Statistical Analysis of Empathy in Virtual Reality: Emotional Responses and Individual Differences

Ivana Trenevska[*†]
Marija Kochova[*†]
ksiar32024@feit.ukim.edu.mk
kti1572024@feit.ukim.edu.mk
Faculty of Electrical Engineering and Information Technologies, SS. Cyril and Methodius University in Skopje
Skopje, North Macedonia

Jasmina Angelevska Kostadinoska
Emilija Kizhevska
Katerina Hadzi-Velkova Saneva
jasminaa@feit.ukim.edu.mk
ekizevska@feit.ukim.edu.mk
saneva@feit.ukim.edu.mk
Faculty of Electrical Engineering and Information Technologies, SS. Cyril and Methodius University in Skopje
Skopje, North Macedonia

Abstract

This paper presents a systematic statistical analysis of a virtual reality empathy dataset comprising 105 participants exposed to immersive 360° emotional videos, examining associations between individual characteristics, trait empathy, emotional responses, and features of the virtual reality experience. The analysis combines correlation and hypothesis testing with multiple-comparison correction, repeated-measures analysis, effect-size estimation, and multiple linear regression. Perceived presence was positively associated with average state empathy, whereas cognitive trait empathy showed a negative association in the multivariable model. Happiness differed significantly from sadness, anxiety, and anger across empathy, arousal, and valence. Self-reported arousal during the Rollercoaster condition also differed significantly from that reported after all four emotion-based scenarios. Overall, state empathic responses were more consistently associated with the VR experience and emotional condition than with the demographic factors examined, highlighting the importance of controlling for multiple comparisons.

Keywords

Virtual reality, Empathy, Hypothesis testing, Multiple linear regression, Nonparametric statistics, Repeated measures, Data analysis

1 Introduction

Empathy is commonly distinguished into two complementary components: *cognitive empathy*, the ability to understand another person's emotional state or perspective, and *affective empathy*, the capacity to share or resonate with another person's emotions. Whereas trait empathy reflects relatively stable individual tendencies, state empathy refers to a situation-specific empathic response elicited by a particular context or stimulus. Understanding how these components manifest under controlled conditions is relevant to psychology, neuroscience, and VR research, where presence has been linked to realistic responses [18] and VR perspective-taking to increased empathic engagement [7]. Recent advances in VR technology allow researchers to study emotional responses in standardized, immersive environments, generating datasets that require rigorous statistical analysis to distinguish genuine relationships from random variation.

The dataset analyzed here originates from a VR study investigating empathy elicited by immersive 360° videos [9, 10]. The original study primarily focused on predicting empathy and other mental states from physiological sensor data using machine learning, while also reporting selected statistical comparisons. However, the associations among demographic characteristics, trait and state empathy, emotional responses, and VR-experience measures have not been systematically examined. This paper addresses this gap using a broader combination of inferential statistical methods.

2 Dataset Description

The dataset originates from Kizhevska *et al.* [9], a virtual reality experiment investigating empathy and physiological responses to immersive 360° VR videos. It consists of 105 participants (mean age 22.43 years; 76% women) who completed a demographic questionnaire and a trait-empathy questionnaire, then viewed four emotional scenarios (sadness, anxiety, happiness, anger), rating state empathy, valence, arousal, and discomfort after each [2, 3, 15, 17] on a 0–4 scale (higher = more empathy, more negative valence, higher arousal, more discomfort), while feelings of presence, captivation, and realism were rated on a 9-point scale. Average Empathy, the main regression outcome, is the mean of the four state-empathy ratings. Participants were assigned to a narrative condition (four versions with an audio narrative preceding each scenario) or a no-narrative condition (one version) [9]. A final Rollercoaster sequence elicited non-empathic arousal; only self-reported arousal is analyzed here, not physiological signals. Ethical approval and informed-consent procedures for the original data collection are reported in [9].

Gender, education (ordinal, primary school to PhD), and field of expertise (15-category classification; STEM and Human Services subgroups used for comparison) were coded following the categorization used in the original questionnaire.

The dataset was screened for missing and out-of-range values: Rollercoaster Arousal was missing for one participant (pairwise exclusion, $n = 104$), and one out-of-range Happiness Valence value (recorded as 9) was treated as a data-entry error and recoded to missing ($n = 104$). All other variables were complete for all 105 participants.

[*]Corresponding authors.
[†]These authors contributed equally to this work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026 , Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.cogni.62

The complete dataset contains 67 variables, including physiological signals not analyzed here; the present analysis uses the demographic, questionnaire-based, and self-reported VR-response variables.

3 Methodology

3.1 Research Questions

The study examines associations between individual characteristics, trait and state empathy, emotional responses, and features of the virtual reality experience through the following research questions:

- *RQ1:* Is trait empathy associated with state empathy experienced during the virtual reality scenarios?
- *RQ2:* Are educational level and field of expertise associated with differences in trait or state empathy?
- *RQ3:* Are gender and age associated with differences in empathy and emotional responses during the VR experience?
- *RQ4:* Do sadness, anxiety, happiness, and anger differ in empathy, arousal, and valence, and which emotional conditions show the most distinct response patterns?
- *RQ5:* Does self-reported arousal differ between emotion-based and non-empathic VR stimuli?
- *RQ6:* Which characteristics of the virtual reality experience are associated with state empathy?

3.2 Statistical Analysis

Statistical procedures were selected based on the measurement scale, study design, distributional characteristics, and assumptions of the corresponding tests. Multiple-comparison corrections and effect-size measures were applied where appropriate. All statistical analyses were performed in R version 4.6.0.

3.2.1 Descriptive Statistics. The analysis begins with descriptive statistics for continuous and ordinal variables, including the mean, standard deviation, median, and interquartile range. For brevity, these measures are reported in Appendix A.

3.2.2 Correlation Analysis. To examine linear relationships between continuous variables, we computed Pearson's correlation coefficient ($r \in [-1, 1]$) [14] and tested $H_0 : r = 0$ against $H_1 : r \neq 0$ at $\alpha = 0.05$. Pearson correlation analysis was performed across 23 prespecified numeric variables, selected a priori based on their direct relevance to the six research questions (demographic, trait-empathy, state-empathy, and VR-experience measures), yielding $\binom{23}{2} = 253$ pairwise tests, which constituted the complete family of hypotheses for Benjamini–Hochberg (BH) correction [1]. Field of expertise (Job), a nominal categorical variable, was excluded because Pearson's correlation coefficient is not appropriate for unordered categorical data. Ordinal variables with limited response categories (e.g., valence, arousal) were treated as approximately continuous, consistent with common practice for Likert-type scales [13]. As a robustness check, Spearman's rank correlation [19] was computed for all 253 pairs; results were consistent with the Pearson correlations in direction and magnitude (mean absolute difference = 0.02, maximum = 0.09), yielding a comparable set of significant associations after the BH correction (62 vs. 68 pairs). Correlation coefficients, uncorrected p-values, and BH-adjusted p-values are reported in Appendix B, limited to correlations judged theoretically relevant to the research questions and excluding definitional or partial–whole relationships (e.g., individual empathy conditions with the average empathy score) and pairs reflecting measurement overlap between closely related instruments (e.g., presence, captivation, and realism of the virtual environment).

3.2.3 Distribution Assessment and Confidence Intervals. Distributional characteristics were assessed using the Shapiro-Wilk test [16] at $\alpha = 0.05$, complemented by graphical inspection of the distributions. Where appropriate, 95% confidence intervals for population means were constructed using Student's t distribution [20] or large-sample approximations [12] for variables with an asymptotically normal sampling distribution. The results are reported in Appendix C.

3.2.4 Independent-Group and Paired Comparisons. Independent-group comparisons were performed by gender, narrative condition, and field of expertise. Equality of variances was assessed using Levene's test [11], followed by Student's t-test when variances were equal and Welch's t-test otherwise, with effect sizes quantified using Hedges' g [6] and Holm correction [8] applied for multiple comparisons. Paired-sample t-tests similarly compared self-reported arousal during the Rollercoaster condition with that reported after each emotion-based VR scenario. Effect sizes for the paired comparisons were quantified using Hedges' g for paired samples. Normality of the within-participant differences was assessed, and the Wilcoxon signed-rank test [21] was used as a nonparametric sensitivity analysis; Holm correction was applied independently to both the t-test and Wilcoxon p-values.

A summary of the analyses is provided in Appendix D, Tables 9 and 10.

3.2.5 Multiple Linear Regression. A multiple linear regression model [12] was fitted to examine which individual and VR-related characteristics were independently associated with average state empathy, computed as the mean of the four condition-specific empathy scores to obtain a single participant-level outcome. The model is specified as

$$\begin{aligned}\text{Average Empathy} = \beta_0 &+ \beta_1(\text{Age}) + \beta_2(\text{Education}) + \beta_3(\text{Gender}) \\ &+ \beta_4(\text{Affective Empathy}) \\ &+ \beta_5(\text{Cognitive Empathy}) \\ &+ \beta_6(\text{Presence in VE}) + \varepsilon.\end{aligned}$$

Categorical variables were coded as follows: Gender was dummy-coded with Female as the reference category (0 = Female, 1 = Male). Education entered the model as a single numeric ordinal variable (coded as described in Section 2), rather than as a set of dummy variables. Model assumptions were assessed before interpretation of the regression coefficients: residual normality using the Shapiro–Wilk test, homoscedasticity using the Breusch–Pagan test, multicollinearity using variance inflation factors (VIFs), and linearity by visual inspection of the residuals-versus-fitted plot. Regression coefficients were evaluated at $\alpha = 0.05$, with 95% confidence intervals reported for the estimated effects. The regression coefficients and corresponding significance tests are shown in Appendix D, Table 11.

3.2.6 Nonparametric Tests. Variables that do not satisfy the assumptions required for parametric procedures are analyzed using nonparametric methods. Since several measurements were obtained repeatedly from the same participants under the four emotional conditions, the observations are dependent. Friedman tests [5] are therefore used to test whether the response distributions differ in location across the four repeated conditions (H_0:

Table 1: Significance Summary: Independent-Sample t-Tests (Before and After Holm Correction)

Variable	Group	Sig. (Uncorrected)	Sig. (Holm)
Average Empathy	Gender	Yes	Yes
Affective Empathy	Gender	Yes	No
Cognitive Empathy	Gender	Yes	No
Age	Gender	Yes	Yes
Sadness Empathy	Gender	Yes	No
Anger Empathy	Gender	Yes	No
Anger Arousal	Gender	Yes	No
Sadness Valence	Gender	Yes	No
Average Empathy	Narrative	Yes	No
Sadness Empathy	Narrative	Yes	No
Captivated by VE	Narrative	Yes	Yes
Realism of VE	Narrative	Yes	No
Affective Empathy	STEM vs Human Services	Yes	No
Cognitive Empathy	STEM vs Human Services	Yes	No

equal location; H_1: at least one condition differs). Kendall's W was reported as an effect-size measure for each Friedman test.

Whenever a Friedman test is significant, pairwise post-hoc comparisons are conducted using the Wilcoxon signed-rank test for paired samples [21], providing a pairwise rank-based comparison of the within-participant responses (H_0: no difference in location; H_1: a difference in location) with Bonferroni correction applied across the six pairwise comparisons within each measure.

The complete results of the Friedman tests are presented in Appendix D, Table 12, while the corresponding pairwise comparisons obtained from the Wilcoxon signed-rank post-hoc procedure are summarized in Appendix D, Table 13.

4 Results and Discussion

4.1 Statistical Results

The analyses in Appendices A–D show several relationships among demographic, empathy, and emotional-response variables. Table 1 summarizes the 14 independent-sample t-tests conducted; all 14 were nominally significant at $p < 0.05$ before correction, while only 3 survived the Holm correction – average state empathy and age (both by gender), and engagement with the virtual environment (by narrative condition). Table 2 shows that self-reported arousal during the Rollercoaster condition differed significantly from that reported after each emotional narrative, with all four comparisons remaining significant after Holm correction in both the paired-sample t-test and the Wilcoxon signed-rank test.

The multiple linear regression model for average state empathy was statistically significant, $F(6, 98) = 5.974$, $p < 0.001$, explaining 26.8% of the variance ($R^2 = 0.268$; adjusted $R^2 = 0.223$). Residuals did not significantly deviate from normality (Shapiro–Wilk $W = 0.986$, $p = 0.325$), the Breusch–Pagan test showed no evidence of heteroscedasticity ($BP = 2.278$, $p = 0.892$), VIF values ranged from 1.14 to 5.46 with no severe multicollinearity (Table 11, Appendix D), and visual inspection of the residuals-versus-fitted plot showed no systematic pattern.

Table 3 summarizes the regression results. Cognitive empathy and perceived presence in the virtual environment were the only variables significantly associated with average state empathy in the multivariable model, with both 95% confidence intervals excluding zero. Correlation analysis (Appendix B) further revealed that all moderate correlations remained significant after Benjamini–Hochberg correction; among the weak correlations judged theoretically relevant to the research questions, 23 remained significant. Table 4 summarizes the Bonferroni-corrected pairwise Wilcoxon signed-rank post-hoc comparisons, showing that happiness differs significantly from all three negative emotions across arousal, valence, and empathy, while no statistically significant differences in arousal were detected among the three negative emotions.

Table 2: Significance Summary: Paired-Sample t-Tests (Before and After Holm Correction)

Pair	Sig. (Uncorrected)	Sig. (Holm)	Wilcoxon Sig. (Holm)
Rollercoaster Arousal vs Sadness Arousal	Yes	Yes	Yes
Rollercoaster Arousal vs Anxiety Arousal	Yes	Yes	Yes
Rollercoaster Arousal vs Happiness Arousal	Yes	Yes	Yes
Rollercoaster Arousal vs Anger Arousal	Yes	Yes	Yes

Table 3: Significance Summary: Multiple Linear Regression

Factor	Significant ($p < 0.05$)	95% CI excludes 0
(Intercept)	Yes	Yes
Age	No	No
Education	No	No
Gender	No	No
Affective Empathy	No	No
Cognitive Empathy	Yes	Yes
Presence in VE	Yes	Yes

Table 4: Significance Summary: Pairwise Wilcoxon Signed-Rank Post-Hoc Comparisons

Pair	Arousal	Valence	Empathy
Sadness vs Anxiety	No	Yes	No
Sadness vs Happiness	Yes	Yes	Yes
Sadness vs Anger	No	No	Yes
Anxiety vs Happiness	Yes	Yes	Yes
Anxiety vs Anger	No	Yes	No
Happiness vs Anger	Yes	Yes	Yes

4.2 Discussion of the Research Questions

The statistical analyses require a revisit to each of the research questions formulated in Section 3.1.

***RQ1:* Is trait empathy associated with state empathy experienced during the virtual reality scenarios?** Correlation analysis revealed a weak negative association between cognitive trait empathy and average state empathy that remained significant after BH correction ($r = -0.247$, $p_{\text{BH}} = 0.043$). Parallel to this, the corresponding association for affective trait empathy did not survive correction ($r = -0.239$, $p_{\text{BH}} = 0.052$). The regression model showed that cognitive empathy was negatively associated with average state empathy ($B = -0.3306$), whereas perceived presence in the virtual environment was positively associated with it ($B = 0.1068$), after accounting for the other variables in the model. Trait empathy was not consistently associated with state empathy experienced during the VR scenarios in this sample. This negative association is counter-intuitive and may reflect a more analytical, less affectively engaged response style among individuals with higher cognitive trait empathy; given the weak effect size, we flag it as unexpected and worth further investigation.

***RQ2:* Are educational level and field of expertise associated with differences in trait or state empathy?** Education showed only a weak, BH-nonsignificant correlation with average state empathy ($r = -0.193$, $p_{\text{BH}} = 0.127$). Field of expertise showed a similar pattern: uncorrected STEM vs. Human Services differences in trait empathy (Hedges' $g = 0.75$ and 0.54) did not survive Holm correction. Overall, education and field of expertise showed weak evidence of independent contribution to empathy after correction for multiple testing.

***RQ3:* Are gender and age associated with differences in empathy and emotional responses during the VR experience?** Age was not significantly associated with state empathy after the adjustment for the other covariates, showing a weak positive correlation with affective empathy ($r = 0.276$, $p_{\text{BH}} = 0.0223$). Age itself differed significantly between genders (Hedges' $g = 1.16$), with male participants older on average, and an unadjusted gender difference in average state empathy was observed as well; however, gender was not significant after adjustment. Nominal differences in trait empathy and emotion-specific

measures did not survive correction. Given the age–gender confound and the unequal group sizes in this sample, these findings should be interpreted with caution.

***RQ4:* Do sadness, anxiety, happiness, and anger differ in empathy, arousal, and valence, and which emotional conditions show the most distinct response patterns?** The Friedman and Wilcoxon signed-rank analyses identified happiness as significantly different from the three negative emotions across arousal, valence, and empathy. The magnitude of the overall condition differences varied by measure, with Kendall's $W = 0.714$ for valence, compared with $W = 0.277$ for arousal and $W = 0.180$ for empathy. Among the three negative emotions, arousal did not differ significantly between any pair, whereas valence differed between sadness and anxiety and between anxiety and anger, and empathy differed between sadness and anger. Happiness was associated with the highest empathy ratings, anger with the lowest, and sadness and anxiety with intermediate values. Thus, the emotional-condition effect was strongest for valence and weaker for arousal and empathic responses.

***RQ5:* Does self-reported arousal differ between emotion-based and non-empathic VR stimuli?** Self-reported arousal differed significantly between the Rollercoaster condition and each of the four emotion-based VR scenarios. All four comparisons remained significant after Holm correction, and the Wilcoxon signed-rank sensitivity analysis yielded the same significance conclusions. Effect sizes ranged from small for anger ($g = 0.33$) to large for happiness ($g = 1.50$), indicating that the magnitude of the difference in self-reported arousal varied across emotions.

***RQ6:* Which characteristics of the virtual reality experience are associated with state empathy?** Feeling of presence was significantly associated with average state empathy ($B = 0.107$, 95% CI [0.044, 0.169]) after adjustment. Captivated by VE correlated weakly with average state empathy ($r = 0.306$) and differed by narrative condition after Holm correction ($g = -1.16$), while differences in empathy and realism did not survive correction. Presence thus appears the most consistently associated VR-related factor.

5 Conclusions and Limitations

The results indicate that empathic responses in immersive VR were more consistently associated with characteristics of the experienced scenario than with the demographic variables examined. Perceived presence was positively associated with average state empathy, whereas the effect of narrative condition was less consistent after correction for multiple comparisons. Self-reported arousal during the Rollercoaster condition differed significantly from that reported after each of the four emotion-based VR scenarios, with effect sizes ranging from small ($g = 0.33$) to large ($g = 1.50$). The emotional conditions also differed in empathy, arousal, and valence, with happiness showing the most distinct response pattern and trait empathy showing no consistent association with state empathy. Several nominally significant associations disappeared after correction for multiple comparisons, underscoring the importance of controlling for multiple testing when interpreting this dataset.

The sample consisted of only 105 participants, with substantial demographic imbalances (most respondents 19–20 years old; 76% women vs. 24% men), limiting generalizability. As a secondary analysis of data collected for a different purpose, the precision of some analyses may be limited, and given the observational design, reported relations should be interpreted as associations rather than causal effects.

Acknowledgements

The setup for data collection and the data collection itself were supported by the Slovenian Research and Innovation Agency (ARIS) through the Young Researcher Ph.D. Program, grant PR-12879.

Funding

The work of Jasmina Angelevska Kostadinoska and Katerina Hadzi-Velkova Saneva was supported by the Ministry of Education and Science of North Macedonia under Grant No. 15-6171/25.

References

[1] Y. Benjamini and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 1, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

[2] M. M. Bradley and P. J. Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25, 1, 49–59. doi:10.1016/0005-7916(94)90063-9.

[3] M. H. Davis. 1980. A Multidimensional Approach to Individual Differences in Empathy. Tech. rep. JSAS Catalog of Selected Documents in Psychology, 85.

[4] James D. Evans. 1996. *Straightforward Statistics for the Behavioral Sciences*. Brooks/Cole Publishing, Pacific Grove, CA.

[5] M. Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32, 200, 675–701. doi:10.1080/01621459.1937.10503522.

[6] Larry V. Hedges. 1981. Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 2, 107–128. doi:10.3102/10769986006002107.

[7] Fernanda Herrera, Jeremy Bailenson, Erika Weisz, Elise Ogle, and Jamil Zaki. 2018. Building long-term empathy: a large-scale comparison of traditional and virtual reality perspective-taking. *PLoS ONE*, 13, 10, e0204494. doi:10.1371/journal.pone.0204494.

[8] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 2, 65–70.

[9] E. Kizhevska, K. Šparemblek, and M. Luštrek. 2024. Protocol of the study for predicting empathy during VR sessions using sensor data and machine learning. *PLoS ONE*, 19, 7, e0307385. doi:10.1371/journal.pone.0307385.

[10] Emilija Kizhevska, Hristijan Gjoreski, and Mitja Luštrek. 2025. Predicting empathy and other mental states during VR sessions using sensor data and machine learning. *Sensors*, 25, 18, 5766. doi:10.3390/s25185766.

[11] H. Levene. 1960. Robust tests for equality of variances. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann, editors. Stanford University Press, 278–292.

[12] D. C. Montgomery and G. C. Runger. 2014. *Applied Statistics and Probability for Engineers*. (6th ed.). Wiley.

[13] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, 15, 5, 625–632. doi:10.1007/s10459-010-9222-y.

[14] K. Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58, 240–242. doi:10.1098/rspl.1895.0041.

[15] R. L. E. P. Reniers, R. Corcoran, R. Drake, N. M. Shryane, and B. A. Völlm. 2011. The QCAE: a questionnaire of cognitive and affective empathy. *Journal of Personality Assessment*, 93, 1, 84–95. doi:10.1080/00223891.2010.528484.

[16] S. S. Shapiro and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52, 3/4, 591–611. doi:10.1093/biomet/52.3-4.591.

[17] L. Shen. 2010. On a scale of state empathy during message processing. *Western Journal of Communication*, 74, 5, 504–524. doi:10.1080/10570314.2010.512278.

[18] Mel Slater. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364, 1535, 3549–3557. doi:10.1098/rstb.2009.0138.

[19] C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15, 1, 72–101. doi:10.2307/1412159.

[20] Student. 1908. The probable error of a mean. *Biometrika*, 6, 1, 1–25. doi:10.1093/biomet/6.1.1.

[21] F. Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1, 6, 80–83. doi:10.2307/3001968.

Appendix A

Table 5: Descriptive Statistics

Variable	Mean	SD	Median	IQR
Age	22.43	5.31	20	6
Education	3.42	0.74	3	1
Pres. in VE	6.43	1.63	7	1
Capt. by VE	6.01	1.77	6	2
Real. of VE	6.09	1.58	6	2
Sadness Val.	2.97	0.58	3	0
Anxiety Val.	2.65	0.60	3	1
Happiness Val.	0.71	0.72	1	1
Anger Val.	3.04	0.71	3	1
Sadness Arou.	2.30	0.77	2	1
Anxiety Arou.	2.14	0.87	2	2
Happiness Arou.	1.18	1.00	1	2
Anger Arou.	2.37	0.98	2	1
RC Arou.	2.71	1.02	3	1.25
Cog. Emp. (trait)	1.95	0.35	2	0
Prox. Resp.	1.94	0.50	2	0.75
Periph. Resp.	2.75	0.34	2.75	0.50
Aff. Emp. (trait)	2.12	0.33	2	0
Sadness Emp.	2.42	0.77	2.70	1
Anxiety Emp.	2.24	0.90	2.30	1.30
Happiness Emp.	2.85	0.75	3	1
Anger Emp.	2.04	0.83	2	1
Avg. Emp.	2.40	0.55	2.40	0.70

Abbreviations: Emp. = empathy; Arou. = arousal; Val. = valence; Avg. = average; Cog. = cognitive; Aff. = affective; VE = virtual environment; RC = Rollercoaster; Resp. = responsivity; Pres. = presence; Capt. = captivated; Real. = realism; Prox. = proximal; Periph. = peripheral.

Appendix B

Table 6: Selected Moderate Correlations ($0.40 \le |r| < 0.60$) [4]

Var. 1	Var. 2	r	p	p_{adj} (BH)
Anger Arou.	Anger Emp.	0.540	<0.001	<.001
Pres. in VE	Sadness Emp.	0.515	<0.001	<.001
Anxiety Arou.	Anxiety Emp.	0.509	<0.001	<.001
Capt. by VE	Sadness Emp.	0.484	<0.001	<.001
Sadness Arou.	Sadness Emp.	0.484	<0.001	<.001
Sadness Arou.	Avg. Emp.	0.457	<0.001	<.001
Anger Arou.	Avg. Emp.	0.437	<0.001	<.001
Happiness Val.	Happiness Emp.	−0.435	<0.001	0.0001

Abbreviations: Var. = variable; Emp. = empathy; Arou. = arousal; Val. = valence; Avg. = average; Pres. = presence; Capt. = captivated; VE = virtual environment.

Table 7: Selected Weak Correlations ($0.20 \le |r| < 0.40$) [4]

Var. 1	Var. 2	r	p	p_{adj} (BH)
Real. of VE	Sadness Emp.	0.382	<0.001	0.0009
Pres. in VE	Avg. Emp.	0.381	<0.001	0.0009
Education	Aff. Emp.	0.373	<0.001	0.0009
Sadness Emp.	Anger Emp.	0.357	0.0002	0.0017
Real. of VE	Avg. Emp.	0.332	0.0005	0.0038
Capt. by VE	Avg. Emp.	0.306	0.0015	0.0097
Anger Arou.	Happiness Emp.	0.306	0.0015	0.0097
Anxiety Arou.	Avg. Emp.	0.296	0.0022	0.0135
Sadness Arou.	Anxiety Arou.	0.295	0.0023	0.0135
Happiness Emp.	Anger Emp.	0.294	0.0023	0.0135
Sadness Arou.	Anger Emp.	0.286	0.0031	0.0174
Sadness Emp.	Anxiety Emp.	0.285	0.0032	0.0176
Pres. in VE	Sadness Arou.	0.281	0.0037	0.0195
Age	Aff. Emp.	0.276	0.0044	0.0223
Age	Prox. Resp.	0.261	0.0072	0.0331
Sadness Arou.	Anger Arou.	0.260	0.0074	0.0334
Sadness Arou.	Anxiety Emp.	0.259	0.0077	0.0342
Anger Arou.	Sadness Emp.	0.256	0.0085	0.0357
Anxiety Emp.	Happiness Emp.	0.255	0.0086	0.0357
Sadness Emp.	Happiness Emp.	0.248	0.0106	0.0426
Cog. Emp.	Avg. Emp.	-0.247	0.0111	0.0432
Anxiety Emp.	Anger Emp.	0.247	0.0109	0.0431
Real. of VE	Happiness Emp.	0.244	0.0121	0.0457

Abbreviations: Var. = variable; Emp. = empathy; Arou. = arousal; Avg. = average; Aff. = affective; Cog. = cognitive; Real. = realism; Pres. = presence; Capt. = captivated; Prox. Resp. = proximal responsivity; VE = virtual environment.

Appendix C

Table 8: Shapiro–Wilk Tests and 95% Confidence Intervals for the Mean

Variable	W	p-value	95% CI (mean)
Age	0.68964	< 0.00001	[21.4134, 23.4438]
Education	0.61391	< 0.00001	[3.2767, 3.5613]
Pres. in VE	0.85961	< 0.00001	[6.1171, 6.7400]
Capt. by VE	0.93690	0.00008	[5.6704, 6.3487]
Real. of VE	0.92303	0.00001	[5.7843, 6.3871]
Sadness Val.	0.74261	< 0.00001	[2.8606, 3.0823]
Anxiety Val.	0.74813	< 0.00001	[2.5321, 2.7632]
Happiness Val.	0.76475	< 0.00001	[0.5732, 0.8499]
Anger Val.	0.82004	< 0.00001	[2.9030, 3.1731]
Sadness Arou.	0.84586	< 0.00001	[2.1477, 2.4427]
Anxiety Arou.	0.84881	< 0.00001	[1.9763, 2.3094]
Happiness Arou.	0.87218	< 0.00001	[0.9901, 1.3718]
Anger Arou.	0.90136	< 0.00001	[2.1834, 2.5595]
RC Arou.	0.88498	< 0.00001	[2.5153, 2.9077]
Cog. Emp. (trait)	0.47313	< 0.00001	[1.8854, 2.0194]
Prox. Resp.	0.95308	0.00096	[1.8443, 2.0367]
Periph. Resp.	0.94196	0.00017	[2.6799, 2.8105]
Aff. Emp. (trait)	0.38542	< 0.00001	[2.0605, 2.1871]
Sadness Emp.	0.94717	0.00038	[2.2715, 2.5666]
Anxiety Emp.	0.96646	0.00922	[2.0670, 2.4130]
Happiness Emp.	0.92815	0.00003	[2.7107, 2.9979]
Anger Emp.	0.98087	0.13410	[1.8822, 2.2054]
Avg. Emp.	0.97844	0.08478	[2.2967, 2.5110]
Capt. by VE(no narrative)	0.91723	0.13260	[3.3666, 5.4569]
Real. of VE(no narrative)	0.90925	0.09716	[4.6742, 6.0317]

Abbreviations: Emp. = empathy; Arou. = arousal; Val. = valence; Avg. = average; Cog. = cognitive; Aff. = affective; Real. = realism; Pres. = presence; Capt. = captivated; Prox. Resp. = proximal responsivity; Periph. Resp. = peripheral responsivity; RC = Rollercoaster; VE = virtual environment.

Appendix D

Table 9: Independent-Group Comparisons with Effect Sizes and Holm-Adjusted p-Values

Variable	Group	Levene (F/p)	t	p	Hedges' g	p (Holm)
Avg. Emp.	M vs F	0.0004/0.9845	-3.3362	0.0012	-0.759	0.0142
Aff. Emp.	M vs F	19.72/2.3×10^{-5}	3.0693	0.0048	1.010	0.0509
Cog. Emp.	M vs F	0.0043/0.9478	2.8320	0.0056	0.644	0.0509
Age	M vs F	15.26/0.0002	3.8584	0.0006	1.163	0.0077
Sadness Emp.	M vs F	0.1637/0.6866	-2.7825	0.0064	-0.633	0.0509
Anger Emp.	M vs F	0.0241/0.8768	-2.8955	0.0046	-0.659	0.0509
Anger Arou.	M vs F	0.0060/0.9386	-2.2040	0.0298	-0.501	0.1488
Sadness Val.	M vs F	1.018/0.3153	-2.1253	0.0360	-0.483	0.1488
Avg. Emp.	No Narr. vs Narr.	0.0505/0.8226	-2.0729	0.0407	-0.545	0.1488
Sadness Emp.	No Narr. vs Narr.	0.5792/0.4484	-2.8873	0.0047	-0.759	0.0509
Capt. by VE	No Narr. vs Narr.	1.529/0.2191	-4.4023	2.6×10^{-5}	-1.158	0.0004
Real. of VE	No Narr. vs Narr.	0.4130/0.5219	-2.1297	0.0356	-0.560	0.1488
Aff. Emp.	HS vs STEM	7.972/0.0067	-2.9848	0.0058	-0.752	0.0509
Cog. Emp.	HS vs STEM	0.3038/0.5838	-2.0145	0.0490	-0.536	0.1488

Note. Welch's t-test was used when Levene's test indicated unequal variances ($p < 0.05$); Student's t-test was used otherwise. Positive Hedges' g values indicate higher scores in the first-listed group; negative values indicate higher scores in the second-listed group. Abbreviations: M = male; F = female; Narr. = narrative; HS = Human Services; Emp. = empathy; Arou. = arousal; Val. = valence; Avg. = average; Cog. = cognitive; Aff. = affective; Real. = realism; Capt. = captivated; VE = virtual environment.

Table 10: Paired-Sample Tests for Rollercoaster versus Emotion-Based Arousal

Pair	t-test ($t/p/p_{\text{Holm}}$)	Hedges' g	Wilcoxon ($V/p/p_{\text{Holm}}$)
RC Arou. vs Sadness Arou.	3.61/0.0005/0.0010	0.463 (small)	2059.0/0.0005/0.0009
RC Arou. vs Anxiety Arou.	4.64/1.0×10^{-5}/3.0×10^{-5}	0.603 (medium)	2248.5/2.9×10^{-5}/8.6×10^{-5}
RC Arou. vs Happiness Arou.	11.35/< 2.2×10^{-16}/< 8.8×10^{-16}	1.495 (large)	3543.0/2.7×10^{-14}/< 0.0001
RC Arou. vs Anger Arou.	2.68/0.0085/0.0085	0.325 (small)	1843.0/0.0110/0.0110

Abbreviations: RC = Rollercoaster; Arou. = arousal. Values reported as statistic/uncorrected p/Holm-corrected p.

Table 11: Multiple Linear Regression Coefficients for Average Empathy

Factor	VIF	B	SE	t	p-value	95% CI
(Intercept)	–	2.7981	0.5400	5.182	1.18 × 10^{-6}	[1.727, 3.870]
Age	4.977	0.0258	0.0201	1.283	0.2024	[-0.014, 0.066]
Education	5.458	-0.1503	0.1503	-1.000	0.3199	[-0.448, 0.148]
Gender	1.520	-0.2242	0.1379	-1.626	0.1072	[-0.498, 0.049]
Aff. Emp.	1.309	-0.2107	0.1654	-1.273	0.2059	[-0.539, 0.118]
Cog. Emp.	1.135	-0.3306	0.1455	-2.271	0.0253	[-0.619, -0.042]
Pres. in VE	1.149	0.1068	0.0315	3.390	0.0010	[0.044, 0.169]

Note. Unstandardized regression coefficients (B) are reported. VIF = variance inflation factor. Abbreviations: Aff. Emp. = affective empathy; Cog. Emp. = cognitive empathy; Pres. = presence; VE = virtual environment.

Table 12: Friedman Tests by Emotion, with Kendall's W

	χ^2_F	p-value	Kendall's W
Arousal	87.356	$< 2.2 \times 10^{-16}$	0.277
Valence	222.74	$< 2.2 \times 10^{-16}$	0.714
Empathy	56.773	2.874×10^{-12}	0.180

Table 13: Wilcoxon Signed-Rank Post-Hoc Pairwise Comparisons by Emotion

	Arousal		Valence		Empathy	
Pair	V	p_{Bonf}	V	p_{Bonf}	V	p_{Bonf}
Sadness vs Anxiety	1461.0	0.6096	1447.5	0.0003	2707.0	0.9963
Sadness vs Happiness	3614.0	< 0.0001	5253.0	< 0.0001	1064.5	0.0001
Sadness vs Anger	948.5	1.0000	382.0	1.0000	3371.5	0.0003
Anxiety vs Happiness	3249.0	< 0.0001	4836.5	< 0.0001	797.0	< 0.0001
Anxiety vs Anger	891.0	0.4146	359.0	0.0001	2449.5	0.2455
Happiness vs Anger	291.0	< 0.0001	0.0	< 0.0001	3897.0	< 0.0001

Note. p_{Bonf} = Bonferroni-corrected p-value, applied across the six pairwise comparisons within each measure.