REDUNDANCY ANALYSIS OF BEHAVIORAL CODING FOR COUPLES THERAPY AND IMPROVED ESTIMATION OF BEHAVIOR FROM NOISY ANNOTATIONS

Md Nasir[1]*, Brian Baucom*[2]*, Panayiotis Georgiou*[1]*, Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, CA, USA
[2]Dept. of Psychology, University of Utah, Salt Lake City, UT, USA.

ABSTRACT

Assessment and quantification of behavior is an important research objective in the recently developed field of behavioral signal processing. This paper focuses on the estimation of behavior from noisy human assessment. It aims to address the redundancy of behavioral descriptors for couples therapy by introducing a lower-dimensional representation of the behavioral space. We present an improved method for estimating the ground truth of behavioral ratings from assessment by multiple experts or annotators. The results show improved estimation performance using the proposed method and provide an insightful analysis of reconstruction error and decorrelation of annotator bias in the reduced behavioral space.

Index Terms— Behavioral Signal Processing (BSP), behavioral informatics, bias, noise, annotator

1. INTRODUCTION

Understanding human behavior is one of the central themes of modern psychological and sociological research and practice. Human expression and perception of behavior highly influence human interactions and relationships on a daily basis. Given the importance and the widespread nature of the problem domain, computational modeling of human behavior is becoming an area of increasing activity bringing together the fields of psychology and social sciences with data analytics [1, 2]. For instance, [3, 4] have presented work relating to understanding couple behaviors, [5] in characterizing autism, and [6, 7] in analysis of Motivational Interviewing based psychotherapy.

Psychologists often use a range of behavioral rating systems to quantify different abstract attributes of human behavior, typically along rating scales (e.g., range 1-7) using constructs such as empathy, blame, humor, anxiety, and satisfaction [8, 9]. These behavioral descriptors are often constructed with the application domain needs in mind, but not necessarily with in-depth understanding of the human limitations in disambiguating behavioral codes from audio-visual observation. Furthermore, behavioral descriptions for a given application often are abstract and subjective and hence tend to have varying amounts of descriptive redundancy leading to correlation between the behavioral codes.

The codes are estimated using observed human behavior data (e.g., audio, video, text) and fundamental questions remain as to the information capacity of such data with respect to behavior codes [1]. Moreover, the inter-relation between the multiple manually-specified codes by experts for a given application is not quantitatively established. To address these questions, we can explore the possibility of conveying as much information on the behavior as possible using fewer codes through a transform of the original codes.

In conventional behavioral analysis, multiple human annotators are employed to derive the behavioral codes from observed data under the guidelines of a specified coding system. The heterogeneity and subjectivity in perception of behavior is reflected in the variability of the derived codes. To fuse the evaluations by multiple annotators and obtain the (latent) ground truth or estimates reflecting the subject's true behavior is still an open problem. While there has been significant research in fusion of discrete categorical labels (*nominal* data), such as classifying emotion types[10] or labeling images[11, 12], in most applications multiple evaluations of ratings (*ordinal* data) on behavioral codes or emotional primitives are simply averaged to get the 'true' estimate. In [13, 14], a more powerful method called *Evaluator Weighted Estimator* (EWE) was proposed which assigns different weights to annotators based on their reliability. However, it has the limitation of not quantifying the bias and tends to fail when all annotators have similar bias trends. This motivates us to estimate the statistical parameters of the annotator-bias to get obtain better estimates of 'true' codes.

In the following sections, we analyze different aspects of representing the codes in a lower dimensional space without significant loss of accuracy. Moreover, we propose a novel and iterative method to estimate 'true' codes from the noisy codes from multiple annotators.

2. HYPOTHESES AND MODEL DESCRIPTION

2.1. Hypotheses

The work presented in this paper is driven by the following hypotheses about the estimation and reduced representation of behavioral codes and the biases associated with the codes.

This work was supported by the National Science Foundation (NSF).

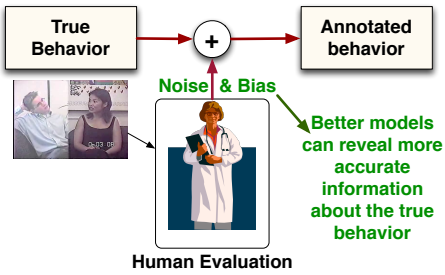

Fig. 1. Human annotation is noisy and subjective. Understanding this process can lead to better estimation of real behavioral states and conditions.

- **H1**: We can get better estimates of the 'true' codes from the noisy behavioral ratings provided by the annotators by identifying the noise-model (akin to bias and reliability) of each annotator.
- **H2**: Behavioral codes can be represented using reduced number of variables using a transformation of the original codes to a lower-dimensional space. Since human annotation is noisy, reducing the dimensionality of the behavioral space can potentially reduce the error introduced by the human-annotation process.
- **H3**: In addition to the behavioral codes, the noise introduced by the annotators in the higher dimensional space is also decorrelated in the lower dimensional space if the same transform is used as in *H2*.

2.2. Model Description

Fig. 1 depicts the human annotation process, which is inherently noisy. Behavioral scales are made by humans for use by humans, and these scales represent average human opinion, however each specific individual is different; someone may have a more negative attitude and can focus on negative aspects of the interaction and provide more negative ratings while a more positive person can have a positive bias. To reduce error in annotated behaviors multiple annotators are often asked to rate each therapy session. Further for each study multiple behaviors are of interest.

Validating H1: We represent this process in Fig.2 on the right: a noisy multiple-input multiple-output framework where the assessment of every behavioral dimension can be treated as a signal transmitted through a noisy channel. Each output block represents the assessment of 'true' codes $\mathbf{x}$ by an annotator k producing his own assessment $\mathbf{y}^{(k)}$ that incorporates his own biases $\mathbf{n}^{(k)}$.

Validating H2: Further, behavioral annotation manuals include a range of behaviors relevant to the study-domain, that are often highly inter-dependent. For instance the couples therapy annotation manuals [15, 16] define behavior codes such as "negative," "positive," "blame," that are very related. We want to investigate the possibility that a lower-dimensional behavioral codeset exists that can minimize annotation cost and reduce the noise-floor of the annotation process. This process is shown in the left block of Fig. 2.

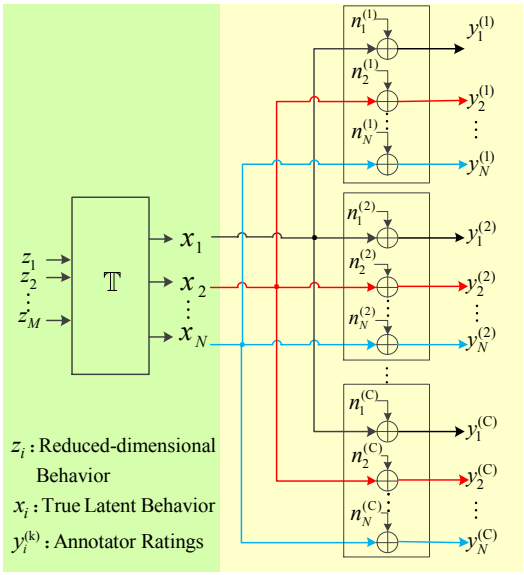

Fig. 2. Model of transmission of behavioral codes analogous to a multi-input multi-output system

Validating H3: After finding the transformation $\mathbb{T}$ we can employ that to transform the individual annotator noise models and investigate how such transformation applies to the individual annotators' internal model.

3. METHODOLOGY

Let us define:

- 'true' codes in N-dimensions as defined by annotation manual, $\mathbf{x} = [x_1, x_2, ..., x_N]^{\mathrm{T}}$
- estimate of 'true' codes in reduced M-dimensions, $\mathbf{z} = [z_1, z_2, ..., z_M]^{\mathrm{T}}$
- bias of k-th annotator, $\mathbf{n}^{(k)} = [n_1^{(k)}, n_2^{(k)}, ..., n_N^{(k)}]^{\mathrm{T}}$,
- 'noisy' codes for each session $\mathbf{y}^{(k)} = [y_1^{(k)}, y_2^{(k)}, ..., y_N^{(k)}]^{\mathrm{T}}$, where the annotator is denoted by $k = 1, 2, 3, ..., C$, where C is the number of annotators for a specific session.

Noise model. We model the biases as additive Gaussian noise that can have non-zero mean. For example, an annotator having positive mean of bias for 'satisfaction' indicates that on an average he or she is more likely to rate 'satisfaction' with a value higher than the actual ones. Moreover, the uncertainty of the bias for every annotator for every code can be modeled as the variance of the bias variable associated with that annotator for that code. According to this model,

$$\mathbf{y}^{(k)} = \mathbf{x} + \mathbf{n}^{(k)} \tag{1}$$

Now based on our Gaussian assumption of bias variables,

$$n_i^{(k)} \sim \mathcal{N}(\mu_i^{(k)}, (\sigma_i^{(k)})^2) \tag{2}$$

where the code index is denoted by $i = 1, 2, 3, ..., N$ and the annotator by $k = 1, 2, 3, ..., C_T$, where C_T is the total number of annotators for the entire dataset.

Mean bias. Behavioral manuals represent the human collective experience: for instance a neutral emotion is what a

very large number of people would consider a neutral emotion. As such, had we obtained a very large number of annotations, we could assume that the average bias would be zero. In this case, we will loosen that assumption to employ all of the annotators of the entire dataset for all codes, *i.e.,*

$$\frac{1}{N}\sum_{j=1}^{C_T}\mu_i^{(j)} \approx \mathbb{E}(\mu_i) = 0 \quad \forall i \in \{1,2,3,...,N\} \tag{3}$$

Transformation of 'true' codes. The transformation $\mathbb{T}$ from $\mathbf{z}$ to $\mathbf{x}$ (from a lower dimensional efficient space to a higher dimensional, over-generated space) is shown in Fig. 2 (left). For ease and tractability in this work we approximate $\mathbb{T}$ by a linear and time-invariant process. So we can represent $\mathbb{T}$ as a transformation matrix $\mathbf{T}$ of size $N \times M$. Therefore,

$$\mathbf{x} = \mathbb{T}(\mathbf{z}) = \mathbf{T}\mathbf{z} \tag{4}$$

Since $\mathbf{z}$, $\mathbf{x}$ are unknown, we first estimate $\mathbf{x}$ from $\mathbf{y}^{(\mathrm{k})}$ and then $\mathbf{z}$ and $\mathbb{T}$ from $\mathbf{x}$. The optimal M (size of $\mathbf{z}$) is also not known and as such we are providing later an analysis of the information loss through this lower-dimensional representation.

3.1. Estimation of 'True' Codes from Noisy Codes

In this section, we discuss how we estimate the 'true' codes from the codes given by various annotators. Our technique is a modification of the Expectation-Maximization (EM) algorithm where in each step we find the noise model of each annotator, then based on this noise model we estimate the ground truth and iterate. We initialize ground truth as the average of $\mathbf{y}^{(\mathrm{k})}$'s. We implicitly utilize (3) as a constraint in our method, which follows:

- Step 1: For every session, initialize $\mathbf{x}(0) = \frac{1}{C}\sum_{k=1}^{C}\mathbf{y}^{(k)}$
- Step 2: For iteration $(t+1)$, use the previously estimated value of $\mathbf{x}(t)$ to get a realization of $\mathbf{n}^{(k)}(t+1)$ for every annotator k as follows:

$$\mathbf{n}^{(\mathrm{k})}(t+1) = \mathbf{y}^{(\mathrm{k})} - \mathbf{x}(t) \quad \forall k \tag{5}$$

- Step 3: Estimate the noise parameters for every annotator using different sessions:
  (i) Find the sample mean (Maximum Likelihood estimate) of biases for each annotator ($\forall k$) using all of L_k instances annotated by him and re-estimate to satisfy collective zero-mean assumption in (3):

$$\widehat{\mu^{(\mathrm{k})}(t+1)} = \frac{1}{L_k}\sum_{l=1}^{L_k}\mathbf{n}^{(\mathrm{k}),(\mathrm{l})}(t+1) \tag{6}$$

$$\mu^{(\mathrm{k})}(t+1) = \widehat{\mu^{(\mathrm{k})}(t+1)} - \frac{1}{C_T}\sum_{j=1}^{C_T}\widehat{\mu^{(\mathrm{j})}(t+1)}$$

  (ii) Compute the sample variance (Maximum Likelihood estimate) of biases $\forall i,k$:

$$[\sigma_i^{(\mathrm{k})}(t+1)]^2 = \frac{1}{L_k - 1}\sum_{l=1}^{L_k}[n_i^{(\mathrm{k}),(\mathrm{l})}(t+1) - \mu_i^{(\mathrm{k})}(t+1)]^2 \tag{7}$$

In (6) and (7) the l in $n_i^{(k),(l)}$ indicates realizations of the same random vector $\mathbf{n}^{(\mathrm{k})}$ in different sessions.

- Step 4: Get new estimates for true codes:

$$\mathbf{x}(t+1) = \frac{1}{C}\sum_{k=1}^{C}(\mathbf{y}^{(\mathrm{k})} - \mu^{(\mathrm{k})}(t+1)) \quad \forall k \tag{8}$$

- Step 5: Go to step 2 unless convergence criterion is met. Due to space limitations the derivation and convergence analysis are omitted.

3.2. Representation of Codes in Lower-dimensional Space

We used the well-known *Principal Component Analysis* (PCA) technique to project the estimated 'true' codes $\mathbf{x}$ onto lower dimensional space.

1. PCA computes a matrix $\mathbf{H}$ such that it transforms $\mathbf{x}$ into $\mathbf{w}$ as $\mathbf{w} = \mathbf{H}\mathbf{x}$.
2. Keep the first M (also, most informative) elements of $\mathbf{w}$ to get the $\mathbf{z}$ and discard the remaining $(N-M)$ (least informative) elements.

Alternatively, we could use a matrix $\tilde{\mathbf{H}}$ by keeping only first M rows of $\mathbf{H}$ and obtain $\mathbf{z}$ from $\mathbf{x}$ directly as $\mathbf{z} = \tilde{\mathbf{H}}\mathbf{x}$. The pseudoinverse of $\tilde{\mathbf{H}}$ is basically the transformation matrix $\mathbf{T}$ introduced in (4), $\mathbf{T} = (\tilde{\mathbf{H}}^{\mathrm{T}}\,\tilde{\mathbf{H}})^{-1}\,\tilde{\mathbf{H}}^{\mathrm{T}}$.

3.3. Transformation of Annotator Noise Model

Having an estimate of the transformation as in Sec. 3.2 from the original behavioral coding space $\mathbf{x}$ described in Sec. 3.1 we can attempt to estimate whether this transformation is also applicable to the noise introduced by the annotators. For every $\mathbf{n}^{(k)} \sim \mathcal{N}(\mu^{(\mathrm{k})}, \mathbf{K}_{\mathbf{n}}^{(\mathrm{k})})$ we obtain $\mu^{(\mathrm{k})}$ and the covariance matrix $\mathbf{K}_{\mathbf{n}}^{(\mathrm{k})}$. Then we project $\mathbf{n}^{(\mathrm{k})}$ using $\tilde{\mathbf{H}}$ to obtain the noise model of the annotators in the reduced-dimensionality behavioral space in terms of their resulting bias mean and covariance matrix.

4. DATASET USED AND EXPERIMENTAL SETUP

For our experiments we used the Couples Therapy corpus ratings [17] defined by the *Couples Interaction Rating System* (CIRS) [16, 15]. It has 13 different codes and every code is rated on a scale from 1 to 9. The database contains a total of 5367 annotations that correspond to 1538 unique sessions from 168 unique couples rated by 17 trained annotators. Ratings include behavioral assessment of husband, wife, or both along 13 behavioral dimensions. Each session is rated by a subset of the annotators ranging from 2 to 9.

5. RESULTS

5.1. Effect on Inter-annotator Agreement

We compare our method with two other commonly used approaches: (i) estimating $\mathbf{x}$ as $\overline{\mathbf{y}^{(\mathrm{k})}}$ (average of annotator ratings) (ii) the *Evaluator Weighted Estimator* (EWE) method [13] which takes a weighted average of annotator ratings based on their reliability.

Due to lack of the real true codes for comparison, we use Krippendorff's α [18] for inter-annotator agreement, where higher value of α means more agreement. We replace the rat-

Table 1. Inter-annotator agreement using Krippendorffs α: (1) Original annotator; (2) Replacing one annotator with our method's estimate of the ground truth, or (3) with mean opinion, or (4) with the EWE estimate of the true code.

Data used	(1) Human coding	(2) Annotator→proposed code estimate	(3) Annotator→mean estimate	(4) Annotator→EWE code estimate
mean α	0.7562	0.9007	0.8661	0.8434
std. dev.	0.1116	0.1087	0.0956	0.1161

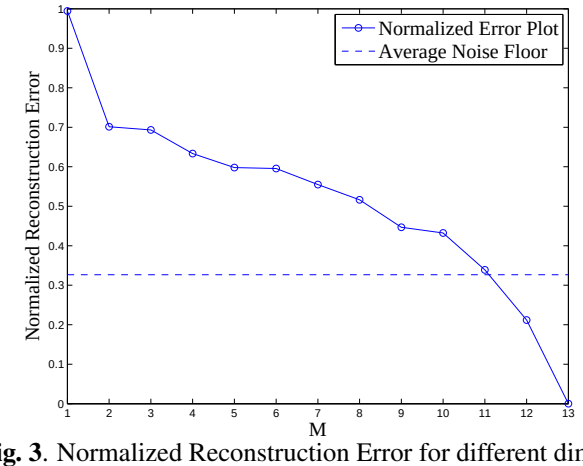

Fig. 3. Normalized Reconstruction Error for different dimensions (M) of $\mathbf{z}$ and average noise due to annotator biases

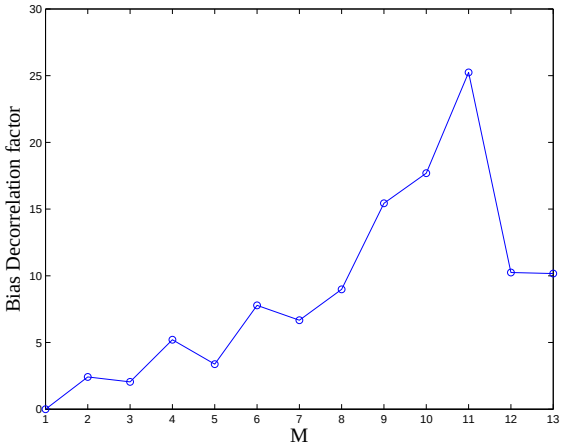

Fig. 4. Variation of Bias Decorrelation Factor of with different dimensions (M) of $\mathbf{z}$ for the noise model biases

ings of one annotator by the estimated 'true' codes and repeat this for all annotators one at time, and report the mean α and the standard deviation for each method. We also compute α for the original human ratis without any estimation. As we see in Table 1, our method performs the best, validating H1.

5.2. Analysis of Lower Dimensional Representation

Given $N = 13$ in our dataset, we have experimented with different values of M. Fig. 3 shows how the normalized reconstruction error (the ratio of the l_2-norm of errors to that of the original codes) varies with M when we reconstruct the higher dimensional codes from $\mathbf{z}$. We also show the normalized 'noise-floor' (the ratio of the l_2-norm of mean biases to that of the original codes) in higher dimensional space which indicates the level of uncertainty in $\mathbf{y}^{(k)}$ with respect to $\mathbf{x}$. So, having a reconstruction error less than the 'noise-floor' is not very meaningful. This justifies H2, the need for consideration of a lower dimensional representation.

5.3. Annotator bias in Reduced Space

To validate our H3, we experimentally show that bias variables also get decorrelated across dimensions in the projected space. We define a *Bias Decorrelation Factor(BDF)* as follows:

$$\text{BDF} = \frac{\text{avg. cross-corr. co-eff. of biases in } \mathbf{x}\text{-space}}{\text{avg. cross-corr. co-eff. of biases in } \mathbf{z}\text{-space}} = \frac{\overline{\rho_{ij}^{\mathbf{x}}}}{\overline{\rho_{ij}^{\mathbf{z}}}}$$

where $i \neq j$ and we consider absolute values of correlation coefficients while taking the average. A high value of BDF indicates a high degree of decorrelation. As we can see in Fig. 4, we get BDF> 1 for all values which signifies the decorrelation of the biases across dimensions in the projected space. Moreover, we get higher values of BDF for $M < 13$, specifically $M = 11$ gives the most decorrelated representation of biases.

6. CONCLUSIONS

We presented a noisy signal-based framework of behavioral codes which we can utilize to obtain reliable estimates of 'true codes' or ground truth using the evaluations of multiple annotators. The proposed method has been shown to increase the inter-annotator agreement compared to other approaches as it utilizes the statistics of the biases of all annotators. We also analyzed the relationship between the reconstruction error and different reduced-size representations and compared it to the noise level arising from the annotator biases. Moreover we observe that annotator biases across different codes are decorrelated when represented in lower dimensional space. Thus we conclude that lower dimensional representations can be useful for minimizing the annotation cost and diluting the impact of annotator biases while retaining most of the information of the original codes. Although not used in this work we intend to analyze reliability information of the various annotators as denoted by the variance (7) in future work. We also intend to employ datasets of higher behavioral dimensionality to show the need of joint annotation-manual definition, annotation code redundancy evaluation, and annotator reliability analysis.

7. REFERENCES

[1] Shrikanth Narayanan and Panayiotis G Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, vol. 101, no. 5, pp. 1203, 2013.

[2] Brian R. Baucom and Esti Iturralde, "A behaviorist manifesto for the 21st century," pp. 1–4, Dec 2012.

[3] Matthew P Black, Athanasios Katsamanis, Brian R Baucom, Chi-Chun Lee, Adam C Lammert, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Communication*, vol. 55, no. 1, pp. 1–21, 2013.

[4] Matthew Black, Athanasios Katsamanis, Chi-Chun Lee, Adam Lammert, Brian Baucom, Andrew Christensen, Panayiotis Georgiou, and Shrikanth S. Narayanan, "Automatic classification of married couples' behavior using audio features," in *In Proceedings of InterSpeech*, Makuhari, Japan, Sept. 2010.

[5] Theodora Chaspari, Daniel Bone, James Gibson, Chi-Chun Lee, and Shrikanth S. Narayanan, "Using physiology and language cues for modeling verbal response latencies of children with asd," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013.

[6] Bo Xiao, Dogan Can, Panayiotis Georgiou, David Atkins, and Shrikanth S. Narayanan, "Analyzing the language of therapist empathy in motivational interview based psychotherapy," in *Proceedings of APSIPA Annual Summit and Conference*, Dec. 2012.

[7] Rahul Gupta, Panayiotis Georgiou, David Atkins, and Shrikanth Narayanan, "Predicting clients inclination towards target behavior change in motivational interviewing and investigating the role of laughter," in *Proceedings of Interspeech*, Sept. 2014.

[8] Nora A Murphy, "Using thin slices for behavioral coding," *Journal of Nonverbal Behavior*, vol. 29, no. 4, pp. 235–246, 2005.

[9] Roger Bakeman, "Behavioral observation and coding," *Handbook of research methods in social and personality psychology. Cambridge University Press, New York*, pp. 138–159, 2000.

[10] Kartik Audhkhasi and Shrikanth Narayanan, "A globally-variant locally-constant model for fusion of labels from multiple diverse experts without using reference labels.," 2013, pp. 769–783.

[11] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie, "The multidimensional wisdom of crowds," in *Advances in neural information processing systems*, 2010, pp. 2424–2432.

[12] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy, "Learning from crowds," *The Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[13] Michael Grimm and Kristian Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*. IEEE, 2005, pp. 381–385.

[14] Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10, pp. 787–800, 2007.

[15] J Jones and A Christensen, "Couples interaction study: Social support interaction rating system," *University of California, Los Angeles*, 1998.

[16] C Heavey, D Gill, and A Christensen, "Couples interaction rating system 2 (cirs2)," *University of California, Los Angeles*, 2002.

[17] Andrew Christensen, David C Atkins, Sara Berns, Jennifer Wheeler, Donald H Baucom, and Lorelei E Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples.," *Journal of consulting and clinical psychology*, vol. 72, no. 2, pp. 176, 2004.

[18] Klaus Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970.