2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) | 978-1-6654-1298-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISMAR-Adjunct54149.2021.00039

# Motion and Meaning: Sample-Level Nonlinear Analyses of Virtual Reality Tracking Data

Mark Roman Miller
Department of Computer Science
Stanford University
Stanford, California, USA

Hanseul Jun
Department of Communication
Stanford University
Stanford, California, USA

Jeremy N. Bailenson
Department of Communication
Stanford University
Stanford, California, USA

**ABSTRACT**

Behavioral data is the "gold standard" for experiments in psychology. The tracking component of virtual reality systems captures data on nonverbal behavior both covertly and continuously at high spatial and temporal fidelity, enabling what is called behavioral tracing. With previous research analyzing this type of data, however, inference has primarily been limited to linear relationships of subject-level aggregates. In this work, we suggest these rough aggregations are often neither the best according to theory nor do they make use of the rich data available from behaviorally traced experiments. We also explore the relationships between motion and subjective experiences with a previously published dataset of 360-degree video and emotion, and we find evidence for nonlinear sample-level relationships. In particular, reported valence relates with head pitch and pitch velocity, among others, and reported arousal relates with head rotation speed and yaw velocity, among others. The role of these sample-level nonlinear relationships for future work are discussed.

**Keywords**: virtual reality, behavior tracing, methodology, nonlinear analyses.

**Index Terms**: • Human-centered computing - Human computer interaction (HCI) - Interaction paradigms - Virtual reality • General and reference - Cross-computing tools and techniques - Measurement • Applied computing - Law, social and behavioral sciences - Psychology

## 1 INTRODUCTION

One of the promises of VR in psychology research is the ability to fully control the sensory stimuli a participant receives and capture all actions a participant makes while in an experiment. Both of these threads have their origins in Sutherland's initial vision of virtual reality as the "ultimate display" [1] and were applied to research in social psychology [2, 3, 4]. Using the same tracking devices used to interact and display a virtual world, virtual reality systems collect a user's physical behavior – commonly, position (X, Y, Z) and rotation (yaw, pitch, roll) of the head and hands – in fine spatial and temporal resolution during experiments.

The collection and analysis of this data has recently been dubbed *behavioral tracing*, defined by Yaremych and Persky [5] as "fine-grained, nearly continuous measurement of physical behavior." Behavioral tracing is of interest to psychologists because of its focus on behavior, considered the 'gold standard' for experimental measures, its covert and continuous collection.



However, a *behavior trace*, a set of spatial measurements taken over the course of an experiment, must be linked to some other variable of interest, often an experimental condition or some questionnaire result. Therefore, there must be some mathematical transformation from the many behavioral tracing values to one (or more) summary variables.

In this work, we provide a preliminary exploration of nonlinear methods and discuss their benefits in a behavioral tracing paradigm. First, we describe the theoretical issues underpinning linear relationships on behavior tracing data (Section 2). We use these insights to describe and justify a potential approach (Section 3). Then, we apply this approach to a previously collected dataset of motion collected during observation of 360-degree video in order to demonstrate its value (Sections 4 and 5). Using these results, we discuss findings and implications for future work (Sections 6 and 7).

## 2 PREVIOUS WORK

First, we position this work relative to a growing area of research using machine learning on what could be described as behavior traces to predict some construct, e.g., cybersickness [6, 7] or mental workload [8]. Future work on interpretable machine learning may provide opportunities in this space.

Returning to behavior tracing, we review some threads of work and their approaches to aggregation methods. Proxemics, the study of interpersonal distance, has been studied extensively in virtual reality using linear methods [9, 10, 11]. We note that these linear results are contradictory at first glance, as some studies show greater interpersonal distance correlating with higher social presence [9, 10], while some show an effect in the opposite direction [11, 12]. We do not argue these are in contradiction, but rather that there is a nonlinear relationship between social presence and interpersonal distance, in an analogue to the "uncanny valley." Overall, this highlights the importance of a nonlinear model to fit the relationship between the two values.

Another approach we mention is work by Yaremych and collaborators [13] that studied parent's behavior choosing lunch for their child while at a VR buffet. Parents' path tortuosity, the degree of indirectness for a path, correlated with a drop in guilt between pre- and post-experiment questionnaires. They suggested path tortuosity captured an aspect of the cognitive effort the parents undertook while putting together their child's plate of food. This work's unique usage of a nonlinear measure enabled by behavior tracing, and more importantly, its framing of the use of path tortuosity as a contribution of the paper, places it as a great example of a nonlinear aggregation method.

The work in the previous literature that we believe to be the best example of sample-level nonlinear analyses is McCall and Singer's work [14] of "proxemic imaging". In a study of fair and unfair virtual players of an economic game, the researchers recorded a participant's distance and gaze relative to the virtual agent. They found that participants kept closer to the fair player compared to the unfair player, and while participants stood directly in front of unfair players more often than they study in front of the fair player. These

978-1-6654-1298-8/21/$31.00 ©2021 IEEE
DOI 10.1109/ISMAR-Adjunct54149.2021.00039

nonverbal behaviors then predicted future behavior (specifically, monetary punishment) towards the virtual human.

In our work, we follow a similar structure to McCall and Singer [14] for binning, but instead of testing the significance of clusters of points between two study conditions, we predict a construct of interest along a reported scale with significance given by a permutation test. Also, our work examines a large sample (over 500 participants) looking at 80 separate pieces of content.

## 3 Sample-Level Nonlinear Analysis

We propose an alternative way to relate behavioral traces to constructs of interest that to our knowledge has not been applied in relation to VR: a sample-level piecewise constant model.

First, we motivate our decision for the type of model (3.1). Then, we describe the model fitting process (3.2), the null hypothesis and the construction of the null distribution (3.3), and finally an instantiation of a permutation test that applies this model that also takes into account dependence of samples within participants (3.4).

### 3.1 Motivation

One description of a classical statistical model is "surface-plus-noise [15]. Underlying any model is a "surface", a relationship describing some scientific truth that is being examined, and "noise", other effects due to unaccounted variables, incomplete knowledge, or errors in measurement that are ignored for efficiency's sake. A choice of model is, in short, a characterization of the space, the surface, and the noise.

The common characterization is Gaussian noise and a linear surface. This characterization can be adequate for most work, but we argue that the rich data afforded by behavioral tracing enables models with more expressive characterizations. In the interest of this expressiveness, the model we use is a step function mapping the motion data to the construct of interest.

### 3.2 Fitting the Model

The goal of the fitting process is to approximate the relationship between a continuous, sample-level variable (e.g., head pitch) and the corresponding session-level subjective rating (e.g., valence). The sample-level variable in session number $s$ of $S$ and index $t$ of $T$ samples per session is denoted by $x_{st}$, and the session-level rating of session $s$ is denoted by $r_s$.

To begin, all samples are sorted into bins of length $w$ by rounding. The bin $b_{st}$ associated with sample $x_{st}$ is

$$b_{st} = w \left\lfloor \frac{x_{st}}{w} + \frac{1}{2} \right\rfloor$$

All samples in a bin are matched with their session-level ratings, then averaged. More explicitly,

$$\hat{f}(b) = \frac{\sum_{s,t} r_s \cdot 1[b_{st} = b]}{\sum_{s,t} 1[b_{st} = b]}$$

Note that each bin can have samples from many participants and each participant can (and often does) have multiple samples in a bin.

Treating all the samples in the same way regardless of participant or session allows the function $\hat{f}(b)$ to be interpreted in a simple way: given the motion value $x_{st}$ (e.g., pitch) of a single randomly chosen sample from all sessions in the experiment, what is the expected value for the subjective rating $r_s$ (e.g., valence) for the session the sample was taken from? Naturally, the best estimate is the average of all subjective ratings with the given motion value.

This mean is performed over the construct values, which are often collected as Likert-scale data. There has been some controversy on whether to treat Likert-scale data as ordinal or interval data, and we refer the reader to discussion in [16] for grounding for this step.

To predict a value for an entire session, rather than a sample, we define the predicted value of a session $\hat{f}(s)$ to be

$$\hat{f}(s) = \frac{1}{T} \sum_{t=1}^{T} \hat{f}(b_{st})$$

i.e., the mean of the predicted samples within the session.

These fitted values are not independent of each other, as they share samples from the same session, and the averages themselves are not composed of some number of independent equally weighted parts, as many samples within a bin come from the same session. These dependencies can complicate null hypothesis testing, discussed below.

### 3.3 The Null Hypothesis

In order to visually and statistically compare against some null hypothesis, it is necessary to specify the null hypothesis and develop a way to calculate a null distribution. We use a Monte Carlo method to estimate the null distribution by running the same curve fitting process described above (binning and averaging the rating) upon data where the session ratings $r_s$ have been shuffled. This maintains the distribution of the rating variable as well as the correlations among the samples from the same session.

We note that this does break the correlations among sessions using the same stimulus or from the same participant. These correlations can be preserved if all ratings are shuffled within each stimulus or within each participant, allowing the null hypothesis to also account for stimulus or participant respectively.

### 3.4 Significance

The statistical test we propose is in essence a permutation test. The essential format of a null-hypothesis test is to measure some quality of interest upon the empirical data and compare its value to a distribution of values drawn under the null hypothesis.

In our case, our quantity of interest is the predictive strength of how much one half of the data is upon the other half of data. First, the data is grouped by session and split into two halves, ensuring all samples from one session remain together. The curve-fitting process is applied to one half of the data, and the fitted function is used to predict the rating in the second half of the data. This prediction is correlated with the empirical ratings, and the correlation's t-value represents the strength of the relationship. This split is made many times – in our results, it is performed 30 times – and all t-values are averaged. The result is one value estimating the predictability of one half of the data upon the other.

The second part of any null hypothesis significance test is the comparison against a null distribution. The process for creating a null distribution is the same as detailed in section 3.3 on shuffling. Once this simulation of data is created, its predictability is estimated using the split-half process above. This reshuffling of session-level ratings is done many times to approximate the null distribution. Finally, the empirical predictability is compared against the predictability values within the null distribution, producing our reported p-value.

## 4 Case Study

We have performed sample-level nonlinear analyses on data previously collected in by Jun and collaborators [17]. Because the methods are described in detail elsewhere, it suffices to be brief about the materials and procedure.

### 4.1 Materials and Procedure

In the previously collected dataset we obtained, a total of 511 participants were collected from two locations: The Tech, a technology museum in San Jose, USA, and the Stanford University campus.

Participants began the study by completing a demographics questionnaire, and then watched five 360-degree videos, each twenty seconds long, randomly selected from a pool of eighty clips. After each video, the participant rated their valence, arousal, presence, simulator sickness, and liking.

The content was displayed using the HTC Vive wired headset, and body motion was tracked using the HTC Vive headset and hand controllers.

### 4.2 Data

The data relevant to the present work fall under two headings that we refer to as *session-level* and *sample-level* data.

Session-level data refers to data uniquely identified by a combination of participant and stimulus. Each stimulus was a twenty-second 360-degree video clip. Session data consists of nine-point Likert scale ratings of valence and arousal, as well as five-point Likert scale ratings on simulator sickness, liking, and presence.

Sample-level data refers to spatial data collected at about 90Hz that includes position (X, Y, Z) and rotation (yaw, pitch, and roll). The conventions follow defaults in the game engine Unity. Position is based upon the the left-handed coordinate system that +Y is up, +Z is forward, and +X is right. Rotation is based upon the intrinsic rotations about the +Y axis (yaw), then +X axis (pitch), and then +Z axis (roll), equivalent to the extrinsic rotations in the opposite order. All rotations follow a left-handed convention.

In addition to the raw data, there are seven additional data streams computed from the raw data. Six streams are first-order differences over time as a finite-sample approximations of velocity. The seventh stream, rotation speed, is the angle between the two consecutive points created by the intersection of the unit sphere and the ray along the direction the participant's head is facing.

### 4.3 Analysis

The seven differential measures (i.e., the six velocities and rotation speed) were produced by taking the first-order difference, then the absolute value, and finally the natural logarithm. The log-transform was performed because the probability density distribution sharply peaked around zero. This process is validated by the fact that values are spread relatively evenly (differing only by 4x) across a huge range of magnitudes (100x between largest and smallest).

The sample-level nonlinear models were fitted using the method described in section 3. The number of bins ranged from 40-100 and was chosen depending on the range and density. We have found the ideal width is the smallest where the "jitter" between bins is just beginning to be visible. This signals the resolution is as fine as possible.

### 4.4 Hypotheses

Because of the exploratory nature of this method, we are comparing all reported subjective, session-level values (arousal, simulator sickness, valence, liking, and presence) against all thirteen measures of motion. We report p-values with Bonferroni-Holm correction made across the 65 comparisons.

## 5 Results

Of the 65 comparisons made, 36 relationships were more predictive in the split-half test than would be expected by chance, and 23 of these relationships are significant while accounting for the multiple comparisons. Due to space concerns, plots will only be shown for selected relationships, but all plots will be available in the supplemental material. Six plots of motion are displayed in Figure 1.

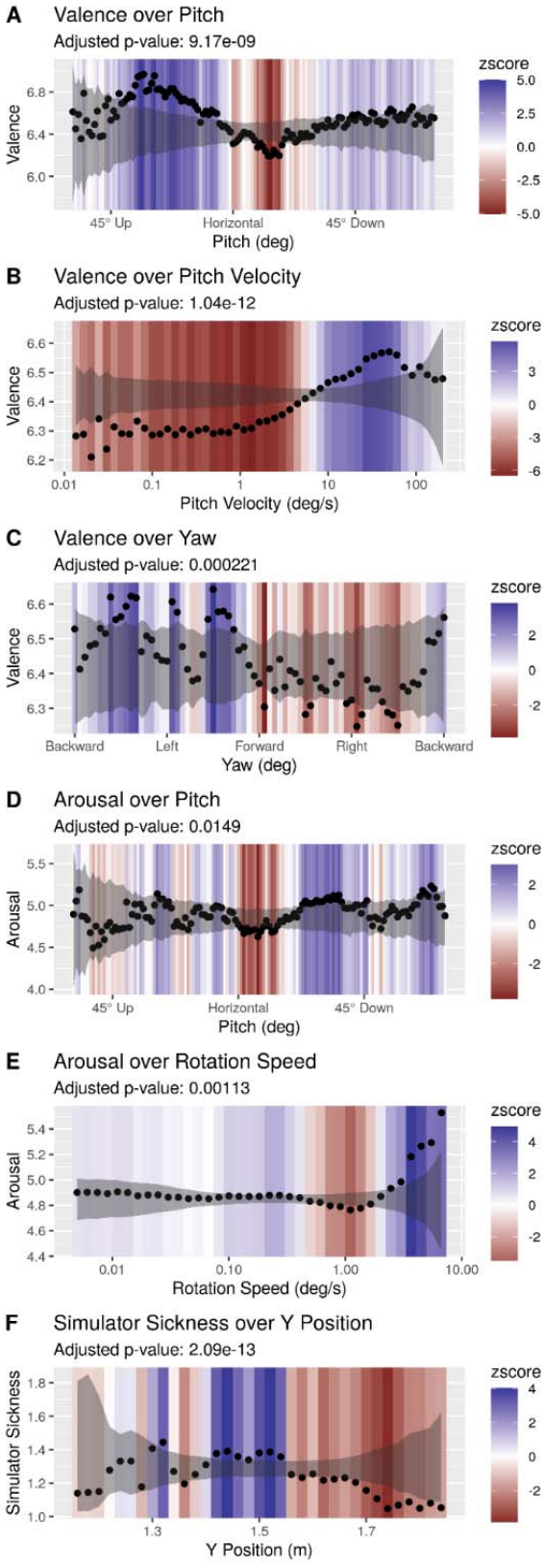


Figure 1. Six plots relating motion to a construct of interest.

In Figure 1, the x-axis is the motion measure, and along it are the 'buckets' in which samples are sorted. For each bucket (column), the black point represents the fitted value, e.g., the mean rating of all samples within the bucket. The transparent ribbon represents the interval of 95% density of values fitted under the null distribution. The color of each column represents the z-score of the empirically fitted value against the null distribution, e.g., with a z-score of 2.6, the fitted value was 2.6 standard deviations above the mean of the null distribution. This color provides visual information of the certainty in the signal, which visually highlights sections with known deviations, even if the deviations are small.

The limits of the x-axis are set such that all buckets displayed have data from at least 25 different participants and at least 1800 samples total (20s of data). This allows the effects not to be dwarfed by large variance in buckets with few samples.

Because of the large number of significant relationships, they will be grouped into categories according to a post-hoc interpretation of the results. The first set of findings relate motion to valence primarily, with a similar but secondary relationship to liking (5.1). The second set of findings relate motion to arousal (5.2). The final set of findings are relationships we attribute the participant as a confounding factor (5.3).

### 5.1 Motion and Valence

There are a total of eight relationships we collect under the heading of valence, which we break down into four subsections. The first is the relationship between valence and pitch, the second collects all relationships between valence and any measure of velocity, the third is two relationships with liking and motion that mimic their respective relationships with valence, and the fourth is a relationship between valence and yaw.

In the plot with pitch (Figure 1, panel A) there are three regions that show a deviation from the null hypothesis. The first is a positive deviation from about 40° above the horizon to 15° above the horizon, meaning that in sessions where participants spent relatively more time facing 15°-40° above the horizon, they rated the video with higher valence. The second is a negative deviation 10° to 15° below the horizon. Finally, the range from about 30° to 70° below the horizon has a small but consistent positive deviation. When ratings were shuffled within videos and within participants, the relationship remained ( $p = 0.02$ and $p = 1.71 \times 10^{-5}$ respectively), i.e., the effect was not merely due to videos or due to participants.

The four relationships with velocity (pitch velocity, roll velocity, Y velocity, and Z velocity) are grouped together because of their similar pattern: lower values in a lower range, higher values in a slightly higher range, and weakly positive values at the highest ranges. Pitch velocity is displayed in Figure 1 Panel B. The crossover point is different among the different segments, but they all show that visible motions are linked with a higher rated valence, while very steady position maintenance indicates lower valence. However, for each measure of velocity this effect flattens out beyond a certain point.

The two relationships with liking follow very similar patterns to the respective valence ratings. Both relationships to pitch have a sharp trough surrounded by higher rounded curves. Liking and pitch velocity shows a relationship similar to valence in that the same speed is the crossover point between a positive and a negative deviation. In addition, valence and liking correlate with a value of $r^2 = 0.59$, which is the second-highest between any two of the five questionnaire values. The only pair larger than this of the other 19 pairs of variables is the pair of liking and presence. Therefore, we interpret this data to suggest that liking is showing this effect due to its connection to valence.

Finally, there is a relationship between valence and yaw (Figure 1, Panel C). The plot has two visual features of note. First, overall, sessions with more samples on the left side also tend to have higher valence. However, when the null hypothesis is modified to include variation due to stimulus, this result no longer holds. Second, the curve is not as smooth as others we have investigated so far. We suggest this result is because yaw varies relatively more across videos than within a video.

### 5.2 Motion and Arousal

The three types of motion that are correlated with arousal are pitch, rotation speed, and yaw velocity. We pair up the rotation speed and yaw velocity because they are both types of motion and have similar curves.

The relationship between rotation speed and arousal Figure 1 panel E is as follows. At the lowest ends of rotational motion, when the participant's effectively stationary, there is no signal, or slightly positive. In the range of from about 0.4°/s to 2°/s, the deviation is negative, but at speeds beyond that, the deviation is positive. One interpretation of this result is the difference between 'wandering' in which the user explores the space slowly and calmly, and 'tracking' in which the user explores the space with focus on one salient object.

The relationship between arousal and pitch is displayed in Figure 1 panel D. Similar to the relationship between pitch and valence, there is a drop in arousal in the range just below horizontal. In contrast to valence, though, high arousal in this dataset is associated with looking downward 20°-40° degrees. In contrast to the relationship between pitch and valence, however, once the video is accounted for, the relationship is no longer significant. We interpret this noting that some videos in the dataset were set high on a balcony or tower and consider that looking down may be more common in these situations.

### 5.3 Relationships Mediated by Participant

Finally, there are twelve relationships that we have judged to be mediated by participant. To explain how the model captures this, let us consider an example of x-position and liking. The span of x-position values varied more across participants than within participants, i.e., participants did not move much horizontally. Therefore, a certain bucket of x-position would be dominated by a few participants. Furthermore, all questionnaire results showed larger variance within each stimulus (due to participant), than within the participants (due to stimulus), even after accounting for differences in sample size.

Because the initial significance test did not account for dependence between sessions due to the same participant or the same video, it is very likely the signal that carries between the train and test set is due to participant. This is corroborated by results showing all relationships of this type did not have significant values once participants' average scores were accounted for.

For almost all of these relationships, we hesitate to interpret a connection between the motion and the construct. However, the relationship between y-position (head height) and simulator sickness stands out. Previous work has shown that men tend to report less simulator sickness than women [18]. In addition to the fact that on average men are taller than women, we interpret this relationship as the relationship between height, gender, and simulator sickness.

## 6 Discussion

The relationship results span a wide range of constructs and motion types. The primary types of nonlinear relationships discovered in this dataset were relationships between valence or arousal and head pitch or motion. There were also many relationships that were statistically significant, but lost significance when accounting for data shared across participants.

### 6.1 Findings

There were 23 relationships significant at a family-wise error rate of $\alpha = 0.05$. These values reject the null hypothesis that the predictive ability on one half of sessions to the other half of sessions was only due to chance. When tested against more reasonable null hypotheses accounting for the dependence of values from the same participant or from the same stimulus, many of these relationships lose their significance. In a post-hoc process, we group these findings together into three clusters: relationships with valence, relationships with arousal, and relationships mediated by participant.

Motion values that have some relationship with valence include pitch orientation; pitch, roll, Y, and Z velocity; and yaw orientation. We also include relationships with liking in this cluster due to liking's high correlation with valence and similarly shaped relationship with motion. The motions that relate to arousal include pitch orientation, rotation speed, and yaw velocity. Finally, there are many significant relationships that we interpret as capturing only the individual differences in this particular set of participants.

### 6.2 On Linearity

One of the goals of psychology is to discover and understand the links between observable behavior and internal state. Progress in understanding can either be made by discovering previously unknown links between construct and behavior, or more deeply understanding links already found, including the study of boundary conditions, mediators, and moderators. Another contribution that falls under the latter set is refining the quantitative values driving the link. We argue that one common impediment to this refinement is the assumption that relationships of interest are likely to be linear.

The value of richer, nonlinear models is twofold: first, it provides more details to generate hypotheses from, and second, these details can distinguish between different theories more effectively than linear models. In fact, Meehl [19] notes this lack of specificity of linear models makes theories *less* falsifiable as measures become more precise.

Furthermore, we opine that many expected relationships between motion and a construct of interest are not linear. To support this preliminary conjecture, consider the relationship between head pitch (tilting the head up and down) and affective valence (positive or negative). This effect is evidenced by common phrases like "hold your head up" and has been empirically verified using photos of winning and losing Olympic athletes [20].

In work performed by Jun et al. [17], among others, a mean over all time points in a trace is related to a construct of interest. In this example, it would be the mean of head pitch over all samples in a linear model predicting the valence the participant reported. Using the variables from section 3.2, and letting $\beta_0$ and $\beta_1$ be the y-intercept and slope respectively, this process is expressed as:

$$\hat{f}(s) = \beta_0 + \beta_1 \cdot \frac{1}{T}\sum_{t=1}^{T} x_{st}$$

We note that, by linearity of the mean, this is mathematically equivalent to first performing the linear transformation from head pitch to valence at the sample level, then averaging across samples.

$$\hat{f}(s) = \frac{1}{T}\sum_{t=1}^{T} (\beta_0 + \beta_1 \cdot x_{st})$$

If we accept the first as reasonable, then this requires us to accept this second interpretation of the pitch having implications for valence at the *sample* level as equally reasonable, and that the proper mapping from pitch to valence is perfectly linear. Taking this a step further, this would mean the valence predicted for a sample in which a participant is looking 50 degrees upward is five times the valence for a participant looking 10 degrees upward.

However, a more mundane intuition of the situation would expect the cause of such great head tilt to be a conversation with someone on a balcony or taking a look at a spider in the corner of the room. A difference of fifty degrees is not five times the effect of ten degrees.

In short, these transformations from measures to constructs are as much a part of how one interprets an experiment as the choice of independent variables, dependent variables, procedure, and stimuli, and should be treated as such.

### 6.3 Limitations and Future Work

Because of the novelty of this work, we consider some of its limitations and encourage directions for future work.

These models are not without their drawbacks. Because of the complexity of the models relative to a linear or quadratic model, there is a greater ability to fit to noise, as indicated by the participant-mediated relationships. In addition, the degrees of freedom available to the researcher in choosing the model may remain hidden if the model types are not preregistered. Furthering the statistical issues, statistical tests on nonlinear models are not well-established. Future work can integrate related statistical methods such as CANOVA [21] and cluster analysis [22].

This process applies population-level features to individual's behavior. This is only fully acceptable in the ergodic case [23], which almost certainly does not apply to these types of human behaviors. Future work can leverage the paradigm of "small data" [24] to focus on individuals and person-level models. This process also assumes, by the use of mean over all samples within a session, that all moments in time are equally important in calculating the rating. Theories of attention would question this heuristic.

In the interest of turning discussion of relationships beyond mere existence and direction and towards characterizations, it is important to stress the effect size between motion on rating is small. In this dataset, the weakest relationship between two questionnaire items, simulator sickness and presence, has an $\eta^2$ value of 0.02, while the strongest relationship between a questionnaire item and a behavioral trace, between head pitch and valence has an $\eta^2$ of 0.0138.

It is worth noting that each of these tests are merely showing that there was some sort of relationship between the motion and the construct, not that the relationship was necessarily nonlinear. While we have developed a potential comparison method, in the interest of space and due to the preliminary nature of the tests, we have not included the results of this process.

## 7 Conclusion

In this work, we propose a simple nonlinear analysis based on fixed-width piece-wise constant functions that has not been applied to virtual reality and behavior tracing. We then apply this method to a previously collected dataset to uncover previously unseen relationships.

We find several relationships between motion and constructs of interest within the dataset from Jun and collaborators [17], highlight especially head pitch and valence, rotation speed and valence, and rotation speed and arousal

Some simple recommendations from this work include visualizing sample-level data, not necessarily immediately aggregated by participant. Additionally, we encourage researchers to explore models beyond a linear model of a session-level mean, especially when either previous work or researcher's intuition indicate something potentially more appropriate.

We hope these methods will improve the efficiency and usefulness of behavioral tracing as a method and reinforce the value of virtual reality as a methodological tool for psychology.

## References

[1] I. E. Sutherland, "The ultimate display," *Proceedings of the Congress of the Internation Federation of Information Processing (IFIP),* p. 506–508, 1965.

[2] M. Slater and S. Wilbur, "A framework for immersive virtual environments (FIVE)," *Presence: Teleoperators and Virtual Environments,* vol. 6, p. 603, 1997.

[3] J. Blascovich, J. Loomis, A. C. Beall, K. R. Swinth, C. L. Hoyt and J. N. Bailenson, "Immersive Virtual Environment Technology as a Methodological Tool for Social Psychology Jim," *Psychological Inquiry,* vol. 13, p. 103–124, 2002.

[4] J. Fox, D. Arena and J. N. Bailenson, "Virtual Reality: A Survival Guide for the Social Scientist," *Journal of Media Psychology,* 2009.

[5] H. Yaremych and S. Persky, "Tracing physical behavior in virtual reality: A narrative review of applications to social psychology," *Journal of Experimental Social Psychology,* vol. 85, no. April, 2019.

[6] R. Islam, Y. Lee, M. Jaloli, I. Muhammad, D. Zhu, P. Rad, Y. Huang and J. Quarles, "Automatic Detection and Prediction of Cybersickness Severity using Deep Neural Networks from user's Physiological Signals," *Proceedings - 2020 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2020,* p. 400–411, 2020.

[7] N. Martin, N. Mathieu, N. Pallamin, M. Ragot and J. M. Diverrez, "Virtual reality sickness detection: An approach based on physiological signals and machine learning," *Proceedings - 2020 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2020,* p. 387–399, 2020.

[8] T. Luong, N. Martin, A. Raison, F. Argelaguet, J. M. Diverrez and A. Lecuyer, "Towards Real-Time Recognition of Users Mental Workload Using Integrated Physiological Sensors into a VR HMD," *Proceedings - 2020 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2020,* p. 425–437, 2020.

[9] J. N. Bailenson, J. Blascovich, A. C. Beall and J. M. Loomis, "Equilibrium Theory Revisited: Mutual Gaze and Personal Space in Virtual Environments," *Presence: Teleoperators and Virtual Environments,* vol. 10, p. 583–598, 2001.

[10] J. N. Bailenson, J. Blascovich, A. C. Beall and J. M. Loomis, "Interpersonal Distances in Virtual Environments," *Personality and Social Psychology Bulletin,* vol. 29, p. 819–833, 2003.

[11] M. Lee, G. Bruder, T. Hollerer and G. Welch, "Effects of Unaugmented Periphery and Vibrotactile Feedback on Proxemics with Virtual Humans in AR," *IEEE Transactions on Visualization and Computer Graphics,* vol. 24, p. 1525–1534, 2018.

[12] N. Norouzi, K. Kim, M. Lee, R. Schubert, A. Erickson, J. Bailenson, G. Bruder and G. Welch, "Walking your virtual dog: Analysis of awareness and proxemics with simulated support animals in augmented reality," *Proceedings - 2019 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2019,* p. 157–168, 2019.

[13] H. E. Yaremych, W. D. Kistler, N. Trivedi and S. Persky, "Path Tortuosity in Virtual Reality: A Novel Approach for Quantifying Behavioral Process in a Food Choice Context," *Cyberpsychology, Behavior, and Social Networking,* vol. 22, p. 486–493, 2019.

[14] C. McCall and T. Singer, "Facing off with unfair others: Introducing proxemic imaging as an implicit measure of approach and avoidance during social interaction," *PLoS ONE,* vol. 10, p. 1–14, 2015.

[15] B. Efron, "Prediction, Estimation, and Attribution," *Journal of the American Statistical Association,* vol. 115, p. 636–655, 2020.

[16] J. Carifio and R. Perla, "Resolving the 50-year debate around using and misusing Likert scales," *Medical Education,* vol. 42, p. 1150–1152, 2008.

[17] H. Jun, M. R. Miller, F. Herrera, B. Reeves and J. N. Bailenson, "Stimulus Sampling with 360-Videos: Examining Head Movements, Arousal, Presence, Simulator Sickness, and Preference on a Large Sample of Participants and Videos," *IEEE Transactions on Affective Computing,* p. 1–1, 2020.

[18] S. Weech, S. Kenny and M. Barnett-Cowan, "Presence and cybersickness in virtual reality are negatively related: A review," *Frontiers in Psychology,* vol. 10, p. 1–19, 2019.

[19] P. E. Meehl, "Theory-Testing in Psychology and Physics: A Methodological Paradox," *Philosophy of Science,* vol. 34, p. 103–115, 1967.

[20] J. L. Tracy and D. Matsumoto, "The spontaneous expression of pride and shame: Evidence for biologically innate nonverbal displays (Proceedings of the National Academy of Sciences of the United States of America (2008) 105, 33, (11655-11660) DOI: 10.1073/pnas.0802686105)," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 105, p. 20044, 2008.

[21] Y. Wang, Y. Li, H. Cao, M. Xiong, Y. Y. Shugart and L. Jin, "Efficient test for nonlinear dependence of two continuous variables," *BMC Bioinformatics,* vol. 16, p. 1–8, 2015.

[22] S. Hayasaka and T. E. Nichols, "Validating cluster size inference: Random field and permutation methods," *NeuroImage,* vol. 20, p. 2343–2356, 2003.

[23] P. C. M. Molenaar, "A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever," *Measurement: Interdisciplinary Research & Perspective,* vol. 2, p. 201–218, 2004.

[24] D. Estrin, "Small data, where n = me," *Communications of the ACM,* vol. 57, p. 32–34, 2014.