Appendix

Example implementation of recommendations in Issues 1 and 2

Implementing the recommendations in Issues 1 and 2 in future research designed to disentangle behavioral associations for related outcomes requires simultaneous consideration of multiple factors. The recommendations in Issue 1 are designed to provide unbiased tests of convergent and discriminant validity while the recommendations in Issue 2 are designed to increase the sensitivity of associations involving observationally coded data by helping researchers determine how many coders are needed to achieve an Interrater Reliability of .7 or higher. These two sets of recommendations can be used in conjunction to determine how to allocate limited resources when creating a new coding system to measure disease/disorder specific behaviors.

Suppose that a researcher wished to develop an observational coding system for measuring specific communication behaviors that partners who are trying to stop smoking exhibit during conflict that are not otherwise accounted for by relationship satisfaction or general communication behaviors. Drawing on the principles of Motivational Interviewing (e.g., Miller & Rollnick, 2012), the researcher hypothesizes that partners who use higher levels of change talk, which is language emphasizing an individual’s reasons for stopping smoking, and lower levels of sustain talk, which is language emphasizing an individual’s reason for continuing to smoke, will be significantly more successful in quitting smoking. In order to test this hypothesis, the researchers plan to test associations between the number of cigarettes smoked per day, sustain talk, and change talk. In doing so, s/he wants to ensure that these associations are not accounted for by general communication behaviors known to be associated with smoking (e.g., demand-withdraw behavior and negative reciprocity; Rohrbaugh et al., 2001) or by relationship
satisfaction.

The researcher uses existing methodological recommendations (e.g., Heyman, 2001; Margolin et al., 1998) to design the observational task in the study and recruits 50 heterosexual couples in which the male partner has smoked at least daily for the past year, the male partner has been trying to quit smoking for no more than one week, and the female partner does not smoke. Couples are asked to discuss their thoughts and feelings about the smoking partner’s attempts to quit for 10 minutes (Heyman et al., 2002). Smoking is measured using a timeline follow-back interview (Brown et al., 1998), and relationship satisfaction is measured using the Couples Satisfaction Inventory, 16-item version (Funk & Rogge, 2007).

The researcher decides to code the 10-minute thoughts and feelings interviews twice, once with the coding system s/he is developing and once with the NORS (K. Baucom et al., 2012). The researcher decides to use the NORS because it measures the behavioral constructs s/he is interested in measuring and because the researcher does not have the time or resources s/he would need to code that recordings with two trained coding systems. The researcher is specifically interested in ensuring that s/he can be confident in obtaining Interrater Reliabilities of .7 or higher for the demand-withdraw and negative reciprocity codes. The researcher consults Table 2 and determines that s/he needs to recruit three naïve coders for the NORS coding. Negative reciprocity has the lower Interrater Reliability value of the two codes and exceeds .7 for coding teams of three coders or larger.

For observational coding of change and sustain talk, the researcher pilot tested the reliability of these trained codes in another sample of male smoker couples. A team of five trained coders generated Interrater Reliabilities of .65 for change talk and .72 for sustain talk. The researcher uses the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910)
presented below to estimate how many additional coders s/he would need to likely obtain an Interrater Reliability of .7 or greater for change talk. Six coders would likely result in an Interrater Reliability of approximately .69 \( ([1.2 \cdot 0.65]/[1+(1.2-1) \cdot 0.65]) \), and seven would likely result in an Interrater Reliability of approximately .72 \( ([1.4 \cdot 0.65]/[1+(1.4-1) \cdot 0.65]) \). Because the Spearman-Brown prophecy formula is an approximate estimate, the researcher decides to recruit seven coders for the trained rating team that will code change and sustain talk.

The researcher’s primary hypothesis involves testing convergent validity between change talk, sustain talk, and the number of cigarettes smoked per day while accounting for shared variance involving woman demand/man withdraw, negative reciprocity and relationship satisfaction. The researcher also wants to perform tests of discriminant validity by testing associations between change talk, sustain talk, woman demand/man withdraw, negative reciprocity and relationship satisfaction. The researcher decides to perform both tests at the same time using the model depicted in Figure A1. The researcher’s sample size is not large enough to generate stable estimates using standard methods for estimating Structural Equation Models so s/he decides to use either a multivariate multilevel model (e.g., Baldwin, Imel, Braithwaite, & Atkins, 2014) or a Bayesian Structural Equation Model (Lee & Song, 2004) to test study hypotheses.

**Details of processing and modeling steps in BSP**

BSP methods for estimating code scores involves a series of data processing and modeling steps\(^1\) that must be accomplished before a code value can be estimated. In this section

\(^1\) BSP can incorporate a wide range of communicative modalities such as lexical (i.e., what was said), acoustic (i.e., how it was said), body posture, kinematics (i.e., physical movement), and visual features. We restrict our discussion to lexical and acoustic features as those have the most substantial body of empirical evidence in couple interactions, even though we have work supporting the behavioral value of the visual modality (e.g., Xiao, Georgiou, Baucom, & Narayanan, 2014).
of the appendix, we describe these processing and modeling steps at a level of detail that is intended to assist researchers who wish to do future research with BSP in making decisions about study design and recording of couple interactions. We also strongly encourage researchers interested in using BSP methods to discuss study design with their Signal Processing collaborators prior to beginning data collection. There are unique challenges to implementing BSP methods that are easily discernable to Machine Learning and Signal Processing researchers that are unlikely to occur to researchers new to BSP. Examples of these kinds of challenges include making decisions about whether and how to collect an acoustic baseline, which recording equipment will provide optimal data quality (e.g., omnidirectional versus directional microphones), and how to position and setup recording equipment (e.g., angle of a video camera relative to participants, how to maximize consistency of acoustic and visual recording conditions across participants, etc.).

Figure A1 displays the data processing steps for the lexical (i.e., what was said) and paralinguistic (i.e., how it was said) modalities which include denoising, speech activity detection, diarization, speaker identification, automatic speech recognition, and feature extraction. Denoising is the first step in the data processing chain and involves removing noise from the audio signal. For example, denoising could involve removing ambient noise such as the hum of an air conditioner from the audio recording. Denoising is important because it improves the accuracy of the second step, speech activity detection. Speech activity detection involves determining when the two partners are speaking and when there are other sources of noise that are not speech (e.g., a phone ringing, construction noise outside a window, etc.). Participants’ speech is then further analyzed while everything else is ignored so the end result of this step is that the periods of time that need further processing are identified. These periods of speech are
then passed to the diarization and speaker identification steps. Diarization refers to determining periods of consecutive time when only one partner is talking, and speaker identification refers to determining which partner is talking during each period of consecutive speech. These two steps are very similar to the “chunking” step in microanalytic coding systems (e.g., Bakeman & Gottman, 1997) where a recording is broken down into smaller intervals prior to being coded.

The next step, automatic speech recognition, refers to generating a transcription of the words that were said during each period of consecutive speech. The transcript generated using BSP methods is different from a transcript that would be generated by having a trained research assistant listen to the recording and type the words that are spoken verbatim. BSP generates a probabilistic transcript called a lattice. A lattice (see Figure A2) generates several possibilities for each spoken word and estimates the probability of each word being spoken in that moment given several sources of information (e.g., what was said immediately prior, what is most likely given the acoustic information during that period of time [i.e., phonetic information], etc.).

Figure A2 shows the words that the ASR considers as possibilities when processing two recordings of the phrase, “I use observational coding to study relationships.” The numbered circles indicate each of these possibilities and the word following the colon above each line is the actual word ASR considers. The check marks in front of these words indicates the words that the ASR determines to be the most likely of the options. For example, in the lattice in the top of Figure A2, the ASR considers the words coding, coal, cold, code, coleen, holding, colluding and according as possibility for the fourth spoken word of the phrase. Likewise, the lattice in the bottom of Figure A2 shows that the ASR considers the words clothing, closing, and coding for the fourth spoken word. The accuracy of the lattice is influenced by numerous factors including the quality of the recording and the similarity between the speech in the recording and the speech
on which the ASR is trained (the left most column of Figure A1).

The final data processing step is feature extraction. Feature extraction refers to quantifying the paralinguistic and lexical information prepared in the prior processing steps. This step is very similar to other kinds of signal processing methods commonly used in psychological science such as psychophysiological research. For example, an electrocardiogram (ECG) is a recording of the electrical activity in the heart that is represented as a complex waveform (i.e., it has several different forms of oscillation that are combined to create a single ECG wave). The ECG wave in and of itself is not meaningful but can be made so by quantifying properties of the wave. For example, heart rate is the number of R waves (i.e., the most pronounced upward inflection in the ECG wave) that occur within a 60 second period of time. Paralinguistic and lexical feature extractions work the exact same way. Some lexical and acoustic features have perceptual correlates (e.g., fundamental frequency of speech is perceived as pitch, a unigram refers to the number of times that different words were spoken during the whole recording, etc.) and others do not. This quality is also true of psychophysiological features. For example, heart rate can be felt by taking someone’s pulse but Respiratory Sinus Arrhythmia (a measure of the respiratory control of the heart, understood to index vagal control; Porges, 2001) can only be calculated with an algorithm. Regardless of whether there is a perceptual correlate for a psychophysiological feature or not, they are understood to provide valuable information about physiological activity; the same is true for lexical and paralinguistic features, and the information they provide about communication.

The modeling step of BSP uses lexical and paralinguistic features estimated in the data processing steps to estimate values for the behaviors in a coding manual. BSP achieves this aim through a form of machine learning known as supervised learning. Supervised learning refers to
using recordings that have been coded to figure out which lexical and acoustic features lead to the highest accuracy of estimated coding values\(^2\). This process occurs in two steps. First, existing recordings that have been coded are broken down into two groups, a training set and a test set. Each recording can only be used in one of these two sets at a time (and we ensure that all recordings from couples in test are removed from training sets to avoid learning couple-specific traits); this two-step process is repeated many times so each recording will be in the training and evaluation set by the end of the modeling. This process, often called “leave-one-out” or “round-Robin,” is employed to achieve maximum statistical validity of the results with small datasets.

The training set is used to learn functions that map the lexical and acoustic features of an existing recording to observational coding values for those recordings. These functions are determined using machine learning techniques like Support Vector Machines, Hidden Markov Models, and Deep Neural Networks. These functions are then used to estimate the coding values for recordings in the test set, and the estimated coding values are compared to the existing coding values for the test set recordings to determine the accuracy of the estimated coding values. This step is very similar to the comparison of estimated values of the outcome variable in a regression, \(\hat{Y}\), to the observed values of the outcome variable, \(Y\), to calculate the proportion of variance in \(Y\) accounted for by the regression, \(R^2\). If a high level of concordance is achieved, the derived functions can be used to estimate coding values for new recordings or existing recordings that have not been coded (i.e., to perform “automated coding”). Ongoing research is investigating the use of unsupervised ML methods for learning behavioral patterns without requiring the availability of a training set (Li, B. Baucom, & Georgiou, 2017). If successful, the availability of unsupervised methods for learning behavioral patterns would offer couple

\(^2\) Hladka and Holub (2015) and Chen and Wojcik (2016) provide accessible introductions to machine learning and other related techniques.
researchers an entirely new method for exploratory analysis of behavior during couple interactions.

A statistical method for estimating how many coders are needed to achieve a desired level of interrater reliability

One question that may arise in considering the possibility of increasing the size of coding teams is how big does the coding team need to be? There are no well-established methods for determining the precise size that a coding team needs to be, but the Spearman-Brown prophecy formula (SBP) provides a method for making a rough estimate that can be used for planning purposes. The SBP estimates the reliability of a test of increased size based on existing psychometric information about the test:

\[
\hat{\rho}_{x,x'} = \frac{n \cdot \rho_{x,x'} }{1 + (n-1) \rho_{x,x'}}
\]

where \(\hat{\rho}_{x,x'}\) is the predicted interrater reliability, \(\rho_{x,x'}\) is the observed interrater reliability, and \(n\) is the proportion of increase or decrease in the number of coders. For example, the average interrater reliability of relationship quality for 2 coders is .66 (see Table 2). The SBP predicts that doubling the number of coders from 2 to 4 would result in an interrater reliability of .80 \([2 \cdot .66] / [1 + (2-1) \cdot .66] = 1.32 / 1.66 = .80\); the observed, average interrater reliability of relationship quality with 4 coders is .80 (see bottom rows of Table 2).

As can be seen by comparing the top and bottom rows of each cell in Table 2, the SBP predicted interrater reliability is more accurate for some codes than for others. For six of the seven codes, the predicted interrater reliability is within .03 of the observed interrater reliability. The discrepancy between the estimated and observed interrater reliabilities for mutual avoidance is larger and approaches .10 in some cases. These estimates are clearly imprecise, but they offer a method for generating data that can be incorporated into decisions about how large a coding
team is likely to be needed when psychometric information for an existing coding scheme is available or when developing a new coding system and scaling up from pilot data collection to full study data collection.
Appendix References


Guilford press.


Figure A1. Model for testing convergent and discriminant validity of smoking specific behavioral codes.

Note. The double-hatched paths indicate tests of convergent validity, and the dashed paths indicate tests of discriminant validity.
Figure A2. BSP pipeline for estimating observational codes.
Human speech: I use observational coding to study relationships

Spoken twice by 2 different speakers

1. ASR Hypotheses:
   i use observational coding to study relationships

2. ASR Hypotheses:
   i use observational clothing to study relationships

Note that alternate hypotheses contain word coding

Figure A3. Sample lattice for the phrase, “I use observational coding to study relationships”