyMon (Sadler et al., 2009) or EMuJoy (Nagel et al., 2007).

Automatic synchronization guarantees that a given set of measurements align precisely with both the media file they describe and with other measurements of the same file (e.g., DARMA measurements from other raters or measurements from external sources such as psychophysiology recordings). Such alignment is essential for analyses of inter-rater reliability and interpersonal (e.g., dyadic or triadic) influence to be meaningful. Furthermore, it enables users to pause and resume their task whenever necessary. This ability allows longer media segments to be measured at one time and prevents data from being lost due to unplanned interruptions.

Annotation review is a powerful tool for observer training and quality control. As suggested by Girard and Cohn

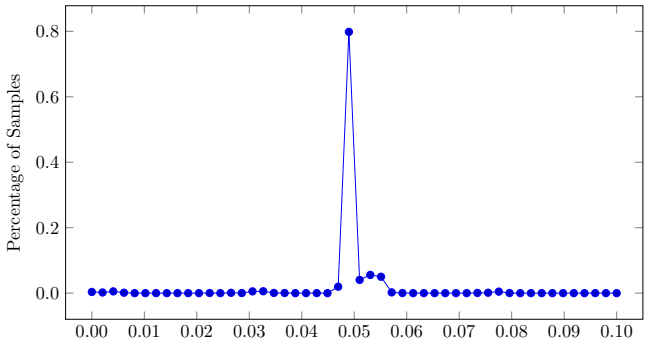

Figure 4. Delay between consecutive joystick samples (in s)

(2016), unreliability between observers can often be identified and resolved using "observer meetings." During such meetings, observers' (i.e., raters') measurements are randomly-selected to be inspected and discussed. This practice provides excellent opportunities for teaching and reinforcing the use of proper *response processes* (i.e., criteria for assigning different measurements). As such, it can increase the validity of measurements in addition to their reliability (Cizek, 2016). DARMA provides several features aimed at enhancing and streamlining annotation review. These features include the ability to visualize multiple time-series alongside the media file, the ability to seek (i.e., jump playback) to different parts of the media file by clicking on the measurement time-series, numerous visualizations of the data, and estimates of measurements' distributions and reliability.

Evaluation

To evaluate the usability of the software for its intended purpose, a group of six users (i.e., undergraduate researchers at the University of Pittsburgh) were closely observed over a period of 3 months in 2014. In the year following this period, while the data for (Ross et al., 2017) was being rated, these users were encouraged to provide feedback on any errors and confusions they encountered. This observation period and the users' feedback uncovered many opportunities to improve the stability and usability of the software; ten minor and three major version updates (numbered v2.00 to v5.00) correspond to these improvements. Furthermore, because of the importance of synchronization between media and annotation files, we conducted systematic testing of the delay between consecutive joystick samples. In order to attain our target sampling rate of 20 Hz, the delay should equal 0.050 s. Our analysis found that 80 % of samples showed a delay between 0.049 and 0.051 s, while 95 % of samples showed a delay between 0.049 and 0.057 s (Figure 4).

To evaluate user satisfaction regarding the software, an online survey was sent to researchers who have used continuous measurement software in the past, most of whom are experts in the field of interpersonal psychology. These researchers were asked to complete the survey themselves, as well as to forward it on to relevant others. The final sample (n = 12) consisted of two professors/research scientists, three graduate students, five undergraduate students, and two respondents who selected "other" as their title. All respondents indicated that they were using DARMA v5.00 or later.

The survey included questions about the software in general and about specific features (e.g., "How satisfied are you with..."). A seven-point scale was used where a score of 1 corresponded to "extremely dissatisfied/unlikely" and a score of 7 corresponded to "extremely satisfied/likely." A text entry box was also provided after each question to collect unstructured feedback.

Responses to the survey indicate that most users (82 %) were extremely satisfied with DARMA overall (M = 6.73). Most users (82 %) also indicated that they were extremely likely to recommend DARMA to other researchers (M = 6.64). Users were, on average, moderately to extremely satisfied with the following aspects of the software: the website (M = 6.18), downloading the software (M = 6.00), configuring the software (M = 6.73), collecting ratings (M = 6.82), reviewing ratings (M = 6.55), and using ratings for statistical analysis (M = 6.36). Users were, on average, slightly to moderately satisfied with the process of installing the software (M = 5.82) and with the software's documentation (M = 5.91).

Future Work

In response to the user satisfaction survey, future updates will focus on improving the clarity of the installation process and making documentation and technical support more accessible to users. Other updates are planned to expand the number of customization options, to add more tools to streamline the analysis and management of annotation files, and to enable the creation of 'queues' of multiple media files to be rated sequentially.

Conclusions

Psychological processes ranging from affective experience to interpersonal communication are being increasingly recognized as dynamic processes that shift over time and across multiple dimensions. DARMA provides a much-needed tool for capturing these—and other—processes continuously and in real-time. A number of essential features set it a step above existing two-dimensional continuous measurement systems such as playback-collection synchronization, annotation review tools, and painless customization.

Acknowledgments

Research reported in this publication was supported in part by the National Institutes of Health under award num-

ber MH096951 and GM105004. The content is solely the responsibility of the authors and does not necessarily represent the views of the National Institutes of Health. Images of Matt Damon and Robin Williams from the film "Good Will Hunting" appear in several figures for demonstrative purposes only; the film was produced by Lawrence Bender Productions and distributed in the US by Miramax Films.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 212–225.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). FEELTRACE: An instrument for recording perceived emotion in real time. *ISCA Tutorial and Research Workshop on Speech and Emotion*, 19–24.

Cowie, R., McKeown, G., & Douglas-Cowie, E. (2012). Tracing emotion: an overview. *International Journal of Synthetic Emotions*, *3*(1), 1–17.

Fournier, M. A., Moskowitz, D. S., & Zuroff, D. C. (2010). Origins and applications of the interpersonal circumplex. In L. M. Horowitz & S. Strack (Eds.), *The handbook of interpersonal psychology* (pp. 57–73). Hoboken, NJ: John Wiley & Sons.

Geringer, J. M., Madsen, C. K., & Gregory, D. (2004). A fifteen-year history of the Continuous Response Digital Interface: Issues relating to validity and reliability. *Bulletin of the Council for Research in Music Education*, *160*, 1–15.

Girard, J. M. (2014). CARMA: Software for continuous affect rating and media annotation. *Journal of Open Research Software*, *2*(1), e5.

Girard, J. M., & Cohn, J. F. (2016). A primer on observational measurement. *Assessment*, *23*(4), 404–413.

Girard, J. M., Wright, A. G. C., Stepp, S. D., & Pilkonis, P. A. (2016, May). *Interpersonal dynamics in couples with personality pathology.* Symposium conducted at the meeting of the Association for Psychological Science, Chicago, IL.

Gottman, J. M., & Levenson, R. W. (1985). A valid procedure for obtaining self-report of affect in marital interaction. *Journal of Consulting and Clinical Psychology*, *53*(2), 151–160.

Gunes, H., & Schuller, B. W. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, *31*(2), 120–136.

Gurtman, M. B. (1994). The circumplex as a tool for studying normal and abnormal personality: A methodological primer. In S. Strack & M. Lorr (Eds.), *Differentiating normal and abnormal personality* (pp. 243–263). New York, NY: Springer.

Heck, R. H. (1999). Multilevel modeling with SEM. In S. L. Thomas & R. H. Heck (Eds.), *Introduction to multilevel modeling techniques* (pp. 89–127). Mahwah, NJ: Lawrence Erlbaum Associates.

Horowitz, L. M., Wilson, K. R., Turan, B., Zolotsev, P., Constantino, M. J., & Henderson, L. (2006). How interpersonal motives clarify the meaning of interpersonal behavior: A revised circumplex model. *Personality and Social Psychology Review*, *10*(1), 67–86.

Juslin, P. N., & Sloboda, J. A. (Eds.). (2011). *Handbook of music and emotion: Theory, research, applications*. Oxford, UK: Oxford University Press.

Kiesler, D. J. (1983). The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological Review*, *90*, 185–214.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, *6*(1), 80–128.

Lizdek, I., Sadler, P., Woody, E., Ethier, N., & Malet, G. (2012). Capturing the stream of behavior: A computer-joystick method for coding interpersonal behavior continuously over time. *Social Science Computer Review*, *30*(4), 513–521.

Markey, P., Lowmaster, S., & Eichler, W. (2010). A real-time assessment of interpersonal complementarity. *Personal Relationships*, *17*(1), 13–25.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46.

Messinger, D. S., Cassel, T. D., Acosta, S. I., Ambadar, Z., & Cohn, J. F. (2008). Infant smiling dynamics and perceived positive emotion. *Journal of Nonverbal Behavior*, *32*(3), 133–155.

Nagel, F., Kopiez, R., Grewe, O., & Altenmuller, E. (2007). EMuJoy: Software for continuous measurement. *Behavior Research Methods*, *39*(2), 283–290.

Ross, J. M., Girard, J. M., Wright, A. G. C., Beeney, J. E., Scott, L. N., Hallquist, M. N., ... Pilkonis, P. A. (2017). Momentary patterns of covariation between specific affects and interpersonal behavior: Linking relationship science and personality assessment. *Psychological Assessment*, *29*(2), 123–134.

Ruef, A. M., & Levenson, R. W. (2007). Continuous measurement of emotion: The affect rating dial. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 286–297). New York, NY, US: Oxford University Press.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172.

Sadler, P., Ethier, N., Gunn, G. R., Duong, D., & Woody, E. (2009). Are we on the same wavelength? Interpersonal complementarity as shared cyclical patterns during interactions. *Journal of Personality and Social Psychology*, *97*(6), 1005–1020.

Schubert, E. (1999). Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology*, *51*(3), 154–165.

Schubert, E. (2007). Real time cognitive response recording. In *The inaugural international conference on music communication science* (pp. 139–142).

Tracey, T. J. G., Bludworth, J., & Glidden-Tracey, C. E. (2012). Are there parallel processes in psychotherapy supervision? An empirical examination. *Psychotherapy*, *49*(3), 330–343.