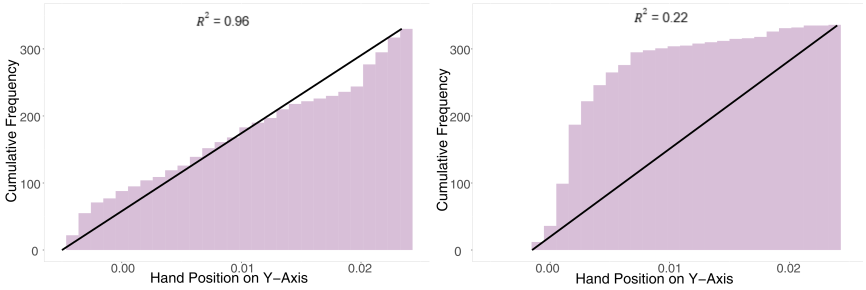
An Evaluation of Movement Data Analysis Techniques for Virtual Reality

Alice Guth[*] Davidson College
Jessica J. Good[†] Davidson College
Eugy Han[‡] University of Florida
Jeremy Bailenson[§] Stanford University
Tabitha C. Peck[¶] Davidson College



(a) Hand motion paths towards a Black avatar holding a gun.

(b) Hand motion paths towards a Black avatar holding a phone.

Figure 1: Cumulative histograms charting the y-axis hand positions of the same user in different trials of a VR experiment. The cumulative frequency where the participant visited each y-location is plotted and a linear regression is fit to the data. The R^2 coefficient of determination of each line is analyzed. a) A cumulative histogram of the y-axis hand positions of making a shooting decision towards a Black avatar holding a gun. b) A cumulative histogram of the y-axis hand positions of making a shooting decision at a Black avatar holding a phone. Both avatars were dressed in high socioeconomic clothing. The R^2 line, in black, is a measure of how well the cumulative frequencies aligned with those predicted by a linear model. The participant had more equally distributed movements when choosing to shoot at an avatar holding a gun compared to an avatar holding a phone.

Abstract

Researchers frequently collect position and orientation data from tracking hardware during virtual reality research studies. These data capture important information about participants' movements during experiments, but often result in large and complex datasets that can be challenging to analyze and interpret. We explore the potential of standard statistical measures that could be used to interpret position and orientation tracking data, and discuss advantages and uses of each measure. We further present three new measurement techniques – 95% range, time spent at mode, and coefficient of determination. We then evaluate the effectiveness of each statistical measure to detect differences in position and orientation on two existing data sets. We find that all of the tested measures are effective in discerning known main effects in position and orientation with varying degrees of sensitivity. This work is intended to guide researchers on future analysis and interpretation of motion tracking data sets.

Index Terms: Tracking, Analysis, Position, Orientation

1 Introduction

User studies are common in virtual reality (VR) research [24] and enable researchers to collect large amounts of data that can be evaluated to better understand both user behavior and VR systems. Data are often collected through multiple modalities including through subjective questionnaire responses such as presence [28] or avatar embodiment [8], measuring response latency or accuracy to cognitive tests [22], measuring proximity to other avatars [18], measuring electroencephalogram (EEG) data [9], or measuring heart rate, skin conductance, or skin temperature [20].

A frequently underutilized metric among researchers involves the collection of data from VR tracking devices, including positional coordinates (x,y,z) and orientation angles $(yaw, pitch, roll)$. These data have the potential to detect a multitude of user experiences including user stability through different gait patterns [10], interpreting gravity perception and stability through head tilt [29], investigating proxemics or gaze patterns [21], or identifying differences in user movement patterns based on self-avatar appearance [16]. Large data sets consisting of tracking data are also required for machine learning, such as for predicting eye movements [33].

Researchers can readily obtain tracking data during experiments from various sources, including the participant's head, hands, elbows, knees, or backs at frequencies around 60 Hz for the duration of the experiment. Despite the ease of acquisition, the statistical analysis of such large datasets can be formidable and complex. For example, using tracking data to determine if participants moved differently between two different experimental conditions may be challenging to interpret. Therefore, an important question is whether there are existing statistical measures that can be employed to effectively analyze and interpret large amounts of tracking data. If so, what types of research questions can each measure be used to analyze?

In this paper, we present an overview of previously used methods for analyzing tracking data, as well as introduce three new metrics, developed based on the existing measures, to fill in gaps in analysis. We discuss when and why each metric could be used to analyze motion tracking data. Finally, we use each measure on two large motion tracking data sets, one where participants were standing and one where participants were sitting. Our results demonstrate each measure's ability to detect significant differences in movement data between groups. This paper is intended to guide future researchers in analyzing underused yet informative motion tracking data sets.

[*] e-mail: alguth@davidson.edu
[†] e-mail: jegood@davidson.edu
[‡] e-mail: eugyoung.han@ufl.edu
[§] e-mail: bailenson@gmail.com
[¶] e-mail: tapeck@davidson.edu

2 Background

According to Brooks [2], a critical component of VR systems requires that the "tracking system... continually reports the position and orientation of the user's head and limbs". Most current VR devices include tracking functionality of the basic head and hand hardware used in VR. The headsets track the head position and orientation of a user and the hand trackers act similarly for hand movement. Functionally, the tracking data collected from hardware are fed back into the system to update a user's point of view and position within the virtual environment in real-time [32]. The collection of tracking data is therefore easily built into most VR devices.

Tracking data is used in real-time to run VR systems, however, it can also be post-processed and analyzed to provide insight into the movement and behavior of individuals within virtual environments (VEs). In this work, we focus on analyzing data after an experiment is run. Currently, motion data collected from VR studies is being used for a multitude of different applications, from analyzing behavior in scenarios created in VE's that are difficult to observe in real life, such as shooter bias [23], to aiding in medical rehabilitation and diagnosis [4, 19]. However, even with the expanding body of literature on applications using tracking data, the research community still lacks a set of tools and techniques to analyze motion data on a larger scale.

2.1 Basic Motion Tracking Metrics

Researchers have commonly used measures of central tendency and variability as metrics for analyzing head and hand tracking data including range, mean, and standard deviation. One of the simplest statistical measures of variability – range – has been used to analyze the extent of motion observed. In a study exploring user experience in a series of 360 degree video simulations [15], researchers track head movement, presence, arousal, and simulator sickness. Jun et al. [15], use the range of rotation on the yaw axis to examine how much of the 360-degree environment a user explored, from no horizontal rotation to full 360-degree exploration. Researchers discovered that new VR users tended to explore the environment less, videos without a focal point typically had a greater range of exploration, and videos that were explored more were preferred less by participants. Range has been used in current literature to determine the range of the environment explored, and focus and engagement with a VE when looking at how much a user turns their head [26].

Arguably the most common measure for analyzing movement in the current VR motion data analysis literature is a more nuanced statistical assessment of variability – the average standard deviation of motion on rotational axes. In a study analyzing the reflection of classroom anxiety in head movement patterns of participants, Won et al. [31] measured scanning behaviors (horizontal head movements), indicative of anxiety, using the standard deviation of head rotation on the yaw axis. They interpret a high standard deviation to denote high amounts of movement and a low standard deviation to reflect minimal movement. Researchers concluded that movement on the yaw axis was positively correlated with anxiety for female participants.

The standard deviation of rotation on hand and head orientation axes has also been used as a measure of overall movement, correlating higher values to more movement, in numerous other works. In a study investigating avatar representation and head movement on prosocial behaviors, Herrera et al. [12] found that head rotation from side to side was positively correlated to pro-social behaviors. In a different work, Herrera et al. [13] connected the measure to the realism of the avatars participants were embodied within, finding that participants in avatars co-located to their movements rotated their hands and head more than those with static avatars.

In a similar study, Li et al. [17] investigated the relationship between head movement in an immersive VR experience and user emotion. While watching a series of videos in VR, researchers measured participant head rotation on the yaw, pitch, and roll axes, finding that the standard deviation of head movement on the yaw axis, positively predicted valence (or emotion) such that more movement related to more emotion. Additionally, there was a relationship between the standard deviation of head movement on the pitch axis and arousal, showing participants moved their heads up and down more when they were excited.

Slater et al. [27] investigated the relationship between body movement and presence in a VE, using rotation on the yaw, pitch, and roll axes, as well as the mean and standard deviation of hand height above the ground as measures of movement. Researchers found that presence was positively associated with whole-body movement, including movement from a standing to sitting position, and changes in head rotation.

In another study, Cho et al. [4], investigated the relationship between head movement in a virtual classroom and reports of ADHD. They measured the magnitude of head movement by averaging the absolute value of orientation data points on the yaw, pitch, and roll axes in a given time frame to get a single measure to represent head movement on each axis. Researchers found a positive relationship between the magnitude of head movement and ADHD as measured in an ADHD questionnaire.

2.2 The Present Research

We have highlighted several ways that researchers have used descriptive statistics such as mean, standard deviation, and range as metrics for analyzing head and hand tracker data within studies assessing movement and behavior in VR. The current literature demonstrates that intriguing differences in behavioral patterns can be detected in motion data sets using these statistical metrics. However, there is currently no guidance within the field on when and for what types of research questions each statistical metric should be used. Researchers must choose a metric for analysis based on an example from past literature or perhaps an assessment of feasibility or ease of analysis. Ideally, researchers would choose a motion tracking metric based on the conceptual variables each metric is shown to operationalize.

In the present paper, we demonstrate the effectiveness of multiple motion tracking metrics, including those already commonly in use as well as three novel proposed metrics. For each metric, we first describe the calculation, proposed meaning, and suggested usage and limitations. We then provide an example of using that metric with existing motion tracking datasets in order to show how findings using that metric might be interpreted. Our intent is to create a guide for future researchers to use in their analysis of VR motion data sets.

3 Measures

In this section, we define and discuss measures for analyzing motion tracking data. These measures include common descriptive statistics (mean, standard deviation, range), previously used measures (P95), as well as new measures (95% range, time at mode, and coefficient of determination).

We define a motion tracking data set with n samples $p_1, p_2, p_3, \cdots, p_n$, where $p_i = (x_i, y_i, z_i, yaw_i, pitch_i, roll_i)$ for $i \in [1, n]$. For each sample, the starting point is normalized such that $p_1 = (0,0,0,0,0,0)$, and p_i is defined as the difference from p_1. Each participant may have multiple motion tracking data sets, for example, one per trial or block. For each of the following measures, one value is computed per motion tracking data set. Statistical analysis would then be performed on the computed measures for each participant.

An R script including example code to compute each measure is included in supplemental material as a resource for researchers.

3.1 Mean

The mean, denoted $\overline{p}$, is a commonly used descriptive statistics measure and is mathematically defined by

$$\overline{p} = (\overline{x}, \overline{y}, \overline{z}, \overline{yaw}, \overline{pitch}, \overline{roll}) = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{1}$$

It is a measure of central tendency that describes the average value of the dataset. When considering motion data, this measure would represent a user's average position or orientation on each axis, to compare across different trial factors. For example, when investigating how people angle their heads during positive and negative interactions [14] an experimenter might ask, "is the mean roll axis different during negative interactions compared to positive interactions?" An example of when the mean may not be a useful measure would be to investigate the average *x*-position when locomoting a maze since the *x*-position would frequently change.

A common limitation of the mean is that it is sensitive to outliers in small data sets. However, motion tracking data sets are often very large making them less susceptible to outliers. A variation to the mean can be 95% mean, where the upper and lower 2.5% of values are trimmed from the dataset before analysis. Researchers should still use caution and look for outliers in the data set when considering using the mean.

3.2 Range

Range, *R*, is a commonly used descriptive statistic that goes largely unused in the analysis of motion tracking data. Range notes the difference between the maximum and minimum value of a data set. Range is mathematically represented as,

$$R = p_{max} - p_{min} \tag{2}$$

Range is a measure of the dispersion in a dataset, or the spread of values represented. When used to analyze motion data, this measure can be helpful in denoting whether participants had a wide range of motion during their trials, or if they remained within a smaller subset of positions. As range is not a measure of central tendency, it shouldn't be used to describe a participant's average movement.

Range may be used when investigating the extent of the virtual environment a user interacts with. For example, did the user utilize the full field of regard, or did the user travel to the edges of the environment? Researchers may also consider range when analyzing differences in behavior between different visual stimuli in a 360-degree video environment and ask "does the range of motion on the yaw axis and room exploration vary depending on the video playing?" [15].

A limitation of range is that it is sensitive to outliers no matter the size of the data set.

3.3 95% Range

The limitation of outliers impacting range led us to propose the 95% range measure. This measure is based on common image processing techniques that remove the extreme values of the signal by trimming the lowest and highest values of the signal [3]. Outliers may appear in the data if tracking is briefly lost, or if the user makes an extreme movement, such as hearing a real-world noise that distracts them from the experiment. These outlier conditions may not be representative of the user's movements and should be removed from the data. However, defining these outliers can be challenging. To prevent outliers from impacting the analysis, following the standard 95% limits, we propose a limited range measure that removes the lower and upper 2.5% of a motion dataset. After removing the lower and upper portions of the dataset, the equation for limited range is identical to Equation 2.

The 95% range should be used in the same way as the full range measure, but may be especially helpful in preventing an inaccurate range calculation as a result of a significantly high- or low-value outlier. Researchers should analyze their datasets for outliers before analysis to determine if they should use the limited range measure rather than the full range.

3.4 Standard Deviation

Standard deviation, σ, is the square root of the average error between observations and the dataset mean. The standard deviation is mathematically defined by,

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - \bar{p})^2} \tag{3}$$

It is measured in the same units as the data making it easy to interpret. When considering motion data, standard deviation is a measure of how far a user strays from their average position/orientation. High standard deviation is an indication of a wider range and distribution of movement. Low standard deviation is an indication of a small range and distribution of movement, i.e. that the user did not stray far from the mean. A low standard deviation is also an indication that the mean is a good representation of the overall movement.

One application in which researchers might use standard deviation is in studying postural sway. For example, an experimenter may ask, "do people with Parkinson's disease have a greater standard deviation than participants without Parkinson's disease when considering their hand position when still?" [25]. In another application, researchers may ask "do students with social anxiety have a different standard deviation of movement on the yaw axis in a classroom setting than those without anxiety?" [31].

Standard deviation is the most common measure used to analyze motion data in the current literature, but one limitation is that it measures difference from the mean, and does not provide exact information about where a participant was positioned in the virtual environment. Standard deviation is therefore often more meaningful when used in tandem with a measure such as the mean, if researchers are analyzing differences in positional locations.

3.5 P95

Principal component analysis (PCA) determines how many components out of the six dimensions of the motion tracking data, are necessary to explain a participant's movement. P95 represents the sum of the variance accounted for by the first principal components that on average accounted for at least 95% of the variance in the data. P95 can be used as a measure of motion complexity, as more complex movements would rely on more axes than simpler movements. P95 gives an overall view of movement since it considers all axes at the same time, compared to the other measures that consider each motion axis separately.

When using P95 a researcher may ask, "does a user's overall movement patterns change based on the appearance of their self-avatar?" [16]. Or, "do overall movement patterns change based on the appearance of an avatar in the scene?" [23]. A limitation of using P95 is that it provides a high-level or gross estimate of movement complexity without zeroing in on a specific movement in one or more axes or orientations.

3.6 Time at Mode

We propose a novel metric that can be used with motion tracking data – the amount of time spent at the modal position. The mode is a common measure of central tendency which represents the most frequent value in a data set. In the context of motion data, the mode would be the positional location on each axis where a user spent the most amount of time. Depending on the precision of the tracking data, it may be necessary to round or bin the data before computing the mode. For instance, if the data is collected with precision to the hundred thousandths place, a single mode may not exist until all data is binned to three or four decimal points.

As a measure of central tendency, mode's use when analyzing motion data is similar to the mean. However, combining mode and duration data provides a unique insight into motion tracking data. Instead of considering the positional value of the mode itself, we ask how much time a user spent at their most frequented position.

We propose the metric as the percentage of time spent at the mode, calculated as the proportion of all samples representing the mode, which can be represented mathematically as,

$$M = \frac{freq(p_{mode})}{n} \tag{4}$$

If a larger percentage of samples are at the most frequented axis position, one can assume there was less overall movement as more time was spent at one position/orientation. In contrast, if a user spends a shorter amount of time at their most frequented position/orientation, more samples must be distributed across other positions.

The time spent at mode is an indication of focus on a target and may be a useful measure for experiments where participants take cognitive tests. For example, consider a study in which a researcher is investigating a participant's ability to focus on a stimulus within a virtual environment under different lighting conditions. In this case, all participants should have a similar mode, centered around where the stimulus is located. When considering the time spent at the mode, the researcher may ask "are there differences in the time spent at the mode in lighting condition A vs. lighting condition B?". A higher time at mode would indicate greater focus within one lighting condition compared to the other. A limitation of time spent at mode that tracking data is continuous and data must be rounded or binned in order to find a mode. We recommend binning data such that the percentage of time at mode is at least 5% of the time.

3.7 Coefficient of Determination

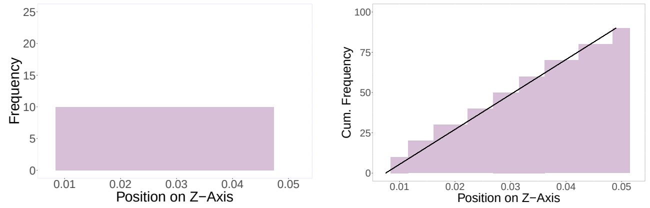
Finally, we propose the coefficient of determination as a novel metric for analyzing motion tracking data. The coefficient of determination, R^2, is a statistical measure that indicates how much of the variance in observations is explained by the independent variable and linear model, and is used as a measure of how well a linear model fits the data set. R^2 is a number between 0 and 1, with 0 indicating the model does not predict the outcome, and 1 indicating the model perfectly predicts the outcome. Mathematically the coefficient of determination is represented as

$$R^2 = 1 - \frac{RSS}{TSS} \tag{5}$$

where RSS is the sum of squares of the residuals of the linear model describing the relationship between motion data and the cumulative frequency at each axis position, and TSS is the total sum of squares of the model.

In the context of motion data, we propose this metric to act as a measure of skewness, or how evenly distributed a participant's time in an environment is across all possible axis positions. We use the coefficient of determination measure to analyze the proportion of time a user spends at each positional location using cumulative histograms. Cumulative histograms map the number of samples in which a participant is found at each location in a cumulative manner. The number of samples at position p_1 is added to the number of samples at position p_2 to get the value of the cumulative histogram at position p_2.

In its application, if each position was visited the same number of times, the cumulative histogram's line of best fit will fit perfectly and have an R^2 value of 1. This scenario is visualized in Figs. 2a and 2b. However, if a participant's movement was not standardized across all positions, such that more time was spent at certain locations over others, the R^2 value of the linear model for the cumulative histogram may be lower. Thus, this measure provides insight on a participant's range of motion and consistency of movement. Variations of R^2 are visualized in Figs. 1a and 1b.



(a) A histogram charting the position of a user's head on the z-axis during a VR experiment and the frequency they visited each location, when all of the frequencies are the same.

(b) A cumulative histogram with a black R^2 line showing how well the cumulative frequencies align with model predictions.

Figure 2: Comparison between standard and cumulative histograms of head position data to visualize the Coefficient of Determination measure.

An example of using the coefficient of determination could be an experiment where a researcher is comparing hand movement for two different grasping or manipulation tasks [7]. The participant is asked to move their hand back and forth along a designated path, or to grab an object. The researcher may ask, "is the participant's coefficient of determination higher when using grasping technique A or technique B?" where the higher coefficient of determination would indicate a more evenly distributed motion path.

4 Methods

To demonstrate the utility of the metrics described above, we applied each metric to the motion tracking data of two existing data sets, with permission from the previous researchers and approval from their Human Subjects Institutional Review Boards. Full details about the experimental designs are included in [11, 23].

Study 1: The full experimental details are presented in [23]. The purpose of the experiment was to assess an established psychological phenomenon called 'shooter bias' within a fully immersive virtual environment. Shooter bias is the documented tendency to more quickly decide to shoot armed Black men compared to armed White men and correspondingly more slowly decide not to shoot unarmed Black men compared to unarmed White men [5].

Participants stood for the duration of the experiment.The experiment was a within-participant user study in which participants were asked to draw a virtual weapon at an avatar that appeared in front of them in the virtual environment, and choose to shoot, by pulling the virtual trigger, or not shoot depending on the item the virtual avatar held. Following each shooting decision, participants were required to re-holster the gun on their right hip before the next avatar appeared. Specifically, participants were required to shoot a virtual gun at armed agents holding guns and not shoot at unarmed agents holding cell phones. We choose to analyze data from this specific study because we expect there to be differences in motion based on the participant's shooting decision and ultimately the item the virtual avatar holds. Therefore, we evaluate the effectiveness of each discussed measure, to detect differences in motion that we know is present based on the expected movements in this study. This study only tracked head and hand motion which, therefore, is what we analyze in this work.

Data was collected from N=101 participants, (55.5% women, 42.5% men, and 2.0% non-binary). Participants ages ranged from

18 - 76 years.

In addition to the item (gun or cell phone), the target agents varied by race, (Black or White), and by socioeconomic status (SES; high or low) which was indicated by the clothing they were wearing (business suit or tank top). Participants completed 160 shoot/don't shoot decision trials.

The focus of the original study was on assessing implicit racial bias in VR. In the current work, we present novel, exploratory analyses of the existing data in order to demonstrate the uses of each metric described in Section 3. For each metric we conducted a 2 (Item: gun or phone) × 2 (Race: Black or White) × 2 (SES: high or low) within-subjects ANOVA for both the hand and head data. The output from all tests is included in the supplemental material. Because the motion to point a (virtual) gun and pull the trigger should be different than the motion to prohibit oneself from pointing a gun and pulling the trigger, main effects of item type should be detected within motion data. As shown in Table 1, all metrics detected the significant main effect of item on at least four of the six degrees of freedom, $(x, y, z, yaw, pitch, roll)$.

Study 2: The full experimental details are presented in [11]. The purpose of the experiment was to assess the impact of spatial room dimensions, namely virtual ceiling height and floor area, on user behavior during networked social interactions.

The study was run at a university and involved students meeting in groups for virtual discussion sessions through ENGAGE, a social VR platform. All students were embodied and represented by virtual avatars. Virtually, students could freely navigate the virtual environment using tools such as teleportation and joystick-based smooth translation. Physically, the students remained seated for most of the duration of the sessions.

The same discussion groups met weekly for four weeks in one of four virtual environments. The virtual environments either had a tall or low ceiling, and a large or small floor area. Each discussion group was assigned at random via a Latin square randomization scheme such that they saw each configuration of the virtual environment once. The sessions consisted of a 20-minute group discussion and activity.

Data was collected from $N = 110$ participants, (55.5% women, 42.7% men, and 1.8% other). Participants ages ranged from $18 - 59$ years. Data from participants who did not complete all four sessions was removed.

The focus of the original study was to investigate the influence of spatial properties of a virtual environment during social interactions. While [11] focused on certain individual motion behaviors (e.g., physical head speed) and social behaviors (e.g., social attention, interpersonal distance), in the current presentation, we present a distinct set of outcome variables that were not previously reported. In the current study, we analyze the existing data by applying each metric described in Section 3. Specifically, we analyzed the motion data of participants during the first two minutes of each session, to determine if there were differences in movement as participants first explored and became acquainted with their environment. For each metric, we analyze physical head data by conducting a 2 (Ceiling: tall or low) × 2 (Floor area: large or small) within-subjects ANOVA.

5 Results

The full analysis is reported in the supplemental material. Due to space limitations we present at most one significant result per measure for each study. If nothing is reported then no significant results were found. In some cases, multiple main effects or interactions were found on different axes (for example, on both the $x - axis$ and the $z - axis$), however, only one will be reported. Assumptions for ANOVA tests, including normality and homogeneity of variance, were checked for all measures and met for most. Due to the large sample data sizes, even when normality is violated, F-Tests are still recommended and are considered robust in terms of type-I errors [1].

5.1 Mean

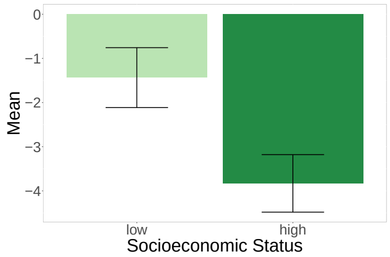
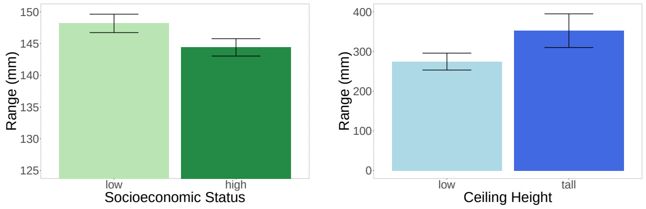


Figure 3: A bar chart showing the difference in the Mean measure on the x-axis hand motion in low-SES trials (light green) compared to high-SES trials (dark green). Measured in millimeters, our results indicate that participants held their hand further to one side in high-SES trials compared to low-SES trials.

Study 1: Using the mean, a significant main effect of SES was found for the hand tracking data along the x-axis, $F(1, 100) = 5.98$, $p = .02$, $\eta^2 = .005$. Participants had a lower mean x-position when making shooting decisions for high-SES avatars (M=-4.01mm, SE=1.14) compared to low-SES avatars (M=-1.49mm, SE=1.28), visualized in Figure 3. A mean of 0 would indicate that the average movement did not change from the starting point. This could indicate that the gun was not moved along the x-axis. Alternatively, a mean of 0 could also indicate that the gun was moved an equivalent distance to the left and right of the starting point. A mean value greater or less than 0 would indicate that on average the gun was moved more to the right, or more to the left respectively, than to the opposite side. These two explanations for mean movement values highlight the need to consider this metric in combination with a metric assessing variability, like range. In isolation, using only the mean could result in an incorrect interpretation of Figure 3, believing that participants moved their hands more when presented with a high-SES target compared to a low-SES target. When combined with a metric of variability, the mean tells a different story (see Section 5.2).

5.2 Range



(a) A bar chart showing the difference in the Range measure on x-axis hand position in low-SES trials (light green) compared to high-SES trials (dark green). Measured in millimeters, our results indicate that participants had a lower range of motion in high-SES trials than in low-SES trials. We altered the y-axis in the figure to start at a value other than 0 to make differences visible.

(b) A bar chart showing the difference in the Range measure on x-axis head position in trials with a low ceiling (light blue) compared to trials with tall ceilings (dark blue). Measured in millimeters, our results indicate that participants had a lower range of motion in low ceiling trials than in tall ceiling trials.

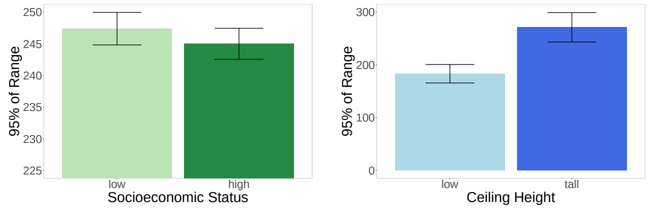
Figure 4: Range measure bar charts for Study 1 (left) and Study 2 (right).

Measures	Hand						Head					
	Main Effects						Main Effects					
	x	y	z	yaw	pit	rol	x	y	z	yaw	pit	rol
Mean	.01	.07	.06		.03	.07	.01	.01	.01	.01	.06	.01
Standard Deviation	.04	.05	.03	.05	.03	.07	.00	.01		.01		.00
Range	.07	.10	.05	.11	.03	.08	.00	.01	.00	.01	.01	.00
95% Range	.06	.10	.05	.06	.04	.09	.00	.01		.01	.01	.00
Time at mode	.07	.10	.06	.06	.03	.07	.02	.03	.02	.04	.03	.07
Coefficient of determination	.09	.29	.15	.17	.09	.27	.01	.01	.02	.01	.01	

Table 1: Study 1: The η^2 effect size for measures where a significant main effect ($p < .05$) of item were identified for each of the x, y, z, yaw, pitch, and roll measures for both hand (left) and head (right) tracking. Negligible effect sizes ($\eta^2 < .01$) are highlighted in white, small effect sizes ($.01 \leq \eta^2 < .06$) in yellow, medium effect sizes ($.06 \leq \eta^2 < .14$) in light orange, and large effect sizes ($.14 \leq \eta^2$) in orange.

Study 1: Evaluation of range identified a significant main effect of SES on the x-axis for hand motion, $F(1,100) = 7.14$, $p = .01$, $\eta^2 < .001$. Participants had a wider range of motion along the x-axis when making shooting decisions for low-SES avatars ($M =$ 150mm, $SE = 7.82$) compared to high-SES avatars ($M =$ 146mm, $SE = 7.47$), visualized in Figure 4a. A higher range indicates that when making the shooting decision, participants had more hand movement, moving the gun from left to right and vice versa, when faced with low-SES avatars than high-SES avatars. This metric in combination with a central tendency metric (like the mean) gives a more complete picture of hand movement. When presented with low-SES targets, participants had greater scanning movements to the left and right sides (indicated by a larger range and a mean closer to 0), but when faced with higher-SES targets, participants had less variable hand movement concentrated to the left of the center start point (indicated by smaller range and negative mean value). Although these effects are small and exploratory, and therefore should be interpreted with caution, one possible explanation is that participants used SES as a cue for danger. The high-SES target may have signaled less danger, and therefore participants focused their attention on the item in the target's hand (participants held the gun in their right hand and moved along the x-axis to the left side, where avatars were holding the gun or cell phone in their right hand). The low-SES target, however, may have signaled more potential danger, leading participants to scan across the visual field to a greater extent (participants moved their hand holding the gun in a greater range to both sides of the starting point).

Study 2: Evaluation of range identified a significant main effect of ceiling on the x-axis for head motion, $F(1,27) = 4.28$, $p = .05$, $\eta^2 = .02$. This suggests participants had a wider range of motion along the x-axis in trials where the ceiling was tall ($M =$ 353mm, $SE = 42.44$) compared to when the ceiling was low ($M =$ 275mm, $SE = 21.26$), visualized in Figure 4b. A higher range indicates that when in a vertically higher environment, participants had more head movement, from left to right, than when in a space with a low ceiling. This finding suggests individuals explore their range of motion more in a larger space.



(a) A bar chart showing the difference in the 95% range measure on the z-axis hand position in low-SES trials (light green) compared to high-SES trials (dark green). Measured in millimeters, our results indicate that participants had a lower range of motion in high-SES trials than in low-SES trials. We altered the y-axis to start at a value other than 0 to make differences visible.

(b) A bar chart showing the difference in the 95% range measure on the z-axis head position in trials with a low ceiling (light blue) compared to trials with tall ceilings (dark blue). Measured in millimeters, our results indicate that participants had a lower range of motion in low ceiling trials than tall ceiling trials.

Figure 5: 95% Range measure bar charts for Study 1 (left) and Study 2 (right)

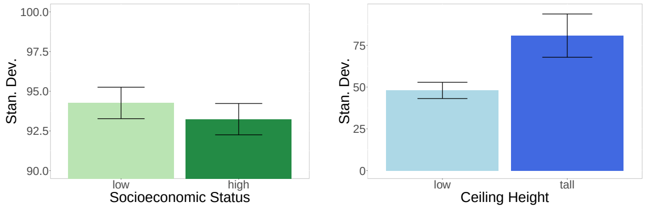
5.3 95% Range

Study 1: Using the 95% range, a significant main effect of SES was found on the z-axis for hand motion, $F(1,100) = 5.15$, $p = .03$, $\eta^2 < .001$. Participants had a higher range of motion along the z-axis when making shooting decisions against low-SES avatars ($M =$ 250mm, $SE = 17.9$) compared to high-SES avatars ($M =$ 247mm, $SE = 17.7$), visualized in Figure 5a. A higher range, even with outliers removed, indicates that when making the shooting decision, participants moved the gun closer to the target (farther from their hip) when presented with low-SES avatars rather than high-SES avatars. This means that across both weapon and non-weapon trials, participants tended to draw their gun more fully when a low-SES target appeared, again perhaps due to stereotypes of criminality or danger. Though significant, this was a very small effect (as indicated by the effect size and mean difference of only 3mm) and therefore, use caution to avoid over interpreting this difference. However, when combined with additional metrics assessing motion variability, greater confidence in detected motion patterns is warranted.

Study 2: Using the 95% range, a significant main effect of ceiling was found on the z-axis for head motion, $F(1,27) = 7.24$, $p =$ .01, $\eta^2 = .06$. Participants had a higher range of motion along the z-axis when the ceiling was tall ($M =$ 271mm, $SE = 27.88$) compared to when there was a low ceiling ($M =$ 193mm, $SE = 17.48$), visualized in Figure 5b. This finding suggests that in addition to an expanded range of motion on the x-axis, as described in Section 5.2, participants also have a higher range of motion forward to backward when in an environment with taller ceilings.

5.4 Standard Deviation

Study 1: Using standard deviation, a significant main effect of SES was found on the y-axis, $F(1,100) = 5.85$, $p = .02$, $\eta^2 <$ .001, visualized in Figure 6a. Participants had greater variation in their movements along the y-axis when making shooting decisions against low-SES avatars (M=95.70mm, SE = 7.48) compared to high-SES avatars (M=94.70mm, SE=7.42). This finding aligns with the effect of SES on the z-axis using the 95% range metric. In these example data, when participants drew their weapon to potentially shoot at a target, they had to move their hand upwards (y-axis) and forward (z-axis) from the starting position at their hip. Greater variability along these two axes suggests that participants tended to fully draw their gun more often for low-SES targets compared



(a) A bar chart showing the difference in the Standard Deviation measure on y-axis hand position in low-SES trials (light green) compared to high-SES trials (dark green). Our results indicate that participants' movement varied further from the mean in low-SES trials than high-SES trials, indicative of more general movement. We altered the y-axis to start at a value other than 0 to make differences visible.

(b) A bar chart showing the difference in the Standard Deviation measure on the x-axis head position in trials with a low ceiling (light blue) compared to tall ceiling trials (dark blue). Our results indicate that participants' movement varied further from the mean in tall ceiling trials than low ceiling trials, indicative of more general movement.

Figure 6: Standard Deviation measure bar charts for Study 1 (left) and Study 2 (right)

to high-SES targets. Note that the correct decision for participants when presented with unarmed targets was to not shoot (not draw their gun). Thus, these motion variability effects suggest that participants tended to fully draw their gun more often for low-SES targets, even if they ultimately decided not to shoot. For high-SES targets, participants may have attended to the item in the avatar's hand (as indicated by the mean metric finding) and more often partially drew their gun (or did not draw at all) when deciding not to shoot.

Study 2: Using standard deviation, there was a significant main effect of ceiling on the x-axis for head motion, $F(1,27) = 6.10$, $p = .02$, $\eta^2 = .05$. Participants had a greater variance in their movements along the x-axis when the ceiling was tall ($M = 81$, $SE = 12.96$) compared to when the ceiling was low ($M = 48$, $SE = 4.86$), visualized in Figure 6b. Similar to our findings for range and 95% range, this significance suggests participants move further from their average position on the x-axis in environments with taller ceilings.

5.5 P95

Study 2: Analysis of P95 identified a significant main effect of ceiling for head motion, $F(1,27) = 4.81$, $p = .04$, $\eta^2 < .02$. A greater percentage of variance was explained for participants in the taller ceiling environment ($M = 97.21\%$, $SE = .26\%$) compared to the lower ceiling environment ($M = 96.65\%$, $SE = .23\%$). This suggests that participants in the low-ceiling environment had more complicated movements.

5.6 Time Spent at Mode

Study 1: Using the time spent at mode, a significant main effect of SES was found on the y-axis for hand motion, $F(1,100) = 4.43$, $p = .04$, $\eta^2 < .001$. Participants spent a smaller percent of time at their most frequented position, indicative of more overall movement, when making shooting decisions against low-SES avatars ($M = 15.7\%$, $SE = 0.61$) compared to high-SES avatars ($M = 16.0\%$, $SE = 0.62$). All values were rounded to the thousandth place before applying the measure. A visualization of these results can be found in Figure 7. A smaller time spent at mode value indicates more time was spent in a variety of locations, as the participant did not hold their hand in one place for very long. Again, this finding aligns with the 95% range and Standard Deviation findings, adding confidence to the interpretation that participants tended to draw and reholster their gun more often for low-SES targets while they were less likely to fully draw (perhaps they paused mid-draw) their gun for high-SES targets. This motion pattern suggests that participants were using avatar clothing (cueing SES) as a signal of possible danger and therefore drew their gun more fully even when ultimately they decided not to shoot unarmed low-SES targets.

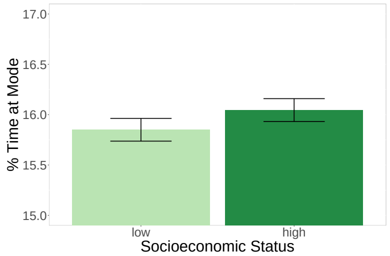


Figure 7: A bar chart visualizing the percentage of time participants spent at the trial mode of their y-axis hand position, depending on whether or not the avatar appearing in front of them was dressed resembling low socioeconomic status (light green, left) or high socioeconomic status (dark green, right).We altered the y-axis in the figure to start at a value other than 0 to make differences visible.

5.7 Coefficient of Determination

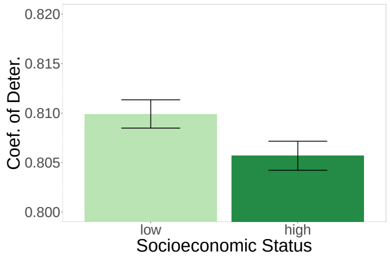


Figure 8: A bar chart showing the difference in the Coefficient of Determination measure on z-axis head position in low-SES trials (light green) compared to high-SES trials (dark green). Measured between 0 and 1, our results indicate that participants had less equally distributed motion in high-SES trials, indicative of less overall movement, than in low-SES trials. We altered the y-axis to start at a value other than 0 to make differences visible.

Study 1: Using the coefficient of determination, a significant main effect of SES was found on the z-axis for head motion, $F(1,100) = 4.23$, $p = .04$, $\eta^2 < .001$. Participants' time during trials was more equally distributed across positions along the z-axis when making shooting decisions against low-SES avatars ($M = 0.810$, $SE = 0.003$) compared to high-SES trials ($M = 0.806$, $SE = 0.003$), visualized in Figure 8. All values were rounded to the tens-thousandth place before applying the measure. Example histograms and cumulative histograms for a participant's hand movements along the y-axis when making shoot/no shoot decisions are shown in Figs. 1a and 1b for comparison against the example Figs. 2a and 2b in Section 3.

Though a less familiar metric for some readers, we propose that the coefficient of determination provides evidence of how smoothly, or equally distributed across space, a movement is executed. Consider that in the example data we have been using, when a participant draws their gun (moving their hand upward and forward), based on physiology their head will move forward as their arm or shoulder moves forward. A larger R^2 value for head movement along the z-axis (forward and back) for low-SES targets corresponds with the greater hand motion variability along the y- and z-axes for these targets, or rather equal time spent across more positions. As participants more often drew and then re-holstered their virtual guns, their heads moved more smoothly forward (draw) and back (re-holster). For high-SES targets, the lower R^2 value suggests that participants moved their head forward and back in a more halting or disjointed manner, or rather spent a disproportionate amount of time at each position. This aligns with the hand motion data to suggest that participants more often partially drew their virtual gun for high-SES targets, likely pausing, and staying in one position, when the target was determined to be unarmed.

6 Discussion

Through our analysis, it is apparent that a wealth of information is held within large tracking data sets collected from VR user studies. The challenge comes in using the appropriate tools to uncover differences in motion and understand what these differences indicate. In this paper, we aim to provide a guide for use of various metrics and demonstrate through analysis on two existing datasets, one with standing participants and one with seated participants, that the metrics we propose are successful in detecting movement differences across trial factors.

Due to the context of our first dataset, collected from a previous study investigating shooter bias, where participants were standing and performed repeated hand motions to make shoot and no shoot decisions based on the item held by an avatar in the scene (gun or cell phone), we expected to find differences in motion based on the item. The significant results and effect-sizes are summarized in Table 1, with almost all measures finding significant item main effects across all axes.

When considering hand movements, effect sizes ranged from small to large. Standard deviation has been a commonly used measurement [13, 17], and was able to identify significant differences between items with small effect sizes on all axes except roll which had a medium effect. When compared to other measures, standard deviation had weaker effects in detecting differences in these hand motions. The newly proposed measures, 95% range and time at mode identified medium effects on the majority of axes, and coefficient of determination identified medium to large effects on all axes. This supports that these new measures may be useful at better identifying differences in movements compared to previously used techniques, especially when analyzing hand motions in 3D user interface tasks.

With this confirmation that our measures were successful in detecting motion differences we expected to exist (main effects of item), our remaining analyses of this dataset were largely exploratory in examining possible effects of target SES on participants' movement.

In addition to the expected findings for item type in the first study, the mean metric was also effective in finding differences in the average hand position for low vs. high-SES target avatars. The mean is a commonly used metric and is helpful when researchers want to know differences in position versus motion as a whole, as it will give a sense of where a participant spent most of their time. However, since the mean is sensitive to outliers and not effective in contexts where there is constant motion, researchers should ensure that main effects and interactions detected by the mean make sense in the context of their study. In the first example dataset used in this paper, the effect of SES on mean hand position along the x-axis is an example of how using the mean in isolation could lead to incorrect interpretation. Measures of central tendency are typically most informative when combined with measures of variability. In our example case, if mean were used in isolation a researcher could incorrectly conclude that participants moved their hand *less* when faced with low-SES targets than with high-SES targets. In combination with additional measures of variability however, it is evident that in reality participants moved their hands *more* when faced with low-SES targets, but simply with greater distribution of movement to the right and left of the center starting point (resulting in a mean closer to zero).

Variability in movement can be assessed with the range, 95% range, and standard deviation. Standard deviation has been commonly used to analyze motion data, but we argue that range and 95% range can provide important additional information. Although all related, and therefore logically showing similar findings, there are cases where these three metrics could have different results. Additionally, the combined use of multiple variability metrics could give researchers greater confidence in their interpretation of a finding, particularly when using multiple axes of motion. In the first example dataset used here, we observed significant main effects of SES on movement variability along the x-axis (range), z-axis (95% range) and y-axis (SD). In fact, range, 95% range and SD all had significant main effects of SES on the x, y, and z axes (full results are reported in the supplemental material). Although these were all small effects, in combination they demonstrate that when presented with low-SES targets, participants tended to draw their virtual gun more fully and focus on the whole target compared to partially drawing or failing to draw the gun and focusing more on the left side (where the target was holding the item in question) when presented with high-SES targets. These findings are consistent with psychological literature demonstrating negative stereotypes of people with low-SES, particularly beliefs about danger and criminality [6].

Due to the context of our second dataset, which explored the effect of environment spatial dimensions on user behavior and movement, while seated, we expected there to be differences in head motion based on either the ceiling (low or tall) or the room (small or large) conditions. In our analysis, we only analyzed movement during the first 2 minutes of this study. We limited our analysis to the time frame participants were initially becoming acclimated with the environment because different movements were expected across different periods of the 20-minute study. Therefore, researchers should determine prior to analysis whether their data should be analyzed cumulatively or in chunks corresponding to specific movements or actions expected during their study, as well as the length of the study. Data collected over a shorter time frame with more standardized movements may produce more significant findings. Numerous significant differences in motion were detected based on the ceiling height along the x-axis (SD, range, 95% range) and z-axis (SD, 95% range) with wider ranges of head motion along the x- and z- axes in the taller ceiling condition. We note that compared to the first dataset, this work was more exploratory as there were no known differences in movement that we were ensuring our measures could find. However, these findings may help contextualize the results reported by [11], particularly when considered in conjunction with self-report measures such as awe, perceived restorativeness, and momentary affective well-being.

Consistent with prior literature, [11] highlighted the positive impact of tall ceilings. In the current study, we examined subtle physical head movements that may offer additional insight into these effects. Notably, the present analysis focuses on the first two minutes of the sessions, when participants were likely acclimating to the virtual environment. The wider range of motion along the x- and z- axes, as well as the greater variance along the x-axis, may reflect participants' exploration of the virtual environment, whereas being in low-ceiling environments may have offered limited visual stimuli for such exploration.

Researchers should determine which measures of variability would be most helpful given their dataset and research question. Range would be helpful in contexts where users are moving between possible extreme points and range of motion is explored, or in contexts where researchers want to determine how much of an environment a user interacted with. One limitation researchers should be aware of with the range metric is the sensitivity to outliers.

To address range's sensitivity to outliers, we propose the 95% range metric (which removes the lower and upper 2.5% of the position data) similar to the statistical practice of trimmed means [30] and autocontrast algorithms for image processing [3]. In our first example dataset, the 95% range had identical results to the full range measure, aside from one main effect on the head's z-axis. The 95% range did not find a main effect of item while full range did find a main effect of item. However, SD also did not find this main effect (see Table 1). When considering the second dataset, 95% range and SD identified differences in movement between ceiling heights in the z-axis while range did not. 95% range should be considered a more conservative representation of range. Researchers should also determine if there are substantial outliers in their dataset prior to deciding which metric to use, and ensure they use the 95% range if necessary. We propose trimming the upper and lower 2.5% of data to mimic previous standards, however different trim values could be used based on the makeup of the datasets.

The use of standard deviation to analyze motion tracking datasets is supported by the previous literature. Standard deviation is useful in contexts when researchers want to know about typical deviations from the mean position rather than the full extent of all motion (for which range would be a better metric). When used in combination with the mean, standard deviation can indicate how closely participants' movements clustered around a focal point on average, even if participants had a few extreme movements within a given session.

Beyond these basic descriptive statistical metrics, measures of movement complexity are also becoming more commonplace in motion tracking data analysis. One metric supported by the literature is P95 which is a measure of how many dimensions of motion tracking data are needed to account for 95% of the variance in the data. This metric is useful when researchers do not have a specific motion pattern predicted, but instead want a higher-level picture of overall motion. When applied to our first dataset, significant effects of Item were found in the P95 measure, but no other effects. Therefore, our results support the ability of the measure to find movement differences (the expected contrasts were identified). Additionally, when applied to our second dataset, significant effects of Room were found in the P95 measure. We suggest using P95 alongside other measures outlined in this paper, as a representation of the overall movement of a participant in VR.

In addition to 95% range, we present another two novel metrics that can be used on their own or to augment researchers' interpretation of findings from other metrics. Both metrics provide indication of movement fluidity or consistency, although with differing interpretations.

Adapting a commonly used descriptive measure of central tendency (mode), we propose analyzing the actual amount of time spent at the modal position rather than the positional value of the mode. In the first dataset, the time spent at the mode metric was effective in finding significance across the item condition on all axes for both hand and head data. It was also effective in finding significant effects across the SES condition, showing more consistently varied movement when presented with low-SES targets compared to high-SES targets. Just as we argued that mean effects were more appropriately interpreted alongside a measure of variability, time spent at mode provides a clearer understanding of a central tendency metric. With this metric researchers can understand how a participant's movements are distributed across both time and space rather than just location (as with range and SD metrics). However, to obtain accurate findings, researchers may need to truncate all values in the dataset to varying degrees of precision, depending on the context of the study. During analysis of our datasets, we repeatedly truncated the values to a decreasing degree of precision until the percent of time at the calculated mode was at least 5%. Truncating values at the thousandth place proved effective. The time at mode measure would be especially helpful in studies where there is an element of focus or attention. In these cases, the time at mode metric would be a higher value in situations of high focus, as the participant spent a substantial amount of time in one location surrounding a stimulus, and lower in situations with lower focus, as the participant failed to remain fixed in one location for long.

Finally, we developed the coefficient of determination (R^2) measure, to explore how equally distributed motion was across possible positions, or "how smooth was the motion from point a to point b". However, we acknowledge that there may be conditions where a participants movements were not "smooth" but simply equally distributed across positional values. Further, in contrast to existing kinematic measures concerned with jerk, movement velocity, etc., R^2 focuses on the equal distribution of movement rather than the fluidity of movement between positions in space. When applied to our first dataset, R^2 was similarly effective in detecting differences in hand and head motion across the item condition with medium to large effects.

Additionally, the R^2 metric identified a main effect of SES along the z-axis for head motion data, showing equally distributed movement or more consistent head movement forward and back in the low-SES condition, when participants tended to move their hand (and therefore arm and shoulder) to a greater extent along the y- and z-axes. More halting or jerky head movement was observed in the high-SES condition, when participants tended to hesitate or pause their hand movement related to drawing their gun. We propose that the coefficient of determination metric, though still in need of greater study, could be especially useful in studies where researchers want to understand users' comfort or expertise with a given motion, or evaluation of different grasping and interaction techniques. For example, after practice we might expect a user's progression through a series of movements to be more consistently distributed compared to initial trials before developing expertise. Or, we might expect a user's motion to be more consistent with a more usable interaction technique compared to a more complicated technique.

7 Conclusion

The purpose of the present research was to provide a guide for researchers working with head and hand motion datasets. We provide an overview of commonly used metrics for analysis (mean, standard deviation, range, P95) as well as three novel metrics (95% range, time spent at mode, and coefficient of determination), noting important considerations for use of each metric. We then utilized two existing motion tracking datasets, one with standing participants and one with seated participants, to provide examples of findings obtained with each metric. These results showcase both the nuance of each metric as well as the value of using certain metrics in combination. Overall, we argue for the importance of researchers basing their choice of analytic metric on their specific research question and the particularities of their motion data. Rather than prescribing wholesale adoption of a set of metrics for all motion studies, this paper provides valuable guidance for researchers on when and why to use certain metrics so that researchers can choose the best metrics for their research.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants 1942146 and 1800922. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] M. J. Blanca Mena, R. Alarcón Postigo, J. Arnau Gras, R. Bono Cabré, and R. Bendayan. Non-normal data: Is anova still a valid option? *Psicothema, 2017, vol. 29, num. 4, p. 552-557*, 2017. 5

[2] F. P. Brooks. What's real about virtual reality? *IEEE Computer graphics and applications*, 19(6):16–27, 1999. 2

[3] W. Burger and M. J. Burge. *Digital image processing: An algorithmic introduction*. Springer Nature, 2022. 3, 9

[4] Y. J. Cho, J. Y. Yum, K. Kim, B. Shin, H. Eom, Y.-j. Hong, J. Heo, J.-j. Kim, H. S. Lee, and E. Kim. Evaluating attention deficit hyperactivity disorder symptoms in children and adolescents through tracked head movements in a virtual reality classroom: The effect of social cues with different sensory modalities. *Frontiers in Human Neuroscience*, 16:943478, 2022. 2

[5] J. Correll, B. Park, C. M. Judd, and B. Wittenbrink. The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, 83(6):1314–1329, 2002. doi: 10.1037/0022-3514.83.6.1314 4

[6] C. Cozzarelli, A. V. Wilkinson, and M. J. Tagler. Attitudes toward the poor and attributions for poverty. *Journal of social issues*, 57(2):207–227, 2001. 8

[7] D. Dewez, L. Hoyet, A. Lécuyer, and F. Argelaguet. Do you need another hand? investigating dual body representations during anisomorphic 3d manipulation. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2047–2057, 2022. 4

[8] M. Gonzalez-Franco and T. C. Peck. Avatar embodiment. towards a standardized questionnaire. *Frontiers in Robotics and AI*, 5:74, 2018. 1

[9] M. González-Franco, T. C. Peck, A. Rodríguez-Fornells, and M. Slater. A threat to a virtual hand elicits motor cortex activation. *Experimental brain research*, 232(3):875–887, 2014. 1

[10] L. Hak, H. Houdijk, P. J. Beek, and J. H. van Dieën. Steps to take to enhance gait stability: the effect of stride frequency, stride length, and walking speed on local dynamic stability and margins of stability. *PloS one*, 8(12):e82842, 2013. 1

[11] E. Han, C. DeVeaux, J. T. Hancock, N. Ram, G. M. Harari, and J. N. Bailenson. The influence of spatial dimensions of virtual environments on attitudes and nonverbal behaviors during social interactions. *Journal of Environmental Psychology*, 95:102269, 2024. 4, 5, 8

[12] F. Herrera and J. N. Bailenson. Virtual reality perspective-taking at scale: Effect of avatar representation, choice, and head movement on prosocial behaviors. *new media & society*, 23(8):2189–2209, 2021. 2

[13] F. Herrera, S. Y. Oh, and J. N. Bailenson. Effect of behavioral realism on social interactions inside collaborative virtual environments. *Presence*, 27(2):163–182, 2020. 2, 8

[14] D. C. Jeong, D. Feng, N. C. Krämer, L. C. Miller, and S. Marsella. Negative feedback in your face: examining the effects of proxemics and gender on learning. In *International conference on intelligent virtual agents*, pp. 170–183. Springer, 2017. 3

[15] 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*, 13(3):1416–1425, 2020. 2, 3

[16] K. Kilteni, I. Bergstrom, and M. Slater. Drumming in immersive virtual reality: the body shapes the way we play. *IEEE transactions on visualization and computer graphics*, 19(4):597–605, 2013. 1, 3

[17] B. J. Li, J. N. Bailenson, A. Pines, W. J. Greenleaf, and L. M. Williams. A public database of immersive vr videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures. *Frontiers in psychology*, 8:2116, 2017. 2, 8

[18] J. Llobera, B. Spanlang, G. Ruffini, and M. Slater. Proxemics with multiple dynamic characters in an immersive virtual environment. *ACM Transactions on Applied Perception (TAP)*, 8(1):1–12, 2010. 1

[19] G. Lugo, M. Ibarra-Manzano, F. Ba, and I. Cheng. Virtual reality and hand tracking system as a medical tool to evaluate patients with parkinson's. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pp. 405–408, 2017. 2

[20] M. Meehan, B. Insko, M. Whitton, and F. P. Brooks Jr. Physiological measures of presence in stressful virtual environments. *Acm transactions on graphics (tog)*, 21(3):645–652, 2002. 1

[21] M. R. Miller, C. DeVeaux, E. Han, N. Ram, and J. N. Bailenson. A large-scale study of proxemics and gaze in groups. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 409–417. IEEE, 2023. 1

[22] T. D. Parsons, C. G. Courtney, B. Arizmendi, and M. Dawson. Virtual reality stroop task for neurocognitive assessment. In *Medicine Meets Virtual Reality 18*, pp. 433–439. IOS Press, 2011. 1

[23] T. C. Peck, J. J. Good, and K. Seitz. Evidence of racial bias using immersive virtual reality: Analysis of head and hand motions during shooting decisions. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2502–2512, 2021. 2, 3, 4

[24] T. C. Peck, L. E. Sockol, and S. M. Hancock. Mind the gap: The underrepresentation of female participants and authors in virtual reality research. *IEEE transactions on visualization and computer graphics*, 26(5):1945–1954, 2020. 1

[25] J. M. Prado, T. A. Stoffregen, and M. Duarte. Postural sway during dual tasks in young and elderly adults. *Gerontology*, 53(5):274–281, 2007. 3

[26] J. A. Schwanewede, A. A. Guth, and T. C. Peck. The impact of environment design bias on working memory. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2

[27] M. Slater, A. Steed, J. McCarthy, and F. Maringelli. The influence of body movement on subjective presence in virtual environments. *Human factors*, 40(3):469–477, 1998. 2

[28] M. Slater, M. Usoh, and A. Steed. Depth of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 3(2):130–144, 1994. 1

[29] Y. Wada, T. Yamanaka, T. Kitahara, and J. Kurata. Effect of head roll-tilt on the subjective visual vertical in healthy participants: Towards better clinical measurement of gravity perception. *Laryngoscope Investigative Otolaryngology*, 5(5):941–949, 2020. 1

[30] R. R. Wilcox. Trimmed means. *Wiley StatsRef: Statistics Reference Online*, 2014. 9

[31] A. S. Won, B. Perone, M. Friend, and J. N. Bailenson. Identifying anxiety through tracked head movements in a virtual classroom. *Cyberpsychology, Behavior, and Social Networking*, 19(6):380–387, 2016. 2, 3

[32] H. E. Yaremych and S. Persky. Tracing physical behavior in virtual reality: A narrative review of applications to social psychology. *Journal of Experimental Social Psychology*, 85:103845, 2019. 2

[33] R. Zemblys, D. C. Niehorster, O. Komogortsev, and K. Holmqvist. Using machine learning to detect events in eye-tracking data. *Behavior research methods*, 50:160–181, 2018. 1