UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Appraisal of Emotional Responses, Mental Workload, and Debriefing Feedback during in-person and Remotely Facilitated Healthcare Simulations: A Natural Experiment Study

Gustavo de Oliveira Almeida

Orientadores
Carlos Eduardo Ribeiro de Mello
UNIRIO
Winslow Burleson
University of Arizona

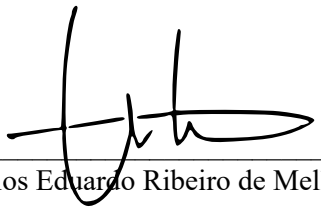
RIO DE JANEIRO, RJ – BRASIL
Maio de 2022

Appraisal of emotional responses, mental workload, and debriefing feedback during in-person and remotely facilitated healthcare simulations: a natural experiment study

GUSTAVO DE OLIVEIRA ALMEIDA

TESE DE DOUTORADO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE DOUTOR PELO PROGRAMA DE PÓS GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:



Prof. Dr. Carlos Eduardo Ribeiro de Mello (orientador) – UNIRIO

Prof. Dr. Winslow Burleson (co-orientador) - University of Arizona

Profa. Dra. Janine Hinton – University of Arizona

Prof. Dr. Leandro Guimarães Marques Alvim - UFRuralRJ

Prof. Dr. Márcio de Oliveira Barros – UNIRIO

Prof. Dr. Sean Wolfgand Matsui Siqueira – UNIRIO

RIO DE JANEIRO, RJ - BRASIL

MAIO DE 2022.

Catalogação informatizada pelo(a) autor(a)

de Oliveira Almeida, Gustavo
d447 Appraisal of Emotional Responses, Mental Workload, and Debriefing Feedback during in-person and Remotely Facilitated Healthcare Simulations: A Natural Experiment Study / Gustavo de Oliveira Almeida. -- Rio de Janeiro, 2022.
406

Orientador: Carlos Eduardo Ribeiro de Mello.
Coorientador: Winslow Winslow Burleson.
Tese (Doutorado) - Universidade Federal do Estado do Rio de Janeiro, Programa de Pós-Graduação em Informática, 2022.

1. interface humano computador.. 2. computação afetiva. 3. simulação médica. 4. experimentos. I. Ribeiro de Mello, Carlos Eduardo , orient. II. Winslow Burleson, Winslow, coorient. III. Título.

Dedication

Of all the people who helped me complete this work, I dedicate this thesis to Lined Ramos and Jacob Almeida, for each taught me how to love, each in a different and complementary way.

Acknowledgements

Completing a thesis is intense work, with countless hours devoted to a work that is never complete or perfect. I thought doing it the second time would be easier, but it was not. Doctoral research, as a large project, is never a task that one can do without significant support in all realms of life.

There are so many whom I would like to thank, but the thesis is already too long to list the name of so many amazing people I met during this journey, including all my PhD colleagues and professors at UNIRIO, NYU and the University of Arizona. A special thank you to my thesis committee and their thought-provoking suggestions.

However, I would like to name a few who were instrumental in this process. I am incredibly grateful for your help and support. I want to thank Dr. Jeremy Rowe for the support, welcome, and ideas during my term at NYU and for introducing me to the critical actors in healthcare simulation.

A big thanks to the University of Arizona (ASTEC) and Yale (YCMS) teams, directors, and staff, for allowing me to collect data during a challenging time (COVID-19). Namely, I would like to thank Melissa Joseph (Yale), David Biffar (ASTEC) for the support. For ASTEC, in my home institution, I could not do it without the tireless support from Lucas Grijalva in our weekly data collection procedure. Thank you so much Lucas!

Another fundamental support I received was from our program Coordinator and my advisor Prof. Mello, who made possible the completion of this work, when I thought I could not complete it. Thank you, Prof. Carlos Eduardo. I'll always appreciate your help. Closing the academic acknowledgment section, I want to thank Professor Win Burleson for his support, mentorship, opportunities, and friendship. He received me as part of his team back at NYU, and invited me to Arizona, for which I am forever grateful. Your guidance during this thesis and my career was more than I could ask for and would not be possible without you!

I would like to thank my family. By my family, I mean my wife Lined and my son Jacob. You have sacrificed so much that this could become a reality. Supporting a spouse during a doctoral thesis is not an easy feat. Lined did it twice, with our year-old baby Jacob, in a foreign country in turbulent and uncertain times during the COVID-19. Without you, I would not have been able to complete this project. I can't thank you enough.

RESUMO

A simulacao realistica é um metodo educacional consolidado para o ensino nas profissoes da saúde. Durante a COVID-19 vários dos maiores centros de simulação adotaram a telesimulacao como uma alternativa educacional para a continuidade da formação de profissionais da area de saúde. A objetivo principal da presente pesquisa foi verificar a nao inferioridade da telesimulacao como um método educacional em comparação com a simulação realizada em pessoa, utilizando como parametros de comparação as respostas emocionais e carga cognitiva nos participantes. O estudo utilizou uma abordagem de experimentos naturais de design misto (pre-post teste e grupos nao equivalentes). O instrumento de pesquisa foi composto por um questionário sócio demográfico e as escalas: Felt Arousal Scale – Feldman, Barret and Russell (2012), a NASA Task Load Index (NASA-TLX) – (Hart, 2006), apreensao emocional–baseada em Scherer & Moors' (2019 e a Simulation Design Scale (SDS student version; Jeffries, 2016). Um total de 134 participantes foram selecionados em sessões de simulação conduzidas remotamente e pessoalmente em dois centros de pesquisa nos EUA. Os dados foram analisados nos Softwares: SPSS 27, PASS 2021, NCSS 2021, JASP 0.361 E Smart PLS 3.3 para a produção de Estatísticas Descritivas, Análises de Confiabilidade, Análise de Variância de Medidas Repetidas (ANOVA), Gráficos, Análise Fatorial Confirmatória e Testes de não inferioridade paramétricos e nao paramétricos utilizando Testes t, e versão nao parametricas destes testes. Os resultados indicaram que telesimulacao é um metodo não inferior comparado com a simulacao em pessoa, na maioria das hipoteses estudadas. Implicações práticas e teóricas, além de sugestões para pesquisas futuras são apresentadas.

Palavras-chave: simulação realistíca; manequins robotícos; interface humano computador; telesimulação;emoções; carga cognitiva; não-inferioridade

ABSTRACT

Healthcare simulation is an established educational method for training health professionals. During COVID-19, most medical simulation centers around the world adopted telesimulation as an educational alternative for the continuity of training of professionals in the health area. The present research objective was to verify the non-inferiority of telesimulation as an educational method compared to the simulation in person, using the participants' emotional responses and cognitive load as comparison parameters. The study used a mixed-design natural experiments approach (pre-post and non-equivalent groups). The research instrument consisted of a sociodemographic questionnaire and the following scales: Felt Arousal Scale – Feldman, Barret and Russell (2012), NASA Task Load Index (NASA-TLX) – (Hart, 2006), emotional apprehension– based on Scherer & Moors' (2019 and the Simulation Design Scale (SDS student version; Jeffries, 2016). One hundred thirty-four participants were selected in simulations conducted remotely and in person at two research centers in the US. Data were analyzed using the software: SPSS 27, PASS 2021, NCSS 2021, JASP 0.361, and Smart PLS 3.3 for the production of Descriptive Statistics, Reliability Analyses, Repeated Measures Analysis of Variance (ANOVA), Graphs, Confirmatory Factor Analysis, and Parametric and Nonparametric non-inferiority testing using t-tests and their non-parametric equivalents. Results indicated that telesimulation is a non-inferior method compared to the in-person simulation for most of the variables studied. Practical and theoretical implications, as well as suggestions for research future research, are presented.

Keywords: healthcare simulation; robotic mannikin; HCI; telesimulation; emotions; cognitive workload; mental workload; non-inferiority

Contents

Chapter 1. Introduction 20

1.1. Research Motivation and Context 20

1.2. Research Questions and Objectives 26

1.3. Personal Motivation 34

1.4. Document Structure 36

Chapter 2. Preliminary Studies 37

2.1. Bibliometric Studies 38

2.1.1. Bibliometric Analysis - Objectives and Questions 42

2.1.2. Bibliometric data analysis 49

2.2. Scoping Review 74

2.2.1. Scoping Review Protocol 78

2.2.2. Scoping Review Objectives 79

2.2.3. Scoping Review Methods 80

2.2.4. Execution of the scoping review 88

2.2.5. Summarization of Results 93

2.2.6. Conclusions for the telesimulation scoping review 101

Chapter 3. Literature Review 103

3.1. Simulation 103

3.1.1. Healthcare Simulation 105

3.1.2. Debriefing 114

3.1.3. COVID-19 Impact on the Healthcare Simulation 116

3.2. Emotions 120

3.2.1. Arousal as a Component of Emotions 124

3.2.2. Emotional Appraisal Process 125

3.2.3. Cognitive Workload 127

3.3. Emotional Responses in Learning 130

3.3.1. Emotions in Human-Computer Interaction 132
3.3.2. Healthcare simulation and emotions 136
3.4. Quasi-Experiments and Natural Experiments 137
3.5. Non-Inferiority Studies 143
Chapter 4. Research Methodology and Methods 149
4.1. Methodological Considerations 150
4.2. Research Design 151
4.3. Research Approach 156
4.4. Research Location 158
4.4.1. ASTEC 158
4.4.2. YCMS 161
4.5. Ethical considerations and IRB evaluation 163
4.6. Research Instruments 166
4.7. Research Procedures 173
4.7.1. Participants and Selection Procedure 182
4.7.2. Software for Data Analysis 182
4.7.3. Statistical Power Analysis 183
Chapter 5. Data Analysis 189
5.1. Sample Description 190
5.2. Exploratory Data Analysis 197
5.3. Missing Data Analysis 198
5.4. Outlier analysis 209
5.5. Data normality tests 213
5.6. Common Method Variance Bias 222
5.7. Confirmatory Factor and Reliability Analysis 225
5.8. Descriptive Statistics 239
5.9. Hypothesis Tests 245

5.9.1. Non-inferiority margins for hypothesis testing 245
5.9.2. Hypothesis 1 248
5.9.3. Hypothesis 2 257
5.9.4. Hypothesis 3 265
5.9.5. Hypothesis 4 280
5.9.6. Hypothesis 5 289
5.9.7. Hypothesis 6 301
5.9.8. Hypothesis 7 312
5.9.9. Hypothesis 8 319
Chapter 6. Discussion of Results and Conclusions 323
6.1. Summary of Results 323
6.1.1. Hypothesis 1 324
6.1.2. Hypothesis 2 327
6.1.3. Hypothesis 3 329
6.1.4. Hypothesis 4 331
6.1.5. Hypothesis 5 333
6.1.6. Hypothesis 6 335
6.1.7. Hypothesis 7 337
6.1.8. Hypothesis 8 338
6.2. Final Conclusions 339
6.3. Limitations and recommendations for future research 341

List of Figures

Figure 1: List of study constructs and their categorizations 29
Figure 2: Number of articles indexed in Scopus (English) – 1864-2021 39
Figure 3: Prisma Diagram 48
Figure 4: Articles published from 1977 to 2021 51
Figure 5: Collaborative Network 59
Figure 6: Top 20 most relevant publication sources 61
Figure 7: Growth of top five journals 61
Figure 8: Co-citation network of sources – Medical Education and Nursing clusters 62
Figure 9: Top 20 countries active in collaborative research 64
Figure 10: Most Relevant authors 66
Figure 11: Productivity of the authors over years 67
Figure 12: Top 20 authors – ranked by number of total citations (TC) index 68
Figure 13: Citation network among top 30 researchers 69
Figure 14: Collaboration network among top 50 researchers 70
Figure 15: Top 10 Trigram words – Title 71
Figure 16: Top 10 Trigram words – Title 72
Figure 17: Top ten most frequent words 72
Figure 18: Most frequently words – Keyword Plus Field 73
Figure 19: Evolution in the number of scoping review studies 77
Figure 20: Literature Review Process 79
Figure 21: Custom Form to Screen References – Level 2 86
Figure 22: Level 1 – Preliminary screening of References 89
Figure 23: PRISM Diagram for the Scoping Review 91
Figure 24: Task Trainer (Part Task) 111
Figure 25:VR Simulation 111
Figure 26: Example of a haptic feedback system 112
Figure 27: Example of Full-scale simulation in a simulation center 112
Figure 28: Example of full-scale simulator – newborn - in a simulation center 113
Figure 29: Example of Full-scale simulator – female mannikin - in a simulation center 113
Figure 30: Margins in superiority, equivalence, and non-inferiority studies. 145
Figure 31: Location of ASTEC and YCMS in the Continental US 158

Figure 32: Technologies Available in ASTEC 159
Figure 33: Simulation involving Robotic Mannekins in ASTEC 160
Figure 34: Medical Students 162
Figure 35: Concept of Presence Tool – Qualtrics 172
Figure 36: Briefing Session before the start of a simulation scenario 173
Figure 37: Research Moments and associated activities 174
Figure 38: Relationship between the instructor's survey and the participants' survey . 175
Figure 39: Sample setup for in-person data collection 177
Figure 40: Prompt for user to engage in the debriefing 179
Figure 41: Design of the Natural Experiment 180
Figure 42: Summary of Missing Values – Block 1 200
Figure 43: Missing Value Pattern – Block 1 202
Figure 44: Missing Value Pattern – Pattern Distribution – Block 1 203
Figure 45: Summary of Missing Values – Block 2 203
Figure 46: Missing Value Pattern – Block 2 205
Figure 47: Missing Value Pattern – Pattern Distribution – Block 2 205
Figure 48: Summary of Missing Values – Block 3 206
Figure 49: Summary of Missing Values – Block 4 206
Figure 50: Missing Value Pattern – Block 4 208
Figure 51: Missing Value Pattern – Pattern Distribution – Block 4 208
Figure 52: Sample of Frequency and PP-PLOT to check for normality 214
Figure 53: CFA for Emotion State Scale 227
Figure 54: CFA for Emotion Appraisal Scale 229
Figure 55: CFA for Fidelity (Realism) and Feedback dimensions 233
Figure 56: CFA for NASA-TLX 235
Figure 57: Emotional Responses during Baseline, Simulation and Debriefing – In-person 249
Figure 58: Cheerfulness during Baseline, Simulation, and Debriefing – In-person 252
Figure 59: Invigoration during Baseline, Simulation and Debriefing – In-person 253
Figure 60: Tranquility during Baseline, Simulation and Debriefing – In-person 253
Figure 61: Arousal Scores during Baseline, Simulation and Debriefing – In-person .. 254
Figure 62: Arousal during Baseline, Simulation, and Debriefing – In-person 255
Figure 63: Emotional Responses during Baseline, Simulation and Debriefing – Telesimulation 258

Figure 64: Emotional Responses during Baseline, Simulation and Debriefing – Telesimulation 261
Figure 65: Invigoration during Baseline, Simulation and Debriefing – telesimulation 261
Figure 66: Tranquility during Baseline, Simulation and Debriefing – in-person..... 262
Figure 67: Arousal during Baseline, Simulation, and Debriefing – telesimulation 262
Figure 68: Arousal during Baseline, Simulation and Debriefing – telesimulation 264
Figure 69: Telesimulation and in-person arousal during baseline, simulation, and debriefing..... 267
Figure 70: Cheerfulness temporal pattern – Telesimulation vs In-Person 270
Figure 71: Invigoration temporal pattern – Telesimulation vs In-Person 270
Figure 72: Tranquility temporal pattern – Telesimulation vs In-Person 271
Figure 73: NASA-TLX Telesimulation vs. In-person..... 281
Figure 74: Raincloud plots for the variables of NASA mental workload vs. Simulation method 287
Figure 75: Arousal Response During Simulation and Debriefing – Low and Medium Complexity 290
Figure 76: Emotional Appraisal Dimensions During Simulation and Debriefing – Low and Medium Complexity 292
Figure 77: Raincloud plots for the variables of Emotional Responses vs Complexity 294
Figure 78: NASA Mental Workload Scale and subdimensions – Low and Medium Complexity 295
Figure 79: Raincloud plots for the variables of NASA mental workload vs Complexity 297
Figure 80: Emotion Appraisal Dimensions– Low and Medium Complexity..... 299
Figure 81: Emotion Appraisal Dimensions– Telesimulation and In-person 302
Figure 82: Raincloud plots for the variables of Emotion Appraisal..... 310
Figure 83: Perception of Fidelity – Telesimulation vs. In-person..... 313
Figure 84: Perception of Presence (0-100) – Telesimulation vs. In-person 315
Figure 85: Raincloud plots for the variables of Fidelity and Sense of Presence 317
Figure 86: Feedback quality – Telesimulation vs In-person 320
Figure 87 - Raincloud plots for the variables of Feedback..... 322

List of Tables

Table 1: Research Hypotheses, simulation session stages, statistical methods, and directionality of the hypotheses 32
Table 2: Steps in conducting the keyword search for the bibliometric analysis 47
Table 3: Main bibliometric indicators of the selected articles 50
Table 4: Type of the articles 50
Table 5: Citation matrix of the published articles 52
Table 6: Top 10 articles with highest citations 53
Table 7: Top 20 countries' production in Healthcare Simulation 54
Table 8: Top 20 countries with highest cited articles in Healthcare Simulation 55
Table 9: Intra-country and intercountry research collaboration 56
Table 10: Top 20 countries in terms of collaboration 57
Table 11: Top 20 countries with most frequent collaborations 58
Table 12: Most Relevant Source 60
Table 13: Top 20 institutions with the highest research productivity 63
Table 14: Author-wise publications 65
Table 15: Comparison between Scoping and Systematic Review 75
Table 16: Sections of PRISMA-ScR 78
Table 17: PICO components for the search question 83
Table 18: Search strings used for the different databases 84
Table 19: Fields used to extract information 87
Table 20: Reference Screening Level 1 90
Table 21: Application of Inclusion and Exclusion Criteria 90
Table 22: Publication Titles 92
Table 23: Simulation Tools and associated fidelity 110
Table 24: Key Differences between traditional, equivalence, and non-inferiority studies 146
Table 25: Summary of research design 154
Table 26: 8 Pair of emotions - Feldman-Barrett and Russell (2012) 168
Table 27: Emotional Appraisal Scale – Items, Sources, and Dimensions 169
Table 28: NASA-TLX – Dimension, endpoints, and descriptions 171
Table 29: Sample invitation text automatically generated by Qualtrics 175

Table 30: Scales, Description and Usage in Pre Simulation (1), Simulation (2), and Debriefing (3) 181
Table 31: Types of Error: Type I, II, and hypothesis rejection 184
Table 32: Sample Power calculation: Numeric Results for an Equal-Variance T-Test 184
Table 33: Sample analysis requirement – non-inferiority tests 186
Table 34: Sample analysis requirement – paired t-test sample vs effect size..... 186
Table 35: Post Hoc statistical power– one-sided t-test sample 187
Table 36: Post-Hoc statistical power– paired t-test sample..... 187
Table 37: List of simulation scenarios and the associated meta-information..... 192
Table 38: Summary statistics of demographic characteristics of subjects in telesimulation and in-person simulation 195
Table 39: Summary statistics of arousal and emotional response for Telesimulation and in-person groups for baseline 197
Table 40: Blocks for testing of Missing Data..... 200
Table 41: Missing Data Summary for Block 1 – Emotional Responses 201
Table 42: Missing Data Summary for Block 2 – Emotional Appraisal..... 204
Table 43: Missing Data Summary for Block 4 – Fidelity and Feedback Quality 207
Table 44: Instrument's minimum, maximum, and range of values..... 210
Table 45: Outlier Testing using Grub's outlier test..... 212
Table 46: Guidelines for normality 215
Table 47: Test of normality for Emotion State Variables 217
Table 48: Test of normality for Feedback and perception of presence 218
Table 49: Test of normality for Emotional Appraisal Process 218
Table 50: Test of normality for NASA Workload Scale 220
Table 51: Summary of Statistical tests for normality 221
Table 52: Common Method Bias Causes and Examples..... 223
Table 53: Single Factor Variance- Harman Test 225
Table 54: Factor Loadings and Factor Weights for the Emotion State Scale..... 228
Table 55: Validity and Reliability for the Emotion State Scale 228
Table 56: Discriminant Validity for the Emotion State Scale 229
Table 57: Factor Loadings for the Emotional Appraisal Process 231
Table 58: Validity and Reliability for the Emotion Appraisal Scale..... 232
Table 59: Discriminant Validity for the Emotion Appraisal Process - Fornell-Larcker Criterion..... 232

Table 60: Discriminant Validity for the Emotion Appraisal Process 233
Table 61: Factor Loadings and Factor Weights for the Fidelity and Feedback dimensions 234
Table 62: Validity and Reliability for the Fidelity and Feedback Scales 234
Table 63: Discriminant Validity for the Fidelity and Feedback Scales 234
Table 64: Factor Loadings and Factor Weights for the NASA-TLX 236
Table 65: Validity and Reliability for the NASA-TLX Scale 236
Table 66: Discriminant Validity for the NASA-TLX Scales 236
Table **67**: Cronbach's Alpha Reliability – NASA-TLX Single Dimension 237
Table 68: Final Constructs Loadings and computations for Study's variables 238
Table 69: Descriptive information for the Feedback and Fidelity items 239
Table 70: Descriptive information for the NASA-TLX Scale 241
Table 71: Descriptive information for the Emotional Appraisal Process 242
Table 72: Variables used in the Hypothesis testing and associated non-inferiority margins 247
Table 73: Summary of included variables and testing procedures for Hypothesis 1 ... 248
Table 74: ANOVA – Repeated Measures (Cheerfulness, Invigoration and Tranquility) – In-person 250
Table 75: PostHoc comparisons – Cheerfulness, Invigoration, and Tranquility – Simulation and Debriefing in-person 251
Table 76: ANOVA – Repeated Measures - Arousal during Baseline, Simulation, and Debriefing – in-person 254
Table 77: Post-Hoc comparisons – Arousal – in-person 255
Table 78: Summary of Results for Hypothesis 1 256
Table 79: Summary of included variables and testing procedures for Hypothesis 2 ... 257
Table 80: ANOVA – Repeated Measures (Cheerfulness, Invigoration and Tranquility) - Telesimulation 259
Table 81: Post Hoc comparisons – Cheerfulness, Invigoration and Tranquility – Simulation and Debriefing - Telesimulation 260
Table 82: ANOVA – Repeated Measures - Arousal during Baseline, Simulation, and Debriefing – Telesimulation 263
Table 83: Post Hoc comparisons – Arousal – Telesimulation 263
Table 84: Summary of Results for Hypothesis 2 264
Table 85: Summary of included variables and testing procedures for Hypothesis 3 ... 266

Table 86: Test for non-inferiority for Telesimulation – Arousal Response during Simulation..... 268
Table 87: Test for non-inferiority for Telesimulation – Arousal Response during Debriefing..... 269
Table 88: Test for non-inferiority for Telesimulation – Cheerfulness response during Simulation..... 272
Table 89: Test for non-inferiority for Telesimulation – Cheerfulness response during Debriefing..... 273
Table 90: Test for non-inferiority for Telesimulation – Invigoration response during Simulation..... 274
Table 91: Test for non-inferiority for Telesimulation – Invigoration response during Debriefing..... 275
Table 92: Test for non-inferiority for Telesimulation – Tranquility response during Simulation..... 276
Table 93: Test for non-inferiority for Telesimulation – Tranquility response during Debriefing..... 277
Table 94: Summary of Results for Hypothesis 3..... 278
Table 95: Post Hoc tests for differences in emotional responses - Independent Samples T-Test Simulation and in-person..... 279
Table 96: Summary of included variables and testing procedures for Hypothesis 4 ... 280
Table 97: Test for non-inferiority for Telesimulation – Mental Workload during Simulation (NASA TLX RAW)..... 282
Table 98: Test for non-inferiority for Telesimulation – Mental Workload during Simulation (NASA TLX RAW - Excluding Physical Demand)..... 283
Table 99: Test for non-inferiority for Telesimulation – Task Demand during Simulation 284
Table 100: Test for non-inferiority for Telesimulation – Task Demand during Simulation (excluding physical demand)..... 285
Table 101: Test for non-inferiority for Telesimulation – Demand on subjects during Simulation..... 286
Table 102: Summary of Results for Hypothesis 4..... 288
Table 103: Summary of included variables and testing procedures for Hypothesis 5 . 289
Table 104: Test of Equality of Variances for Arousal Responses..... 291
Table 105: Arousal Responses vs Scenario Complexity – Independent T-Test..... 291

Table 106: Test of Equality of Variances for Emotional Appraisal Dimensions 293
Table 107: Emotional Responses – Simulation and Debriefing vs. Complexity – Independent T-Test 293
Table 108: Test of Equality of Variances for the Mental Workload Scale and Subdimensions (NASA-TLX) 296
Table 109: Mental Workload Scale and Subdimensions (NASA-TLX) vs Complexity – Independent Samples T-Test 296
Table 110: Summary of Results for Hypothesis 5 298
Table 111: Emotional Appraisal Process – Comparison between Medium and Low Complexity – Independent Samples T-Test 300
Table 112: Summary of included variables and testing procedures for Hypothesis 6 . 301
Table 113: Test for non-inferiority for Telesimulation – Pleasantness 303
Table 114: Test for non-inferiority for Telesimulation – Goal Congruence 304
Table 115: Test for non-inferiority for Telesimulation – Certainty 305
Table 116: Test for non-inferiority for Telesimulation – Novelty 306
Table 117: Test for non-inferiority for Telesimulation – Predictability 307
Table 118: Test for non-inferiority for Telesimulation – Situational Control 308
Table 119: Test for non-inferiority for Telesimulation – Importance 309
Table 120: Summary of Results for Hypothesis 6 311
Table 121: Summary of included variables and testing procedures for Hypothesis 7 . 312
Table 122: Test for non-inferiority for Telesimulation – Perception of Fidelity during Simulation 314
Table 123: Test for non-inferiority for Telesimulation – Perception of presence during Simulation 316
Table 124: Summary of Results for Hypothesis 7 318
Table 125: Test of differences in the sense of Presence and Fidelity between in-person and telesimulation 318
Table 126: Summary of included variables and testing procedures for Hypothesis 8 . 319
Table 127: Test for non-inferiority for Telesimulation – Feedback Quality - Debriefing 321
Table 128: Summary of Results for Hypothesis 8 322

List of Appendixes

Appendix 1: Research Instrument 374
Appendix 2: Frequency distribution of study variables..... 378
Appendix 3: Frequency distribution and Skewness and Kurtosis analysis of study variables..... 399
Appendix 4: Variables Transformation 405

Chapter 1. Introduction

This chapter presents the main aspects that motivated the research, contextualizing the healthcare simulation relevance in the national and international context. Then, we describe the research objectives and briefly describe the study and its associated questions. Next, we share the researcher's statement and motives for conducting this research. Finally, we present the organization of this thesis.

1.1. Research Motivation and Context

The first patient simulators were created in the late 1960s and have seen widespread use since the 1980s and 1990s. Modern patient simulators use automatic control of features guided by physiological and pharmacological models to look and respond to human intervention and interaction with increasing realism using human-like robotic devices, such as mannikins. Likewise, high-fidelity patient simulators are increasingly used in the education and training of diverse healthcare professionals, including nurses, physicians, and emergency care personnel (Good, 2003).

Simulation and computerized devices are widespread in healthcare education in the United States and other developed countries (Shin, Park, & Kim, 2015). For instance, as of December 2021, the Society for Simulation in Healthcare (SSH) SIM Center Directory lists 600 simulation centers in the United States (https://www.ssih.org/Home/SIM-Center-Directory/Area/US).

Simulation-based education has evolved as a critical component of healthcare trainees' pre-licensure education, with the ultimate objective of creating learning experiences that are simultaneously safe and realistic (Bogossian et al., 2019). Bradley et al. (2019) support that simulation is widely used in nursing education, including more than half of the Boards of Nursing in the United States (30 out of the 50 states). These boards have regulations for simulation use and define a proportion of clinical hours that

can be replaced with simulation. Thirteen state Boards allow up to 50% of clinical hours to be replaced with simulation.

Likewise, Schlottmann et al. (2020) surveyed Departments of Surgery (DoS) in the United States. They found that 85 of 87 (98%) respondents indicated the presence of a simulation center in their institution, while 60 of 86 (70%) noted that simulation centers were in their department. Most DoS had incorporated a formal simulation curriculum for trainees (83 out of 87), with 90% offering protected time for simulation education for their residents with 76% reporting engaging with the simulation center at least monthly.

Ziv et al. (2003) present the simulation-based medical simulation as an "ethical imperative" that can balance the fundamental dilemma between training new health professionals and patient treatment, safety, and well-being. According to the authors, simulation-based learning in medical education allows health professionals to develop critical knowledge, skills, and attitudes without the risks associated with learning with actual patients. Simulation-based healthcare learning can provide engaging and relevant learning opportunities using simulated clinical experiences which are accurate and 'life-like.' (Cantrell, Franklin, Leighton, & Carlson, 2017).

Several studies indicate that students using computerized health simulators demonstrated improved diverse and critical professional and academic outcomes – such as critical thinking, confidence, clinical competence, and theory-practice integration (Kaddoura, Vandyke, Smallwood, & Gonzalez, 2016; Raman et al., 2019). Simulation is considered a key component in healthcare education curricula, demonstrating positive results for students and educational institutions (Hayden et al., 2014).

A primary goal of high-fidelity healthcare simulation is to prepare learners to respond to real-world situations, and activities that evoke strong emotional responses (Pawar et al., 2018). Likewise, Dunnington (2015) describes another objective of healthcare simulation: to promote real-life learning opportunities that induce a sense of presence for students to feel like they are in a natural clinical environment.

A high-fidelity simulation should be as real and challenging as possible to allow training in complex or rare situations to foster expertise that can transfer to clinical environments without the risks associated with learning on the job with actual patients (e.g., errors, malpractice). However, the computerized healthcare simulation's effects, implications, and effectiveness are not yet well understood and represent a potentially rich research area that raises many important, complex research questions.

During the COVID-19 pandemic, several universities transitioned the healthcare simulation sessions to a remotely facilitated, referred to as telesimulation or virtual simulation (Ray et al., 2021; Diaz, 2021; Roach & Okrainec, 2021). The Society for Simulation in Healthcare (SSH) released a *Position Statement on Use of Virtual Simulation* during the Pandemic to support virtual simulation as a viable replacement method for clinical hours for students currently enrolled in health sciences. The statement also attests that virtual simulation is an effective teaching method used for over ten years and has yielded improved learning outcomes (Kononowicz et al., 2019; Foronda et al., 2020).

Virtual simulation (VS) is a broad term that encompasses multiple and frequently conflicting definitions. It has been used in various health care professions to provide learning opportunities to experience real-life clinical scenarios with interactive, computerized, virtual patients. Virtual simulations include strategies that present situations that resemble reality in a digital setting to allow clinical training in decision-making, motor control, and communication skills, among other outcomes (Padilha, Machado, Ribeiro, Ramos, & Costa, 2019).

Turrise et al. (2020) identify studies that aimed to establish the efficacy of virtual simulations. For instance, they cite Cobbett and Snelgrove-(2016) Clarke's study, which revealed no significant differences in participant knowledge or confidence between VS and face-to-face simulation. While most studies reviewed did not find significant differences in outcomes while using VS, increases in knowledge, problem-solving, and critical thinking were observed, and positive student feedback regarding virtual simulations.

Several studies use the term virtual simulation without adequately defining it, causing confusion and hindering the discussion and meta-analysis of its effects and affordances across studies. Given the lack of specificity of this term, Cant et al. (2019) found seven different types of virtual simulation described in the specialized medical simulation literature:

a) Immersive simulation - A simulation of a real-life situation involving the participants' senses, emotions, thinking, and behavior; creating an immersive simulation depends on alignment with learning objectives and the simulation's fidelity (physical, conceptual, and emotional) and participant's perception of realism.

b) Screen-based simulation - Presented on a computer screen using graphical images and text, like a gaming format, where the learner interacts with the interface using a keyboard, mouse, joystick, or other input devices.

c) Serious games - Implemented as a game designed with an explicit and defined academic or professional purpose

d) Virtual reality – Representation of an interactive 3D world in which objects have a sense of spatial presence; that is, objects that are not physically present but appear to be so through computer software.

e) Virtual simulation and virtual patients - A simulation involving people operating simulated systems. Virtual simulations may include surgical simulators used for on-screen procedural training, usually integrated with a haptic device (s) to interact with the system.

f) Virtual reality simulation - Simulations that incorporate an array of physical or interfaces such as a computer keyboard, a mouse, speech/voice recognition, motion sensors, or haptic devices to replicate real-life situations and healthcare procedures;

f) Web-based simulation - A screen-based simulation with the simulation program operating from a website via an interface with the Internet.

Throughout this research, we use the concept of virtual simulation to signify the use of video conferencing tools such as Zoom, Skype, Google Meets, and other tools to conduct healthcare simulations without the need for the learner's physical presence (co-location) in a simulation center.

Some traditional simulation resources such as computer terminals, vital signs terminals, and robotic manikins are used in this modality. However, participants can be distant from the simulation center (even in different countries). This type of simulation has been described in the emerging literature as telesimulation.

In 2010, Okrainec et al. (2010) described telesimulation as a new concept that involved using the Internet to connect an instructor and a learner in separate locations during a simulation session. However, McCoy et al. (2017) discuss the limitation of this simple definition, which did not convey the cognitive and affective aspects of learning involved in the medical simulation. After a throughout discussion, McCoy et al. submitted a definition to the Society for Healthcare Simulation for the official term in SSH dictionary:

> *"We define telesimulation as a process by which telecommunication and simulation resources are utilized to provide education, training, and/or assessment to learners at an off-site location. Off-site location refers to a distant site that would preclude the education, training, and/or assessment without the use of telecommunication resource"* (McCoy et al. p.133 - 2017)

Preliminary evidence supports the effectiveness of both co-located and virtual simulations in several dimensions (e.g., technology acceptance, satisfaction, skills, clinical reasoning, knowledge, and effective management of care, among others) (Duff et al., 2016 Kaddoura, et al., 2016; Raman et al., 2019). However, additional studies are needed to understand important but less explored aspects of the learning experience, such as learners' emotional responses in healthcare simulations.

An established field of research seeks to understand the use of computers in educational settings through the human-computer interaction perspective (Ahn & Clegg, 2018). A more specific and novel research agenda that aims to understand how humans interact with the various computerized devices from an affective computing perspective was described by Hudlicka (2017).

Affective computing is a computing paradigm that considers emotions, detecting, eliciting, and responding to human emotions, with diverse and promising applications for computer-assisted learning, perceptual information retrieval, arts and entertainment, and human health and interaction (Piccard, 2000; Jeon, 2017). Cognitive Sciences have been the dominant paradigm of human factors and human-computer interaction research, given the focus on the information processing approach in the field (Jeon, 2017). Only recently have emotions begun to receive attention from mainstream human-machine research.

The term "affect" has different connotations regarding the evaluative aspect of attitudes and the tripartite model with cognitive, affective, and connotative/behavioral components (Cohen, Pham, and Andrade, 2008). Cohen et al. (2008) explain that an explicit or implicit "liking" for some object, person, or position is considered an evaluative judgment rather than an internal feeling state. The essential characteristic of this definition of affect is related to induced feeling states instead of a liking or a descriptive cognition.

Likewise, “emotion” has various definitions, frequently with non-specific meanings. In an influential paper, Clore et al. (1987) tested the validity of a taxonomy of affective concepts to distinguish emotions from similar terms in the affective lexicon, which according to the authors, is the subset of words related to affect or affective conditions.

This research has adopted the concept of emotion as a “valenced affective reaction to perceptions of situations” (Richins, 1997, p.127), separating the concept of emotion from other types of cognitions without valence evaluation, such as interest and surprise, or purely physical states such as sleepy and droopy. General and subjective personal evaluations such as feeling self-confident or abandoned are not contained in that definition of emotion (Richins, 1997).

It is well established that emotions have a pivotal role in education (Phye, 2007; Pekrun, R., & Linnenbrink-Garcia, 2014). Given the increase in use, understanding the impact of emotion in simulation-based healthcare education is urgent and important (LeBlanc, 2019). Jeon (2017) states that emotions affect human cognition on various levels and influence fundamental processes: attention, perception, and memory.

Emotions also influence situational assessment, decision-making, goal management, planning, and learning. Emotional responses to educational situations are associated with important outcomes such as student persistence in learning, satisfaction, and academic achievement (Henrie, Halverson, & Graham, 2015). Cohen, Pham, and Andrade (2008) state that the arousal intensity of an affective experience increases immediate and long-term memory for the experience, suggesting a better consolidation of memory traces under high emotional arousal conditions.

Fraser and McLaughlin (2019) indicate that emotions in healthcare simulation affect learning in many ways, with broad evidence of its impact on motivation, level of cognitive processing and problem-solving; retrieval of learned material; and extraneous cognitive load. However, the relationship between emotions and learning is complex. Negative emotions may hinder learning, while positive emotions may improve learning due to increased motivation and a heightened problem-solving capacity (Fraser et al., 2012).

A 2019 editorial published in the Journal of the Society for Simulation in Healthcare (SSH) - “The Relationship Between Emotions and Learning in Simulation-Based Education” called for additional research on emotional responses in healthcare simulation.

The editorial indicated that a better understanding of emotional responses and the effects of emotion on learning would provide a foundation for creating a new generation of simulation-based curricula, which can improve learners' clinical practice. Understanding emotional responses and the role of emotion in simulation is both an essential and urgent need (LeBlanc, 2019).

If we employ a more precise virtual simulation definition of telesimulation (McCoy et al., 2017; Cant et al., 2019), the effectiveness and outcomes of telesimulation are much less studied than in-person simulation, especially regarding learners' emotional responses.

As the COVID pandemic impacts simulation centers, many instructors and learners accustomed to in-person high fidelity simulation centers may have been involuntarily introduced to virtual simulation. The shift to a remote format (telesimulation) transforms learners' and instructors' dynamics and experience of the simulation process, creating the need to understand the new educational processes from the perspectives and circumstances of both instructors and students/participants.

Investigating learners' affective and cognitive responses in these new settings is essential to understand better telesimulation and its effectiveness as an alternative educational method. It is essential to verify whether telesimulation impacts learners' emotional responses and mental workload, similarly to in-person simulation.

Comprehending students' learning experience through the theoretical lens of emotional responses can improve educational outcomes and future development of interventions that can deal with stress and anxiety before they occur in clinical settings and maximize learning experience effectiveness.

The proposed research aims to understand: the structures in which learners' experience unfolds in telesimulation and in-person simulation; and emotional responses and evaluation of experiences in a telesimulation environment using the in-person simulation as a control variable. As a secondary objective, we will assess the feasibility and applications of telesimulation as a comparatively challenging and emotion-evoking educational method mediated by remotely facilitated methodologies.

1.2. Research Questions and Objectives

As will be demonstrated in the bibliometric, scoping, and literature review, telesimulation is a new and rapidly growing field in medical education, explicitly

emerging as a response to the COVID-19 closure of simulation centers but having its first appearance long ago. However, telesimulation also has longer-term applications. Telesimulation is closely related to telemedicine and increased access to simulation center expertise in developing countries and rural settings.

This study aims to understand better how participants' emotional and cognitive response is affected during remotely facilitated simulation healthcare scenarios and how it compares to the in-person simulation as a viable and acceptable methodology for healthcare education. We will assess emotions and cognitive responses by combined use of surveys that appraise learners' responses while they undergo remotely facilitated and in-person training in clinical simulation centers.

Our overarching research question:

Is Telesimulation equally engaging, challenging, and emotion-inducing learning experience as in-person simulation?

To answer the question, we conducted a natural experiment comparing telesimulation and in-person simulation to determine the extent to which learners' emotional and cognitive responses are affected on telesimulation and in-person simulation in low and medium complexity scenarios.

Understanding cognitive and emotional responses in remotely facilitated healthcare simulation can help determine the efficacy and feasibility of telesimulation as a viable alternative to conducting healthcare training, even in the absence of a pandemic, thus having both practical and theoretical impacts in the long term.

This investigation builds upon Fraser et al. (2012), Pawar et al. (2018), and Fraser & McLaughlin (2018)'s studies measuring the workload and emotional response of healthcare professionals during simulation training and debriefing. The current research innovates by adding new components to understand emotional and cognitive responses in a novel setting. First, it includes recent developments in theories of emotional appraisal processes (Scherer and Moors, 2019) instead of relying only on discrete emotion models as used in Fraser & McLaughlin (2018).

In addition, this study also collected these additional responses during the critical moments of pre-simulation, simulation, and debriefing, with the potential to present findings on the dynamics of emotional responses of learners related to healthcare

simulation medical and nursing training with implications in human-computer interaction.

This research is a direct and specific response to the 2019 SSH call to consider and better understand the role of emotions in healthcare simulation. Based on previous research, we have added scenario complexity as an additional variable, along with perceptions of fidelity and sense of presence, to strengthen internal research validity.

Finally, this research compares the emotional response of learners in telesimulation with in-person simulation. The role of a critical aspect of learner experience – participants' emotional responses - must be better understood in both in-person and telesimulation settings. This need is more pressing since this alternative form of simulation will likely remain in use, and likely increase, even after the COVID-19 pandemic is overcome.

We have proposed secondary questions that answered to understand the various mechanisms and subcomponents of the emotional and cognitive responses:

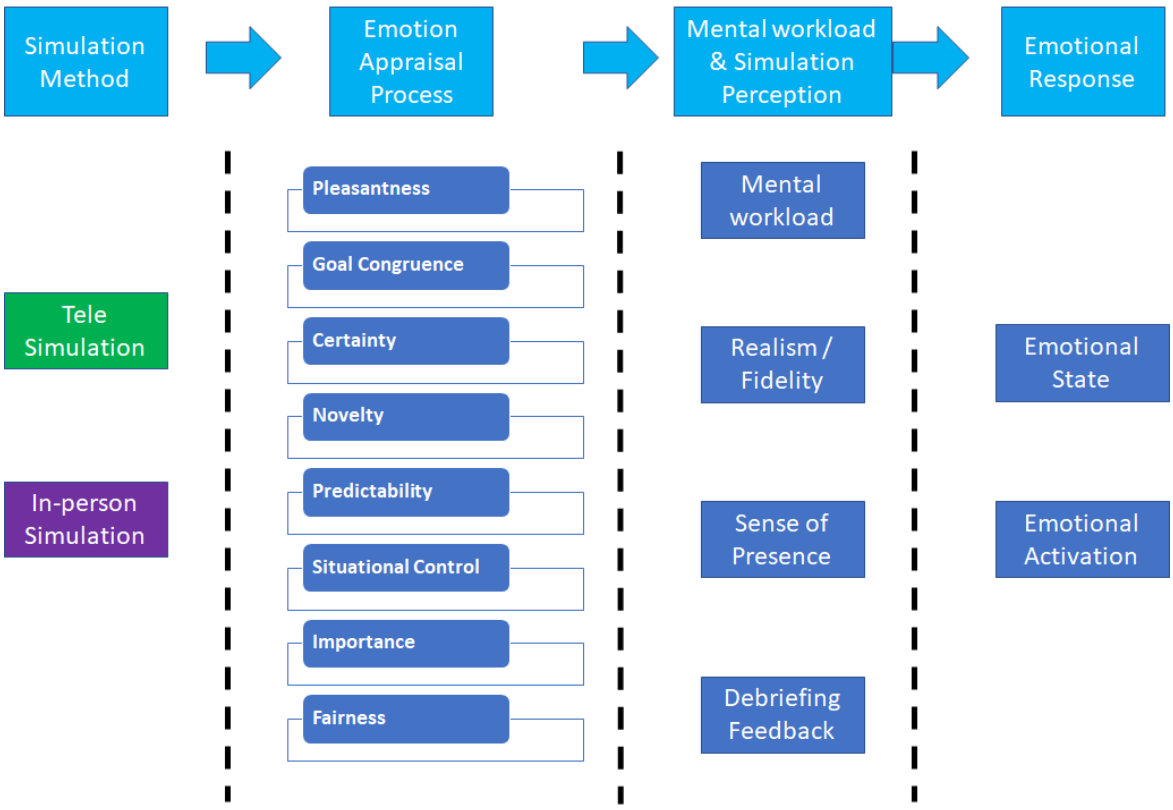
- Does in-person simulation evoke emotional responses between baseline, post-simulation, and post-debriefing stages?
- Does telesimulation evoke emotional responses between baselines and post-simulation and post-debriefing?
- Are emotional responses in telesimulation non-inferior compared to in-person?
- Is the mental workload non-inferior in telesimulation compared to in-person?
- Does scenario complexity have a role in explaining the emotional responses and cognitive responses?
- Is the emotional appraisal process non-inferior in telesimulation compared to in-person?
- Is the perception of realism, fidelity, and sense of presence non-inferior in telesimulation compared to in-person simulation?
- Is telesimulation a non-inferior method to provide debriefing feedback after simulation sessions compared to in-person?

The secondary questions supported the specification and conduction of bibliometric analysis, scoping review, and literature review, presented in the following chapters, which defined and explored the key constructs that needed to be formally defined and the inter-relationships.

The secondary questions were derived from the bibliometric analysis, scoping review, and literature review, presented in the following chapters, in which we define and explore the key constructs and inter-relationships.

Figure 1 presents the main concepts and their categorization in a) the simulation methods (including control and treatment), b) variables related to the emotion appraisal process, c) a list of mental workload and simulation perception, and d) emotional response.

Figure 1: List of study constructs and their categorizations



Source: Compiled by the author to generate and base the study hypotheses

Kothari (2004) argues in science that experiments are designed with the objective of testing hypotheses. A hypothesis is a critical instrument in research, leading to new experiments and observations. Still, according to Kothari (2004), in everyday language, when someone talks about a hypothesis, they usually mean a simple assumption or a guess

that must be proven or denied. However, a hypothesis is a formal inquiry that a researcher wants to answer in academic contexts.

A hypothesis is a proposition put out as an explanation for the occurrence of a specific phenomenon, either as a provisional presumption to guide additional research or as a highly probable explanation in light of known facts. A research hypothesis is frequently a predicted statement that connects an independent variable to a dependent variable and may be tested using scientific procedures (Kothari, 2004).

However, there are requirements for defining a hypothesis. Browner et al. (2001) provide a clear and precise list of requirements for a good hypothesis. According to them, a hypothesis should be based on a good research question, be simple, specific, and be stated in advance. Kothari (2004) presents a more comprehensive list with eight requirements:

- Be clear and precise
- Be testable (logically or statistically tested)
- Should state relationships between variables (for relational hypothesis)
- Be limited in scope and be specific
- Be stated in simple terms to be understandable to the largest possible audience
- It should be consistent with most known scientific evidence
- Must explain the facts that originated the necessity of the hypothesis testing (the underlying problem must be based on facts and empirical references)

Another essential aspect of the creation of a hypothesis is the notation. More specifically, in terms of statistical analysis, the null and alternative hypotheses are employed and mutually exclusive. The null hypothesis usually takes the form of H0, while the alternative hypothesis usually is expressed as H1 or Ha.

The null hypothesis is the formal mechanism to test statistical significance, usually establishing whether there is no association or a difference between the study's variables. On the other hand, the alternative hypothesis proposes an association (or difference) between the variables. The researcher's objective is to reject the null hypothesis and accept the alternative hypothesis by exclusion (given that it is not directly tested). Therefore, we formalized our primary and secondary research questions in a formalized set of hypotheses. The central hypothesis of the research is as follows:

- **Null Hypothesis (H0)** - Telesimulation is an inferior educational approach, measured by emotional and cognitive responses, compared to in-person simulation.

- **Alternative Hypothesis (H1)** - Telesimulation is a non-inferior educational approach, measured by emotional and cognitive responses, compared to in-person simulation**.**

Following Kothari's (2004) guidelines regarding the requirements for a hypothesis, we decomposed the central hypothesis into a set of hypotheses, operationalizing variables and simplifying the central hypothesis into the secondary hypothesis that can be statistically tested with the appropriate tests. The collection of generated hypotheses and hypothesis-testing methods are presented in Table 1:

Table 1: Research Hypotheses, simulation session stages, statistical methods, and directionality of the hypotheses

Research Hypothesis	Temporal Stage (Simulation or Debriefing)	Statistical Method	One-tailed / two-tailed or non-inferiority
H1 – In-person simulation does not evoke an emotional response during simulation or debriefing	-	-	Two-tailed
H1a: In-person simulation does not evoke an emotional state response during simulation and debriefing (time factor) as measured by the Emotional State Scale (ESS) (Feldman, Barrett and Russell, 2012)	SIMULATION AND DEBRIEFING	Paired t- tests / Repeated measures ANOVA	Two-tailed
H1b: In-person simulation does not evoke an arousal response emotional state response during simulation and debriefing (time factor) as measured by Felt Arousal Scale (FAS) (Svebak & Murgatroyd, 1985)	SIMULATION AND DEBRIEFING	Paired t- tests and Repeated measures ANOVA	Two-tailed
H2 – Telesimulation does not evoke an emotional response during simulation or debriefing	-		Two-tailed
H1a: Telesimulation does not evoke an emotional state response during simulation and debriefing (time factor) as measured by the Emotional State Scale (ESS) (Feldman, Barrett and Russell, 2012)	SIMULATION AND DEBRIEFING	Paired t- tests and Repeated measures ANOVA	Two-tailed
H1b: Telesimulation does not evoke an arousal response emotional state response during simulation and debriefing (time factor) as measured by Felt Arousal Scale (FAS) (Svebak & Murgatroyd, 1985)	SIMULATION AND DEBRIEFING	Paired t- tests and Repeated measures ANOVA	Two-tailed
H3 - The emotional response evoked by telesimulation is inferior to in-person simulation during simulation and debriefing		Independent Samples t-test	Non-inferiority
H3a: Arousal response (FAS) between baseline and post-simulation in telesimulation is inferior to in-person simulation	SIMULATION	Independent Samples t-test	Non-inferiority
H3b: Arousal response (FAS) between post-simulation and post-debrief in telesimulation is inferior to in-person debriefing	DEBRIEFING	Independent Samples t-test	Non-inferiority
H3c: Emotion state response (ESS) between baseline and post-simulation in telesimulation is inferior to in-person simulation	SIMULATION	Independent Samples t-test	Non-inferiority
H3d: Emotion state response (ESS) between post-simulation and post-debrief in telesimulation is inferior to in-person debriefing	DEBRIEFING	Independent Samples t-test	Non-inferiority
H4 – The mental workload (as measured by NASA-TLX) in telesimulation is inferior to in-person simulation	SIMULATION	Independent Samples t-test	Non-inferiority
H4a: Task demand on the subject (Mental, Physical and Temporal) in telesimulation is inferior to in-person simulation	SIMULATION	Independent Samples t-test	Non-inferiority
H4b: Subject response to the task (Perceptions of effort, performance, and frustration) in telesimulation is inferior to in-person simulation	SIMULATION	Independent Samples t-test	Non-inferiority

H5 –Scenario complexity does not affect the level of emotional and cognitive responses during telesimulation and in-person simulation	SIMULATION	Independent Samples t-test	One-tailed
H5a: Higher complexity scenarios are not associated with stronger emotional state (ESS) in-person simulation	SIMULATION	Independent Samples t-test	One-tailed
H5b: Higher scenario complexity is not associated with stronger arousal response (FAS) in in-person simulation	SIMULATION	Independent Samples t-test	One-tailed
H5c: Higher scenario complexity is not associated with a higher task demand on subjects (NASA-TLX) in-person	SIMULATION	Independent Samples t-test	One-tailed
H5d: Higher scenario complexity is not associated with a higher subject response to the task (NASA-TLX) in-person	SIMULATION	Independent Samples t-test	One-tailed
H6 - The scores of emotional appraisal process (Pleasantness, Goal Congruence, Certainty, Novelty, Predictability, Situational Control, Importance and Fairness) in telesimulation is inferior to in-person simulation	SIMULATION	Independent Samples t-test	Non-inferiority
H7: The perception of fidelity/realism and sense of presence in telesimulation is inferior to in-person simulation	SIMULATION	Independent Samples t-test	Non-inferiority
H7a: Perception of fidelity/realism (Simulation Design Scale – Student Version (Jeffries, 2016) in telesimulation is inferior to in-person simulation	SIMULATION	Independent Samples t-test	Non-inferiority
H7b: Sense of presence (Dinh, Walker, Song, Kobayashi, & Hodges, 1999) in telesimulation is inferior to in-person simulation	SIMULATION	Independent Samples t-test	Non-inferiority
H8: The debriefing feedback score in telesimulation is not inferior to in-person simulation	DEBRIEFING	Independent Samples t-test	Non-inferiority

1.3. Personal Motivation

The present research and my interest in the theme arose from complex and interconnected sequences of events and experiences as a student, a scientist, and an educator. My student experience during this doctoral course triggered my interest in human-computer interfaces and how humans interact with computer interfaces.

A crucial moment was when I enrolled in an elective class in UNIRIO – Human-Computer Interaction in 2017. Participating in this class prepared me to understand critical concepts of HCI that later proved essential to awaken my interest and curiosity in the theme. I found a new world of multidisciplinary theories, methodologies, and approaches available in Information Sciences, Psychology, Education, Engineering Business, Informatics, and Computer Science.

Given that newfound interest, I pursued one opportunity to engage in research at New York University as a Visiting Research Scholar, mentored by Professor Burleson. The new opportunity was to be part of the National Science Foundation[1]'s Major Research Instrumentation Program (MRI): Development of Experiential Supercomputing: Developing a Transdisciplinary Research and Innovation Holodeck.

This opportunity was in a new country, a different scientific area, in the College of Nursing, meaning many changes and a steep learning curve. It took me a while to realize the importance of multidisciplinary work and how Informatics was essential for the College of Nursing and its Research agenda. When I visited the NYU Clinical Simulation Center, it immediately became clear to me the potential and connections between the seemingly distant fields.

I noticed all the robotic manikins, monitors, and networks and how the technology was a central part of Nursing, Medicine, and Health Sciences in general. Simultaneously, I became involved in projects that investigated participants' affect as a critical component to understanding human-computer and social interactions, so I became aware of the Affective Computing approach and the role of emotions in developing next-generation computerized systems.

[1] "The National Science Foundation (NSF) is an independent federal agency created by Congress in 1950 "to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense..." https://www.nsf.gov/about/glance.jsp

Taking advantage of the hands-on training I received as a Visiting Scholar in the Holodeck lab, I planned to use the NYU-X laboratory affective sensors to understand users' affect and emotions during simulation sessions. This setting drove me to study and experiment with a wide selection of heart rate sensors, galvanic skin responses, electroencephalogram (EEG), and motion capture for almost a full year in NYU.

However, my mentor accepted a position in Arizona, on the other side of the country, and offered me a postdoctoral position to join him and his new lab, which I gladly accepted. Fortunately, the University of Arizona had a state-of-the-art Medical Simulation Center, and I continued to pursue my research interests in this new area. It took me a reasonable amount of time to establish contacts with the leadership of the simulation center, better define the research protocols, and select the sensors to be employed.

However, when I believed I was ready to submit the IRB protocol, the United States (and the world shortly after) was struck by the COVID-19 pandemic. At first, like most people, I waited and hoped that it was temporary, a matter of a couple of weeks. However, as time passed, it became more evident that the research protocol needed to be adjusted since Simulation Centers were closed all around the nation, and new forms of education were employed.

The new form, telesimulation, was a creative and bold initiative led by some of the Nation’s leading universities to maintain the education of much-needed professionals to deal with the COVID-19 epidemic. Out of the many essential services, we could not afford, as a society, to postpone medical and healthcare education. One frequent question was whether telesimulation was an acceptable replacement for the in-person simulation.

As I investigated more about the issue, I found that various studies have tried to answer that question in the most straightforward way possible: by inquiring students and instructors about their perception of whether this new modality was acceptable. Although necessary, I felt that something was missing in this approach.

Using my realization of the importance of emotions could be one precise and valuable proxy for the effectiveness of telesimulation compared with in-person simulation, I have adapted my research to compare the two types of simulation: in-person and telesimulation, to evaluate the emotional response using a series of surveys and design a natural experiment considering the limitations of conducting research with medical trainees during a pandemic.

1.4. Document Structure

This first chapter described the research context, research questions and hypothesis, and personal motivation to develop this research. Chapter 2 – Preliminary studies - presents an overview of studies in both simulation and telesimulation using bibliometrics and scoping review as methods.

Chapter 3 – Literature Review - presents a literature review summarizing critical concepts related to simulation, healthcare simulation, emotions, and the emotional appraisal process, the uses and importance of emotions in education, HCI, and healthcare simulation. The third chapter introduces two fundamental methodological approaches in this study: natural experiments and non-inferiority studies.

Chapter 4 – Methodology - presents the research in more detail, including the methodological approach, research site, instruments employed, and other methodological selections.

Chapter 5 – Data Analysis - presents the data analysis, including descriptive statistics, missing values, scale validation, statistical tests, and hypothesis testing using non-inferiority and group comparisons. Finally, Chapter 6 – Discussion and Conclusions - discusses the hypothesis testing results and presents the study conclusions, discussing the study's limitations and suggestions for future research.

Chapter 2. Preliminary Studies

For the current thesis, we have performed two independent but related reviews – one bibliometric study to define and explore the field of medical simulation. In contrast, the second review, a scoping review, unfolds the recent scientific and practical field of telesimulation. The two complementary reviews are the outcome and document the iterative process employed to define and propose the research question.

The initial research question was broadly based on healthcare simulation and then evolved into a natural experiment involving telesimulation after the COVID-19 pandemic and the role of telesimulation as a promising method for healthcare simulation with implications beyond the COVID-19 pandemic.

After the initial review of the healthcare simulation, we recognized the importance and prevalence of HCI approaches in the field to understand users' or participants' affect when interacting with robotic and informatics systems. Thus, the present chapter is divided into two main sections: A bibliometric analysis of the field of medical simulation, followed by a scoping review of the new field of simulation – telesimulation, presented in sequence.

2.1. Bibliometric Studies

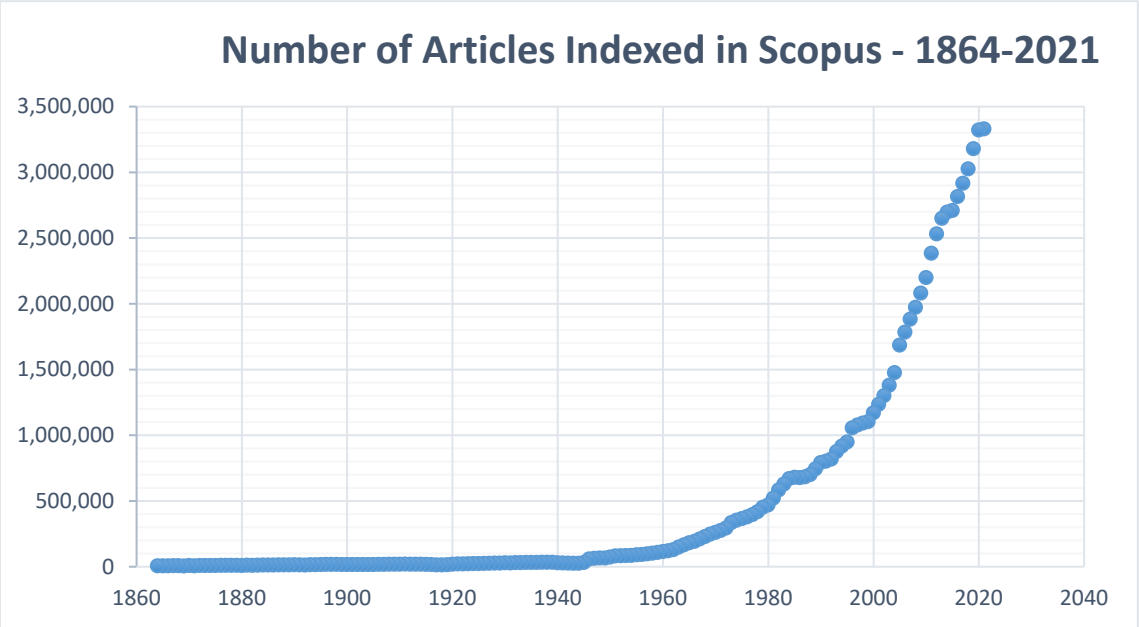
The scientific research output produced in most scientific domains has reached a production level much higher than its members can effectively read. For instance, the SCOPUS database (www.scopus.com), one of the leading scientific databases with summary and full text of theses, articles, and abstracts, has 76.8 million core records, more than 16 million author profiles, and 70,000 academic affiliations profiles (SCOPUS, 2020).

Scopus organizes titles into four broad subject categories (life sciences, physical sciences, health sciences, social sciences, and humanities), divided into 27 major subject areas and 300+ minor subject areas, distributed in 39743 journal titles (SCOPUS, 2020).

A SCOPUS query for English-language articles conducted in December 2021 retrieved 3,330,447 documents published in 2021, distributed in 2,492,794 articles, 231,266 reviews, and 72,104 other documents. Areas such as Medicine (922,236), Engineering (601,703), and Computer Science (404,054) have reached yearly articles production comparable to all scientific research a few decades ago.

To illustrate, the same query (English language documents) retrieves 791,650 documents for 1990, effectively 85% of 2021 production for the medical field alone. In 30 years, the total of articles has grown by 320 %, reflecting a significant increase in the volume of publications published globally, confirming the trend for this number to rise significantly in the following years. We can verify the growth in Figure **2**:

Figure 2: Number of articles indexed in Scopus (English) – 1864-2021



The exponential growth in the number of articles published emphasizes the necessity to study scientific production in an automated and streamlined manner to identify the main articles and journals in any academic field. The term statistical bibliography, known as Bibliometrics, was first introduced in 1923 by E. Wyndham Hulme, referring to the methods and techniques used to understand scientific and technological processes by counting documents and publications (Guedes & Borschive, 2005).

One of the first documented discussions of bibliometric analysis was provided by Lotka (1926), presenting it as a "method for measuring researchers' productivity," widely known as Lotka's Law, describing the productivity of authors. The law states that there is diminishing productivity among scientific authors.

A significant number of authors will limit their contribution to one article, while a select number will be responsible for creating multiple articles (Budd, 1988). Simultaneously, Gross and Gross (1927) reportedly conducted the first citation analysis to define which scientific periodicals would best prepare students in chemistry majors, analyzing 3,633 references distributed among 247 different journals.

In his seminal article "Statistical Bibliography or Bibliometrics" published in 1969, Allan Pritchard popularized the term bibliometrics to designate the field of studies that uses mathematical and statistical methods to investigate and quantify the processes of written communication, including books, articles, and other media (Guedes & Borschiver,

2005). Bibliometric analysis is a research methodology that has been widely applied in library and information science and uses statistical tools to analyze published academic studies (Liang and Liu 2018).

According to Zupic & Čater (2015), although not a new method, bibliometrics has only recently gained considerable attention due to the growth of readily available online databases with citation data, such as SCOPUS, Web of Science, Google Scholar. Another factor is the simultaneous emergence of various software for conducting bibliometric analyses (e.g., HistCite, BibExcel, Bibliometrix, VOSVIEWER, among many others).

Definitions of bibliometrics are plentiful. Bibliometrics study the scientific production, citations, and content produced by the scientific community. It is concerned with quantifying written communication through data collecting, processing, and presenting procedures. Publications, authors, keywords, institutions, country of origin of researchers and institutions, number of citations, and leading journals are the main factors analyzed in a bibliometric study (Guedes & Borschiver, 2005)

According to Sainero (2000), Bibliometrics is a method used to gather methodological knowledge through quantitative methodologies to examine scientific information production, communication, and use. Bibliometrics is one of the most potent tools information professionals and managers use to identify critical indicators for appraising scientific output. It is also a tool for interaction between researchers, institutions, and areas of knowledge, allowing the construction and visualization of complex relationships between various levels of analysis.

Glanzel (1996) presents bibliometrics as a sophisticated research field that, while generally designated as social science, is influenced by pure and technical sciences. Any methodological classification requires well-documented data processing methods based on a precise definition of the sources and indicators. Furthermore, the selection and successful integration of technologies used in various investigations is required.

Porter et al. (2002) define bibliometric analysis as research profiling to denote the broad scanning of related literature by investigating topical relationships and research trends with the potential to connect efforts by diverse research domains. More importantly, research profiling allows identifying the central issues in a scientific field, which techniques are generally used, the main topics, and how the field evolves. They point out that the (then-current) bibliometric analysis has evolved into text mining, going beyond the traditional citation analysis to become the full-scale investigation of abstracts or even complete document sets.

Bibliometric analysis can represent a project on its own. However, the authors defend that bibliometric analysis, or in their words, research profiling, should be used to present the context for all the primary research. They emphasize that objective of research profiling is not to replace a literature review but to strengthen it (Porter et al., 2002).

Still, according to Porter et al. (2002) it is possible to derive a different type of knowledge from a body of literature than the careful reading of individual articles used in the formal literature review. The literature review approach has a micro-focus (papers as the unit of analysis), a narrow range (with a limited number of references), is usually strongly constrained to the topic, and is presented as a text discussion.

However, the research profiling approach has a macro focus on finding patterns in the literature). Moreover, it encompasses a much higher range of literature (usually over thousands of references). It includes the topic and related research domains and questions, and the results can be presented in text, numerical and graphical depictions.

Likewise, Zupic, I., & Čater (2015) share the view on bibliometric analysis compared to the traditional theoretical literature review, citing how the volume of research has increased in the management and organization sciences domains, with direct implications for the feasibility of covering the relevant literature. They support bibliometric analysis given its capacity to deal with the abundance of data available to filter research and discover the impact and foundational and vital articles.

Also, they point out that the bibliometric approach can increase review quality by introducing a systematic, transparent, and repeatable review process. In the same essence presented by Porter et al. (2002), Zupic, I., & Čater (ibid) defend those bibliometric approaches can help with literature reviews by directing the researcher to the most influential publications and mapping the research field without subjective bias.

As a result, according to Zupic, I., & Čater (ibid), science mapping adopting bibliometric techniques provides a unique insight into the domain. Instead of a structured literature review, science mapping has a macro focus and tries to discover patterns in the literature as a body of work. While a conventional literature review can provide depth, bibliometric approaches can address many studies (hundreds or thousands). This analysis method can also provide a graphic representation of a research field.

Finally, Zupic, I., & Čater (ibid) emphasize that bibliometric methods are not a replacement for traditional review methods but rather a complement. Even if employed without all the methodological rigor, the analysis can convey valuable information about the research field, such as which publications and authors are relevant and the field's

structure. Consequently, bibliometric methodologies can be used to enhance conventional literature reviews.

2.1.1. Bibliometric Analysis - Objectives and Questions

During this chapter, we have employed bibliometric analysis to study and characterize the field of healthcare simulation. This chapter evaluated the healthcare simulation scientific production as indexed in Web of Science (WOS) database. Our objective is to present telesimulation, identify research trends in healthcare simulation, and demonstrate the field's development, central questions, and evolution over time. Stated as a research question, our purpose is:

RQ. What is bibliometric analysis's contribution to reviewing and developing the literature on healthcare simulation?

The sub-questions are:

RQ1. What is the number of articles published on healthcare simulation?

RQ2. What structure is formed by the publications and citations in healthcare simulation?

RQ3. Who are the most cited authors, and how do they influence and collaborate in healthcare simulation?

RQ4. Which research documents are cited the most frequently by authors in healthcare simulation?

RQ5. Which journals give the most attention to the topic and are the most important?

RQ6. Which keywords do authors on healthcare simulation use the most?

RQ7. What are the most critical research institutions producing research in healthcare simulation?

RQ8. What are the most important countries producing research papers in healthcare simulation, and how do they inter-collaborate?

Language

In carrying out this bibliometric analysis, we chose the English language to build the search query since it is the language adopted by most international journals related to

the research subject. It is also the primary language used by most electronic bases listed in the CAPES Portal.

Additionally, during the keyword selection process, we did not find articles written in other languages in which the author is proficient (Portuguese or Spanish) among the top 100 citations in the field. More than 99% of articles found were in English. However, given that most periodicals require the production of their abstracts and keywords in English, relevant articles could still be found in additional languages and were included whenever applicable.

Types of documents

We considered the following categories in the database Web of Science reviews, journal articles (full articles), and conference proceedings papers.

Keywords

The definition of keywords is a fundamental part of the literature review process. The correct definition of keywords is essential for defining search strings suitable since the strings determine the search results (Biolchini et al., 2005).

We selected keywords for each fundamental construct, using synonyms for each element of the research protocol. For instance, specifically for healthcare simulation, we have used medical simulation, nursing simulation, and healthcare simulation, the three main ways this field is referred to in the specialized literature. The following sections will describe the complete list of keywords employed in more detail.

Inclusion and exclusion criteria

The present bibliometric analysis included documents found with the use of the methods described, which met at least one of the following criteria for inclusion (IC):

- **IC1** – Publications that describe a study related to healthcare simulation
- **IC2 -** Publications that are reviews, meta-reviews, or systematic reviews regarding healthcare simulation
- **IC3** - Publications that present theoretical or empirical contributions related to healthcare simulation

We excluded from the analysis documents that met at least one of the following exclusion criteria (EC):

- **EC1** - Publications in which the search keywords appear in the title, summary, or in the publication's text, but are not related to healthcare simulation, for instance, documents that mention simulation and healthcare but not in the sense of "healthcare simulation" as adopted in this study.
- **EC2** - Publications where the keywords are used just casually mention healthcare simulation but do not have it as its primary focus.
- **EC3** – Editorial, promotional or commercial documents related to healthcare simulation but do not represent studies or theoretical contributions to the field.
- **EC4** – Similar articles (same author and similar theme) to other articles already selected.

Article Database Selection

One of the crucial steps in conducting a bibliometric analysis is to select the databases to provide access to articles. Several databases aggregate citations and compute co-citation analysis, journals, and productivity, among other bibliometric metrics. For the selection of sources, the following conditions were used as conditions for digital libraries:

a) Existence of a search element that allows logical expressions or equivalent mechanisms, including advanced search operators (such as Proximity Operators).

b) Existence of a search element that allows the search of the full text or in specific fields of publications.

c) Publications must belong to one of the publishers listed in the CAPES Portal or the University of Arizona Library services.

d) Search engines must ensure unique results by searching the same keywords.

e) The results and citations can be exported in formats readily usable by the principal bibliometric software available.

Given the criteria, we selected the Web of Science portal, published by Thomson Reuters Scientific.

The Web of Science is a multidisciplinary database composed of three citation indexes: Science Citation Index Expanded (available back to 1990), Social Sciences Citation Index (with coverage going back to 1990), and Arts & Humanities Citation Index (available since 1900). Social Sciences Citation Index (archives since 199). The Web of

Science claims to be the world-leading citation database, with more than 12,000 high-impact research publications globally providing authoritative and diverse coverage. [2].

Bibliometric Software

For this analysis, we employed two software to describe the literature in healthcare simulation: HistCite and Bibliometrix. Histcite version 12 was used for bibliometric analysis and information visualization. The software's primary goal is to make bibliometric analysis and visualization more accessible. It uses fields such as authors, titles, dates, bibliographic references, author and institution locations, and various other variables to investigate any topic of scientific research.

Bilbiometrix, a R package, is the other software utilized. It provides a set of tools for quantitative study in bibliometrics and scientometrics. Bibliometrix allows the importation of data from a variety of scientific databases. Using its user-friendly interface, we used the R program BiblioShiny to generate all the required bibliometric analyses.

Execution of the bibliometric review

After completing the planning phase based on the systematic review model, we defined a protocol and initiated the survey. Details of how the bibliometric survey was performed are presented in the results section of this chapter.

Results of application of the queries of systematic review

The execution of queries was conducted in December of 2021, using the field topic available on the base of WoS. This field includes the title, summary or abstract, keywords of the author, and a unique field of the database, called *KeyWords Plus*, added by journal editors to highlight relevant additional keywords that the authors did not initially list.

We will demonstrate the detailed keyword search process incorporating the knowledge about the synonymous, coincidental keywords and the filtering process. Given that the objective was to present a broad overview of the field, we performed heuristics to remove records unrelated to the theme. This approach proved challenging and labor-intensive as several subfields of science employ simulation (e.g., statistical simulation, numerical or computational simulation, tissue simulation, among many others).

[2] Web of Science - Statistics of access and number of indexed journals http://wokinfo.com/media/pdf/SSR1103443WoK5-2_web3.pdf

The iterative process was deemed necessary when the three general terms (medical, nursing, and healthcare) coupled with the string "simulation" and telesimulation yielded 42,348 records. Then, we employed the NEAR operator – available in the Web of Science Portal - to restrict the results to instances where the words were at most one word apart. This adjustment provided 3.067 documents with higher relevance to the theme.

The following steps (7 and 8) in the process consisted in the inclusion of keywords frequent in healthcare simulation, often used to describe the type of simulation or associated technologies such as: *'"high fidelity", "low fidelity", "human patient", "Standardized Patient", "Skills Training", "Task Trainer", "Manikin", "Virtual Reality".*

These keywords were combined with terms related to healthcare simulation and medical and nursing fields, again using the operator NEAR. The 8th step created a more complete search term by adding variations of the term and the wildcard (*) operator, totaling 4080 records.

In the sequence, the query refinement was performed towards filtering out articles not relevant to the bibliometric review, removing the coincidental appearance of the terms and instances where the keyword appears in the same field but is not related to healthcare simulation.

This process was executed in two complementary approaches. In Step 9, we removed keywords related to different simulation types (e.g., mathematical, statistical, animal model) among other scientific approaches. Next, we removed areas of knowledge that are unrelated to the study theme in Step 10. This process was also iterative, with the unrelated field marked initially filtered, then the resulting articles were inspected before they were marked for exclusion. The 10th step produced a total of 3266 documents.

Finally, the last step (11) in the query selection process was to select only documents with the type of ARTICLES, PROCEEDINGS AND REVIEW ARTICLES. This procedure produced a total of 2,851 documents. The Steps, the keywords, and the associated numbers of the queries are available in Table 2:

Table 2: Steps in conducting the keyword search for the bibliometric analysis

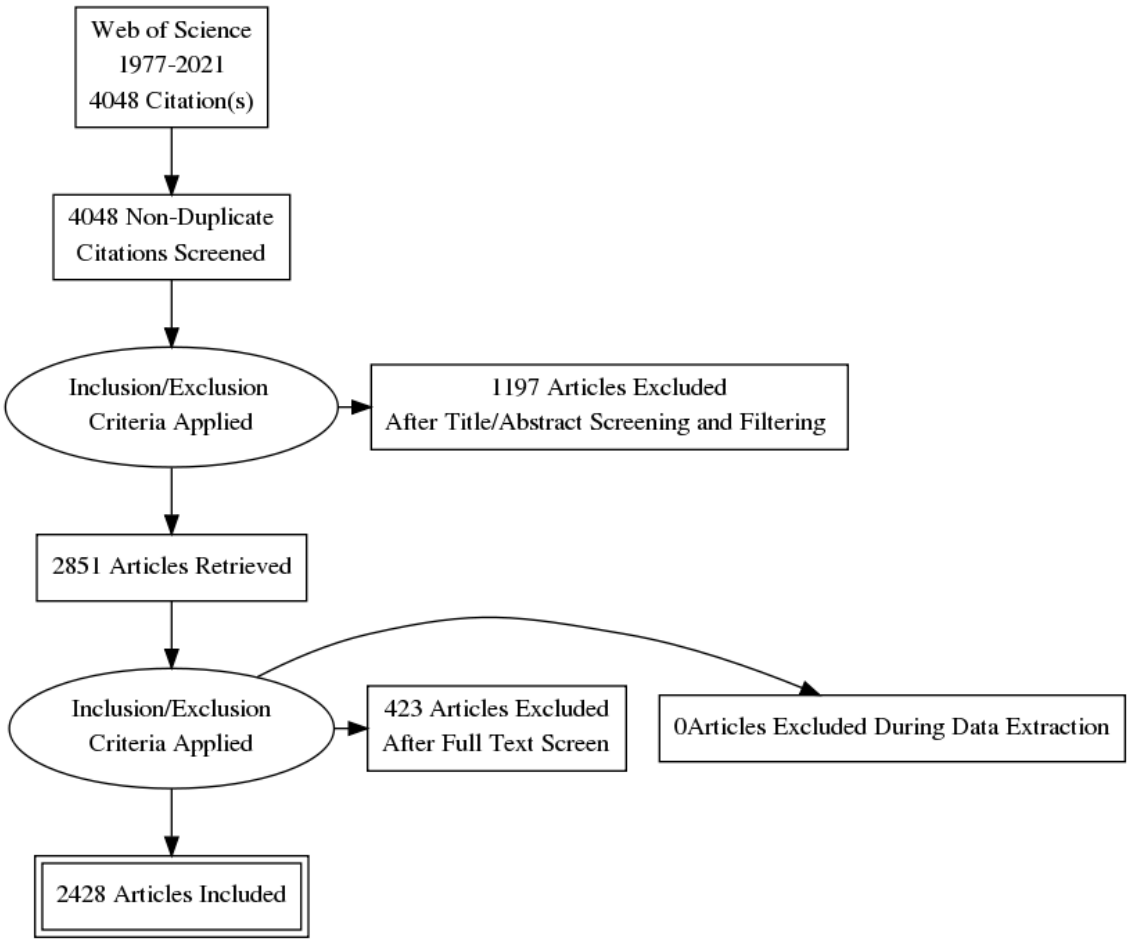
Step	Keyword search	#Articles
1	(healthcare AND simulation)	7,169
2	(medical AND simulation)	32,417
3	(nursing AND simulation)	7,174
4	(“remotely facilitated simulation” OR telesimulation or tele-simulation)	60
5	*Joined the terms into a single query* (("remotely facilitated simulation" OR telesimulation OR tele-simulation) OR (medical OR nursing OR healthcare)) AND (simulation)	42,348
6	*Used operator NEAR – since many results had healthcare and simulation separated, not being useful for the current review.* (("remotely facilitated simulation" OR telesimulation OR tele-simulation) OR (medical OR nursing OR healthcare)) NEAR/1 (simulation)	3,067
7	*After initial analysis of 3067 articles, we found that ‘"high fidelity", "low fidelity", "human patient", "Standardized Patient", "Skills Training", "Task Trainer", "Manikin", "Virtual Reality" are one of the top keywords* ("high fidelity" OR high-fidelity OR "low fidelity" OR low-fidelity OR "human patient" OR "Standardized Patient" OR "Skills Training" OR "Task Trainer" OR "Manikin" OR "Virtual Reality") NEAR/1 (("remotely facilitated simulation" OR telesimulation OR tele-simulation OR medical OR nursing OR healthcare))	3,500
8	*Added wildcards (*) to expand the extent of articles and account for different spellings -e.g., nurse, nursing* (("remotely facilitated simulation" OR telesimulation OR tele-simulation OR medic* OR nurs* OR healthcare OR "healthcare")) NEAR/1 (simulation) (Topic) or ("high fidelity" OR high-fidelity OR "low fidelity" OR low-fidelity OR "human patient" OR "Standardized Patient" OR "Skills Training" OR "Task Trainer" OR "Manikin" OR "Virtual Reality") NEAR/1 (("remotely facilitated simulation" OR telesimulation OR tele-simulation OR medic* OR nurs* OR healthcare OR "healthcare")) (Topic)	4,048
9	*After initial analysis, we produced a list of terms not related to healthcare simulation* Previous query and: not ("stochastic simulation" OR "computer simulation study" OR "medical simulation devices" OR "imaging simulation" OR "image simulation" OR "imaging research" OR "Operations Research" OR "dynamic simulation" OR "medical imag*" OR "Monte Carlo" OR "discrete event simulation" OR " Discrete Choice" OR "porcine" OR "swine" OR "rat" OR "animal" OR "simulation model" OR "computational simulation*" OR "numeric* simulat*" OR "Hotspot Health Physics" OR "Structural Health Monitoring" OR "brain simulation" or "microbiology" OR "finite element" OR "Simulink" OR "Scilab" OR "VisSim" OR "Octave" OR "MATLAB" OR "simulation packages") (Topic)	3,559
10	*Following analysis of the results, we removed categories of research not related to healthcare simulation.* Interdisciplinary Applications Or Chemistry Multidisciplinary Or Endocrinology Metabolism Or Engineering Aerospace Or Nanoscience Nanotechnology Or Biophysics Or Biotechnology Applied Microbiology Or Veterinary Sciences Or Nuclear Science Technology Or Statistics Probability Or Engineering Mechanical Or Mathematics Applied Or Acoustics Or Physics Applied Or Cell Biology Or Construction Building Technology Or Film Radio Television Or Fisheries Or Regional & Urban Planning Or Automation Control Systems	3,266

11	*We have selected the following types of articles: full articles, proceeding and review articles.* *FILTER TYPES OF ARTICLES – ARTICLES, PROCEEDINGS AND REVIEW ARTICLES (remove early access articles) 33 EARLY ACCESS PAPERS – final query 2851 ARTICLES*	2,851

Source: Compiled by the author

Finally, before importing the results into the bibliometric software, we reviewed each article's title and abstract to apply the inclusion and exclusion criteria for the bibliometric analysis, removing a total of 432 articles. The result of the overall process is displayed in Figure **3** in the form of a Prisma diagram:

Figure 3: Prisma Diagram



The application of the filters and the inclusion and exclusion criteria yielded 2428 articles to be included. We transferred five files with 500 records each and one with 428 records due to the 500-item limit of the Web of Science portal. The plain text format was the export format with all fields available in the database, including abstracts and citations.

Later, these files were added to the HistCite software version 12.03, from the first to the fifth file, totaling 2428 records without any error during the import process. The corresponding process was performed on Bibliometrix, using a .ZIP file with the five files.

2.1.2. Bibliometric data analysis

The quantitative analysis of the selected papers is presented in yearly analysis, source analysis, geographical analysis, institutions, and author contributions on the topic of the healthcare simulation. In sequence, the analysis and the process are described.

Bibliometric aspects

Our first analysis is the overall bibliometric indicators of the selected articles, which will be discussed in the upcoming sections. Table 3 has an overview of the main aspects of the bibliometric data:

Table 3: Main bibliometric indicators of the selected articles

Description	**Results**
MAIN INFORMATION ABOUT DATA	
Timespan	1977:2021
Sources (Journals, Books, etc.)	955
Documents	2428
Average years from publication	6.82
Average citations per documents	14.07
Average citations per year per doc	1.47
References	44777
DOCUMENT TYPES	
Article	1773
article; proceedings paper	69
proceedings paper	356
review	230
DOCUMENT CONTENTS	
Keywords Plus (ID)	2225
Author's Keywords (DE)	4473
AUTHORS	
Authors	8914
Author Appearances	11169
Authors of single-authored documents	191
Authors of multi-authored documents	8723
AUTHORS COLLABORATION	
Single-authored documents	207
Documents per Author	0.272
Authors per Document	3.67
Co-Authors per Documents	4.6
Collaboration Index	3.93

Documents and source types

In the sequence, we tabulated the results to understand the pattern of publications regarding the type of outlets. Table 4 shows the type of articles in healthcare simulation published from 1977 to 2021.

Table 4: Type of the articles

Document Type	Frequency	Percentage
Article	1773	73.0
Proceedings Papers	425	17.5
Review Articles	230	9.5
Total	2428	100.00

Table 4 shows that 2428 articles were published in healthcare simulation from 1977 to 2021. One thousand seven hundred seventy-three articles were published in

journals or were presented at conferences or symposiums. Four hundred twenty-five articles were proceeding papers published in conferences, symposia, seminars, colloquia, workshops, and conventions, and a total of 230 articles were review articles and surveys of previously published literature.

Most of the articles, representing 73%, were published in journals, indicating an established and robust field with a more permanent and usually high-impact outlet than proceeding papers.

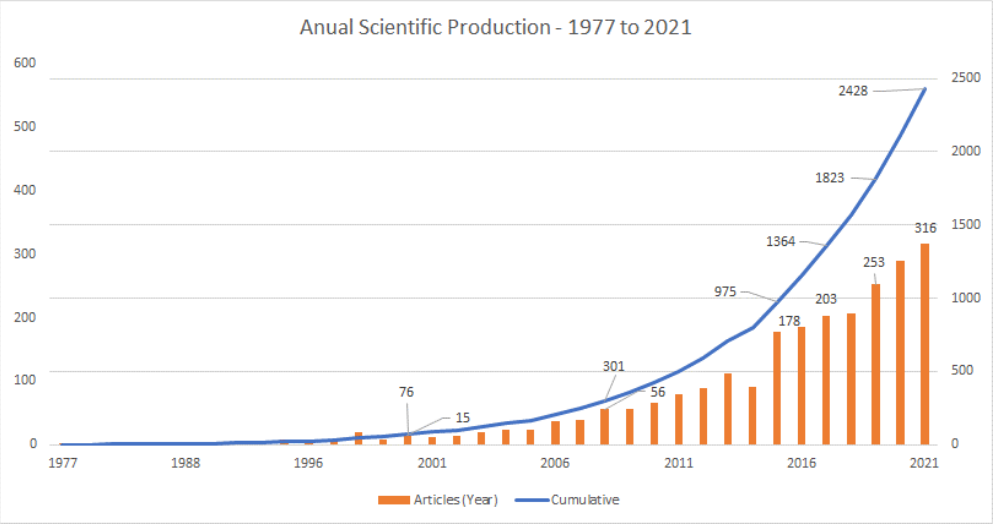
The yearly output of articles

One of the first analyses was related to the yearly output of selected scientific articles. We verified the evolution of articles published from 1977 to 2021. We found 14 articles from 1977-1993 and 17 from 1995-1997. More recently, 289 articles were published in 2020 and 316 in 2021.

An increasing trend in publishing research articles in healthcare simulation can be observed from 2001 onwards. From 2011 onwards, a significant increase in published articles was verified.

We found 172 articles published in the last 12 years, far greater than those published before 2011. The year 2021 had the most significant number of articles published (n=316; 13.1%), showing a growing interest in the field. Figure **4** illustrates the evolution of publications in healthcare simulation from 1977 to 2021.

Figure 4: Articles published from 1977 to 2021



Citation analysis

Another fundamental analysis to understand a given field using bibliometric analysis is to unveil the pattern of citations and co-citations over time and its distribution in different journals, countries, and authors (Zupic & Čater, 2015). The impact of a research document is measured as to how it has been helpful to other researchers (Bornmann et al., 2008). Citations are a most-widely and robust indicator of the usage and influence of a scientific publication (and authors); thus, the number of citations serves as a determiner of the influence of research (Bornmann & Daniel, 2007).

This sub-section will present the principal information about the citation analysis performed. The first analysis performed was conducted to show the general aspects of the data with articles published from 1977 to 2021, regarding citations statistics presented in Table 5:

Table 5: Citation matrix of the published articles

Metrics	Data
Publication years	1977-2021
Citation Years	44 (1977-2021)
Papers	2.428
Authors	9.255
Citations	34,166
Citations/year	776.5
Citations/paper	14.07
Citations/author	3.69
Papers/author	0.26
Authors/paper	3.81

Considering all documents, 9255 authors wrote the 2428 articles. The published articles received in total 34,166 citations. On average, the body of articles received 776 citations per year, with 14 citations received per article. Each author received four citations. Finally, on average, papers were authored by four authors.

Most cited articles

In the sequence, we analyzed the ten most cited articles published from 1977 to 2021. The article “Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review”, written by Issenberg et al., in 2005, received 1858 citations, with an average of 116 citations per year, which makes it the highest cited article in our sample.

In the sequence, the paper authored by Mcgaghie et al. (2010) received 833 citations, averaging 76 citations per year, making it the second most cited article. Finally, the third article in terms of citations received was also written by Mcgaghie with collaborators in 2011, and received 771 citations to date. The list of the ten most cited articles is available in Table 6:

Table 6: Top 10 articles with highest citations

Rank	Author (Year)	Title	Citations	Avg. cites/year
1	Issenberg et al., (2005)	Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review	1858	116.125
2	Mcgaghie et al., (2010)	A critical review of simulation-based medical education research: 2003-2009	833	75.73
3	Mcgaghie et al., (2011)	Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence	771	77,1
4	Ziv et al., (2003)	Simulation-based medical education: an ethical imperative	510	28.33
5	Okuda et al., (2009)	The utility of simulation in medical education: what is the evidence?	484	40.33
6	Bradley, P (2006)	The history of simulation in medical education and possible future directions	463	30.86
7	Motola et al., (2013)	Simulation in healthcare education: a best evidence practical guide. AMEE Guide No.82	347	43.37
8	Steadman et al., (2006)	Simulation-based training is superior to problem-based learning for the acquisition of critical assessment and management skills	284	18.93
9	Aggarwal et al., (2010)	Training and simulation for patient safety	275	25.00
10	Shapiro et al., (2004)	Simulation based teamwork training for emergency department staff: does it improve clinical team performance when added to an existing didactic teamwork curriculum?	273	16.06

Geographical analysis

In the sequence, we analyzed the geographical distribution of the countries according to the author's affiliations to the records evaluated. In Table 7, we present the top 20 countries in terms of published articles:

Table 7: Top 20 countries' production in Healthcare Simulation

Country	Frequency	Percentage (N= 2428)
USA	1161	47.82
Canada	237	9.76
United Kingdom	195	8.03
Australia	132	5.44
South Korea	82	3.38
China	80	3.29
France	75	3.09
Germany	72	2.97
Brazil	49	2.02
Netherlands	46	1.89
Italy	44	1.81
Spain	44	1.81
Poland	38	1.57
Japan	35	1.44
Turkey	33	1.36
Denmark	30	1.24
Switzerland	29	1.19
Singapore	27	1.11
India	26	1.07
Israel	26	1.07

The analysis indicated that the USA was the country of affiliation with 47.82% of the articles, making it the uncontestable leading academic producer in healthcare simulation. Canada published 9.76% of the overall articles and achieved the second position in articles published. The United Kingdom published 8.03% of the overall articles and achieved 3rd position.

The results support the view that USA, Canada, and the United Kingdom are critical contributors to healthcare simulation studies. These three countries' publications are higher than all other countries combined. It is worth mentioning that the publication of healthcare simulation studies is limited to high and mid-higher-income countries, probably related to the high costs and unavailability of simulation centers in mid and lower-income countries.

In the subsequent analysis, we produced a geographical analysis, using the criteria of the number of citations attributed to each country. Table 8 shows the top 20 countries with the highest cited articles in healthcare simulation from 1977 to 2021.

Table 8: Top 20 countries with highest cited articles in Healthcare Simulation

Country	Total Citations	Average Citations
USA	19932	19.037
United Kingdom	3297	19.509
Canada	2005	10.897
Australia	1256	11.849
Korea	904	11.74
China	723	8.407
Israel	584	27.81
Sweden	402	21.158
Italy	395	10.676
Germany	392	6.125
Netherlands	371	12.367
France	353	5.883
Denmark	327	15.571
Switzerland	311	17.278
Ireland	265	12.619
Singapore	263	13.15
Brazil	234	5.707
Japan	197	6.567
Austria	171	13.154
Norway	163	10.867

Table 8 shows that the articles published by the USA received overall 19932 citations to date, which makes it the country with the highest number of citations, aligned with its position as the first ranked country in terms of the number of publications. The articles published by the United Kingdom received 3297 citations overall, which makes it the second country with the highest number of citations.

Articles published by Canadian authors received 2005 citations and ranked 3rd among the countries with the highest citations. The data shows that authors with affiliations in the USA generally published high-quality articles and received the highest number of citations, far more significant than all of the top 20 countries combined.

After determining the countries with the most publications and citations, we analyzed the collaboration network using two metrics available in Bibliometrix. SCP and

MCP indicate the proportion of articles classified as single country publications (SCP) and multiple country publications (MCP).

Articles classified as SCP indicate that no co-authors from other countries participated in the article. In contrast, MCP indicates that at least two countries were involved (determined by the author's affiliation). Table 9 lists the proportion of SCP and MCP research healthcare simulation from 1977 to 2021:

Table 9: Intra-country and intercountry research collaboration

Country	Articles	SCP	MCP	MCP Percentage
USA	1047	969	78	7%
Canada	184	146	38	20.7%
United Kingdom	169	137	32	18.9%
Australia	106	81	25	23.6%
China	86	68	18	20.9%
Korea	77	65	12	15.6%
Germany	64	59	5	7.8%
France	60	50	10	16.7%
Brazil	41	35	6	14.6%
Italy	37	25	12	32.4%
Poland	36	23	13	36.1%
Spain	34	29	5	14.7%
Japan	30	28	2	6.7%
Netherlands	30	23	7	23.3%
Turkey	30	27	3	10.0%
Denmark	21	15	6	28.6%
Ireland	21	20	1	4.8%
Israel	21	20	1	4.8%
Singapore	20	11	9	45.0%
India	19	17	2	10.5%

Starting from the country with the highest number of articles, the USA published a total of 1047 articles, with 969 single country publications (SCPs), whereas 78 were multiple country publications (MCPs), corresponding to 7% of the overall publications, one of the lowest rates of inter-country collaboration in the top 20 cohort. The same pattern was present in Canada and United Kingdom publications.

Canadians published 184 articles, with 146 single country publications (SCPs) and 38 multiple country publications (MCPs), representing 20.7% of publications in collaboration with another country. The UK published 169 articles, with 137 (SCPs), whereas 32 were multiple country publications (MCPs), representing 18.9% of the

publications. We found that USA, Canada, and the United Kingdom are the top three countries in article output. However, their MCPs are low, with possible relation to the vast amount of possible collaboration in their own country.

Countries with fewer citations had a different ratio of intercountry collaboration; for instance, Italy published 37 articles (SCP=25; MCP=12), with 32.4% of articles produced with foreign collaborators. Poland also displayed a similar pattern with 36 articles, with 23 SCP and 13 classified as MCP, with 36.1% of their publications using collaborators from other countries. Finally, Singapore had 20 articles published, with 11 SCP and 9 MCP, 45% attributed to MCP.

After determining the ratio of countries that publish with foreign collaboration, we have tabulated pair-wise collaborations, including all the countries in our dataset. The subsequent analysis is diverse since it counts the total of collaborations and not the proportion of articles published in collaboration, as displayed in Table 10:

Table 10: Top 20 countries in terms of collaboration

Country	**Frequency**
USA	273
United Kingdom	97
Canada	69
Australia	42
China	24
Netherlands	18
Poland	18
Germany	16
Italy	16
France	15
Spain	15
Denmark	13
Brazil	10
Belgium	6
Saudi Arabia	6
India	5
Ireland	5
Switzerland	5
Japan	4
Korea	4

Table 10 shows the top 20 countries having the highest number of collaborations. We found 696 inter-country collaborations. Although having a small proportion of

articles produced in collaboration with other countries, still the USA was the country with the highest number of collaborations. The UK had 97 collaborations, followed by Canada with 69 collaborations. To better explain the network of collaborations, we have cross-tabulated the collaboration pairs to underline the collaboration network. Table 11 lists the top 20 pairs of collaborations between countries in healthcare simulation:

Table 11: Top 20 countries with most frequent collaborations

From	To	Frequency
USA	Canada	56
USA	United Kingdom	36
United Kingdom	Australia	18
USA	Korea	13
USA	China	11
Canada	United Kingdom	10
USA	Australia	10
USA	Poland	10
USA	Switzerland	10
United Kingdom	China	8
USA	France	8
USA	Netherlands	8
USA	Singapore	8
Australia	China	7
Canada	Colombia	7
Canada	France	7
USA	India	7
USA	Italy	7
Canada	Australia	6
Netherlands	Denmark	6

Table 11 shows that the USA published 56 articles collaborating with Canada, making them countries with the highest number of collaborative publications. In the sequence, the second most common collaboration was between the USA and UK, with 36 articles in collaborations between the USA with the United Kingdom. The United Kingdom and Australia collaboration were the third most frequent, with 18 collaborations.

Finally, we plotted the collaboration network among the countries selected using the Bibliometrix – Collaboration Structure feature, where each node represents a country. The size of the node represents the number of collaborative publications. The results of the procedure are presented in Figure 5:

Figure 5: Collaborative Network

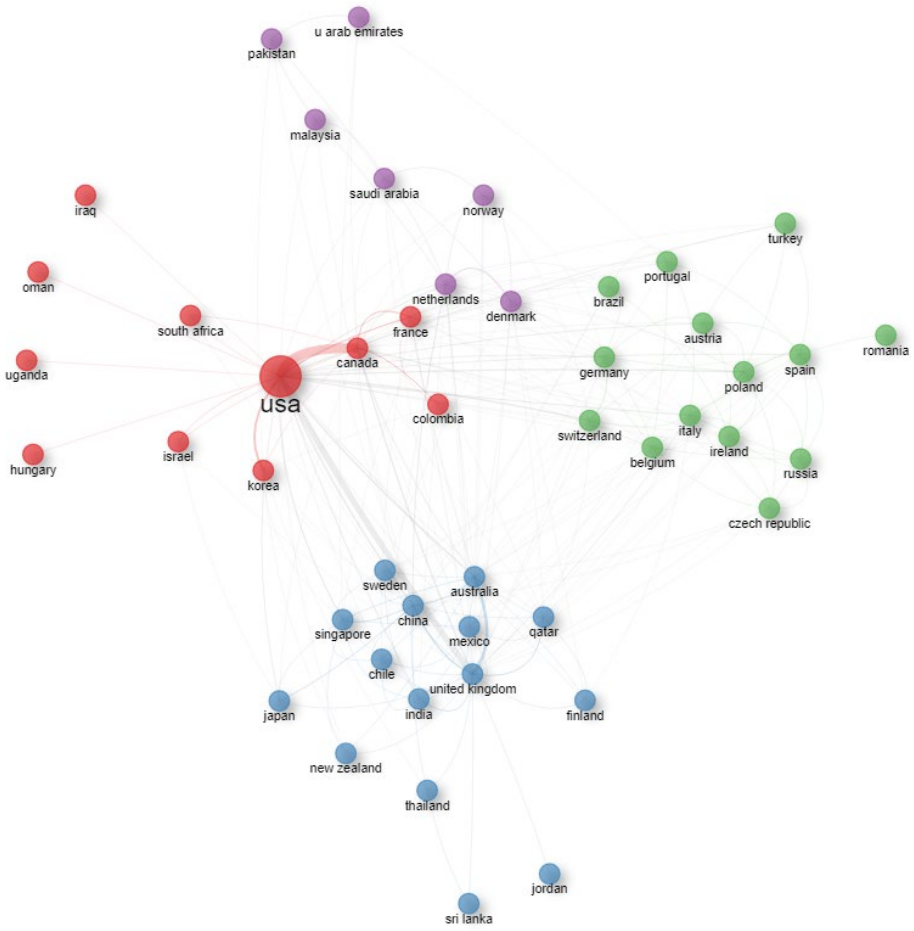


Figure 5 shows that the USA, the United Kingdom, and Canada are influential countries in collaborative efforts with solid ties. The larger node shows that the USA has the highest number of collaborative publications. However, it is interesting to note the cluster of countries represented by the blue color, including the United Kingdom and countries with historical linkages such as New Zealand, Australia, India, and the green and purple countries with less strong ties.

Source analysis

This section has analyzed the primary outlets for publications in the healthcare simulation domain. In the sequence, the top 20 most relevant publication sources in healthcare simulation are available in Table 12:

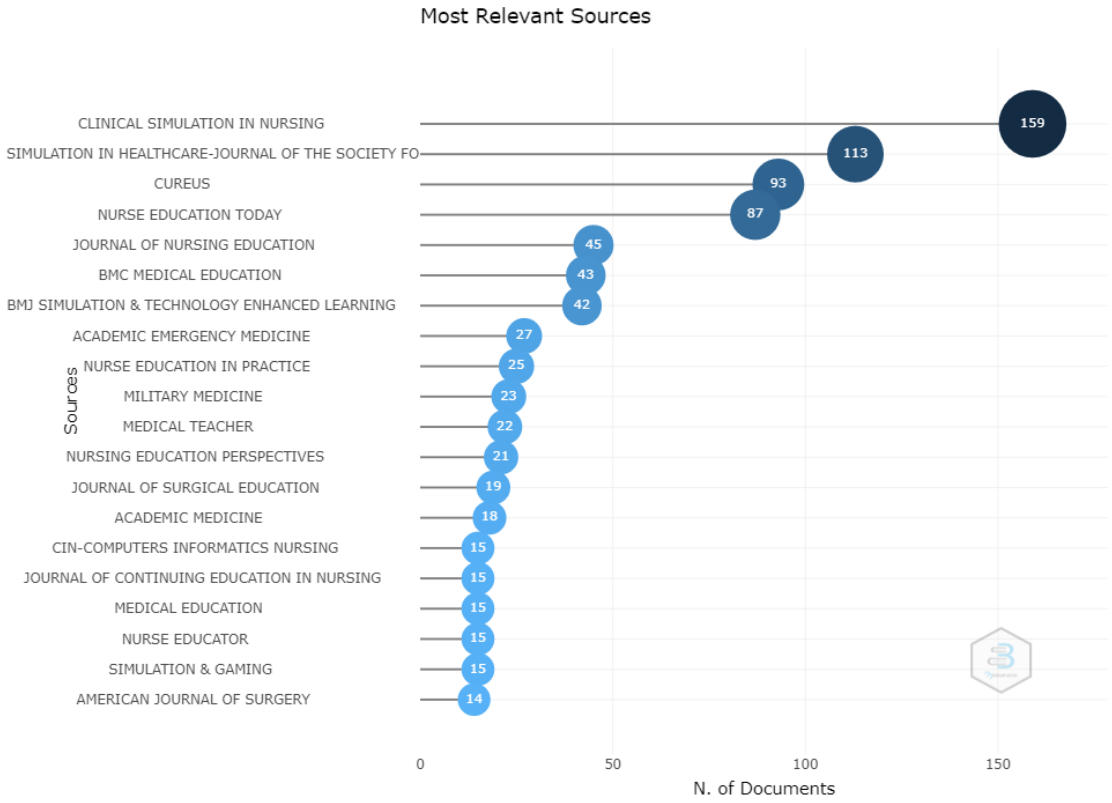
Table 12: Most Relevant Source

Source	Rank	Articles	Cum
Clinical Simulation in Nursing	1	159	159
Simulation in Healthcare-Journal of the Society for Simulation in Healthcare	2	113	272
CUREUS	3	93	365
Nurse Education Today	4	87	452
Journal of Nursing Education	5	45	497
BMC Medical Education	6	43	540
BMJ Simulation & Technology enhanced Learning	7	42	582
Academic Emergency Medicine	8	27	609
Nurse Education In Practice	9	25	634
Military Medicine	10	23	657
Medical Teacher	11	22	679
Nursing Education Perspectives	12	21	700
Journal of Surgical Education	13	19	719
Academic Medicine	14	18	737
CIN-Computers Informatics Nursing	15	15	752
Journal of Continuing Education in Nursing	16	15	767
Medical Education	17	15	782
Nurse Educator	18	15	797
Simulation & Gaming	19	15	812
American Journal of Surgery	20	14	826

The top 20 sources had a total of 826 articles published. The “Clinical Simulation in Nursing” was the outlet with more publications, with 159 articles. “Simulation in Healthcare-Journal of the Society for Simulation in Healthcare” published 113 articles, which ranked it on number 2. “Cureus Journal of Medical Science” published 93 articles, which ranked it on third.

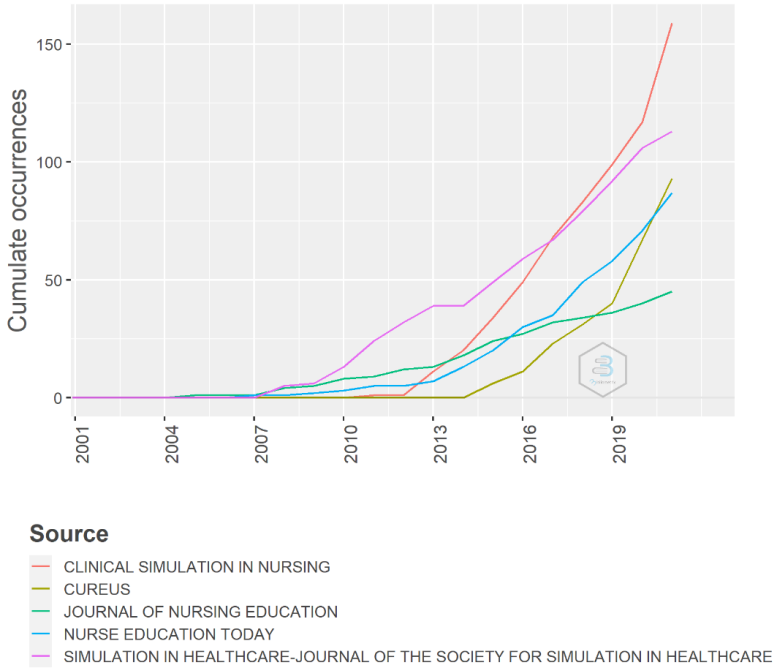
Figure 6 illustrates the same information in a graphic format called “Most Relevant Sources” in Bibliometrix:

Figure 6: Top 20 most relevant publication sources



After unveiling the leading journals containing the selected articles, we have performed an analysis to illustrate the growth of the top five journals in healthcare simulation from the years 2001 to 2021, as displayed in Figure 7:

Figure 7: Growth of top five journals



In Figure 7, each line represents one of the top five journals that host articles related to healthcare simulation. Most journals were not established or had a significant number of publications until 2007.

Publications in "Clinical Simulation in Nursing remained constant from 2001 to 2012. The number of publications greatly increased from 2012 onwards, and 159 articles were published by 2021. Likewise, publications in "Nurse Education Today" increased moderately from 2008 onwards, with 87 articles in 2021.

Articles in the "Journal of Nursing Education" increased from 2006 onwards, and 45 articles were published by 2021. The same pattern was observed in "CUREUS", constant from 2001 to 2014, while the output increased from 2014 onwards, with 93 articles published in 2021.

Finally, we produced a network using the co-citations between journals to understand the dynamics between the different sources, using as the field the source (journals or conferences) with the following parameters: Clustering algorithm – Louvain, Number of nodes = 30, Repulsion force, 0.3. Figure 8 illustrates the co-citation network of sources using Collaboration Structure using the source the basic unit for clustering:

Figure 8: Co-citation network of sources – Medical Education and Nursing clusters

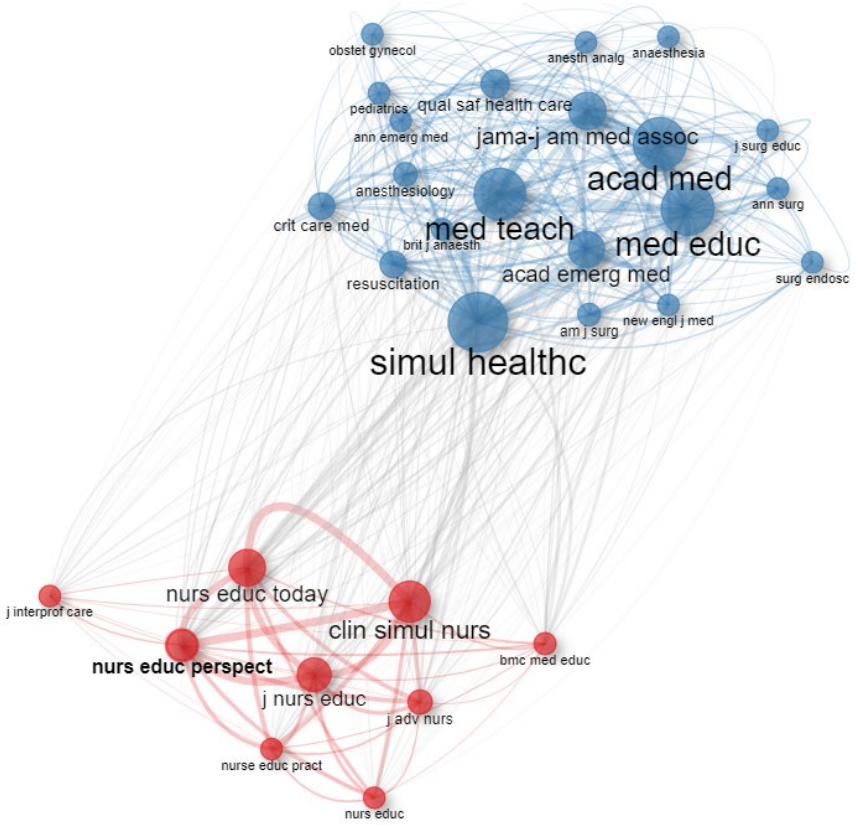


Figure 8 illustrates a co-citation network of sources, where each node represents a journal or a conference. The node size represents the number of published articles, with a larger node representing the more significant number of articles.

We can visualize two well-formed clusters, separating the journals into what we named “Nursing Journals” with journals led by “Clinical Simulation in Nursing”, “Nursing Education Today” and “Nursing Educational Perspective” featured in red.

We labeled the blue cluster as “Medical Education and Healthcare Simulation” since it includes the leading journal in the field: “Simulation in Healthcare-Journal of the Society for Simulation in Healthcare”, along with other prominent journals that are focused on healthcare simulation or medical education like “Medical Education”, “Medical Teacher”, “BMC Medical Education”, and “Academic Emergency Medicine”.

Institutional analysis

This section analyzes the institutional publication and the inter-relationship between the institution, including collaboration networks. In the first analysis, we sought to describe the institutions with the highest product in healthcare simulation. Table 13 shows the data of the top 20 institutions with the highest research productivity in the healthcare simulation field:

Table 13: Top 20 institutions with the highest research productivity

Affiliations	**Articles**	%
Memorial University of Newfoundland	94	1.4%
Northwestern University	87	1.3%
University Toronto	71	1.0%
Harvard University	49	0.7%
University Washington	49	0.7%
Harvard Medical School	46	0.7%
Stanford University	46	0.7%
Johns Hopkins University	43	0.6%
Massachusetts Gen Hosp	42	0.6%
University Of Calgary	42	0.6%
University Of Pittsburgh	41	0.6%
Monash University	38	0.6%
University Cent Florida	38	0.6%
Mayo Clinic	37	0.5%
University Michigan	35	0.5%
University Miami	34	0.5%
Queens University	33	0.5%
Uniformed Serv University Health Sciences	32	0.5%

University Illinois	32	0.5%
Boston Children's Hospital	31	0.5%

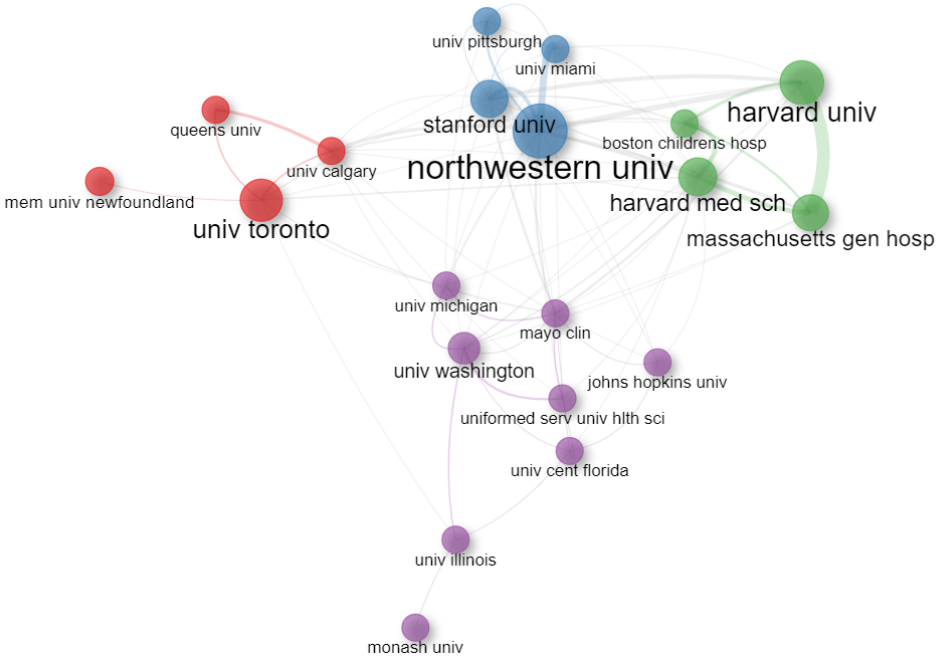
Notes: Total Institution Count = 2,467 unique institutions using the BIBLIOMETRIX's Affiliation Name Disambiguation option –
Total number of articles– 6,827 (one article is counted more than once according to the number of authors and affiliations)

The analysis demonstrates that the three top institutions are in Canada, despite the USA being the leader in the field. The top university regarding research output is the "Memorial University of Newfoundland" with 94 articles, closely in the second place, the "Northwestern University" published 87 articles, followed by the "University Toronto" with published 71 articles.

The remaining institutions are primarily in the United States, including Harvard University, the University of Washington, and Harvard Medical School. This concentration of first-level research universities as leaders in the field speaks to the importance and relevance of the scientific field. In the sequence, we generated a collaboration analysis using the institution as the level of analysis.

We have limited the number of institutions to 20 to visualize better the collaboration network based on the author's affiliation and article co-authorship. The parameters were 20 nodes, repulsion force = 0.2, and the clustering algorithm was Louvain with normalization association with the results available in Figure 9:

Figure 9: Top 20 countries active in collaborative research



Like in the previous analysis, each node represents an institution, and the links between nodes represent a collaboration (operationalized by co-authored articles). The size of the node represents the number of articles published.

As expected from the frequency analysis, we can apprehend that “Northwestern University”, “University Toronto”, and “Harvard University” are prominent institutions in publishing collaborative articles. However, the collaborative network allows additional information about the collaborations among institutions.

We have identified four clusters: the first contains institutions based in Canada (depicted in red), while the second cluster, in blue, represents a closely-knit group led by Northwestern University and Stanford. The third cluster, in green, unveils a collaboration congregating institutions geographically close in the Boston – Massachusetts area

Authorship Analysis

This section analyzes the authorship in the healthcare simulation field, performing analysis by productivity, citations, and collaboration pattern. The first analysis. The first analysis performed was to uncover the number of authors per article or size of the author team, displayed in Table 14:

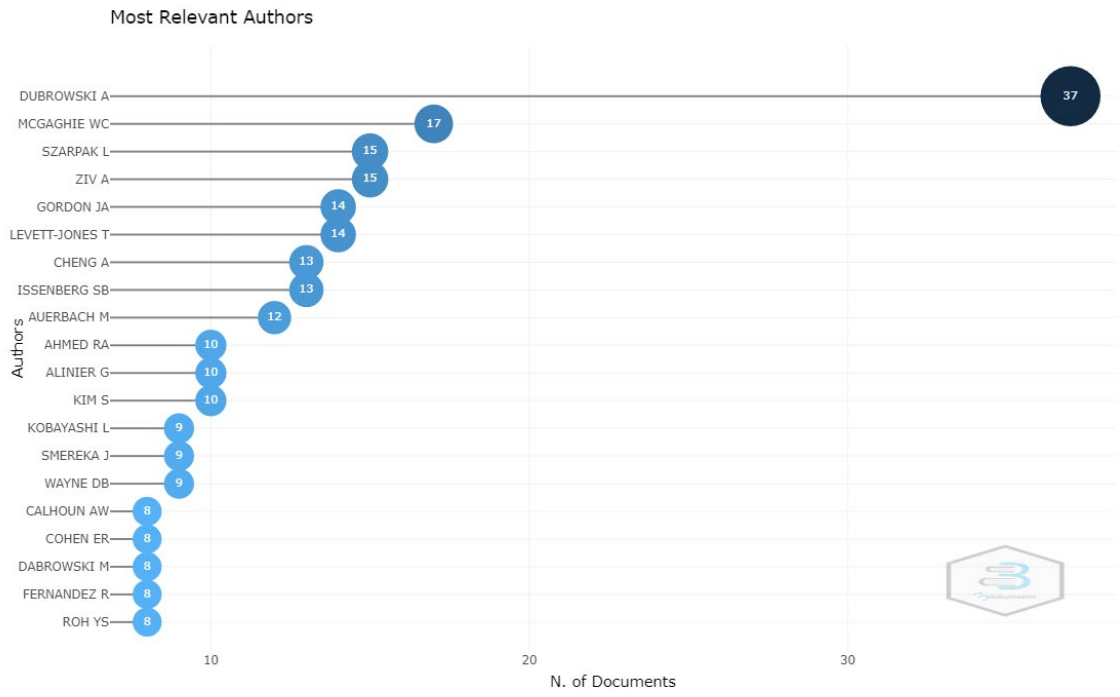
Table 14: Author-wise publications

Author Count	Frequency	% (N = 2428)
1	207	8.53
2	378	15.57
3	395	16.27
4	377	15.53
5	353	14.54
6	236	9.72
7	189	7.78
8	109	4.49
9	53	2.18
10	35	1.44
11	34	1.40
12	13	0.54
13	11	0.45
14	9	0.37
15	12	0.49
16	4	0.16
17	5	0.21
18	3	0.12
20	3	0.12
22	1	0.04
		100.00

Articles authored by one author accounted for 8.19%, while 15.84% of the articles were authored by two and 16.59% of the articles were written by three authors, cumulatively responding for 40% of the articles. The most frequent configuration was to have a group of three authors, although one publication had 22 authors listed.

Next, we executed a frequency analysis to understand the most relevant authors. In Bibliometrix, the most relevant authors are listed according to the number of publications authored, as displayed in Figure 10:

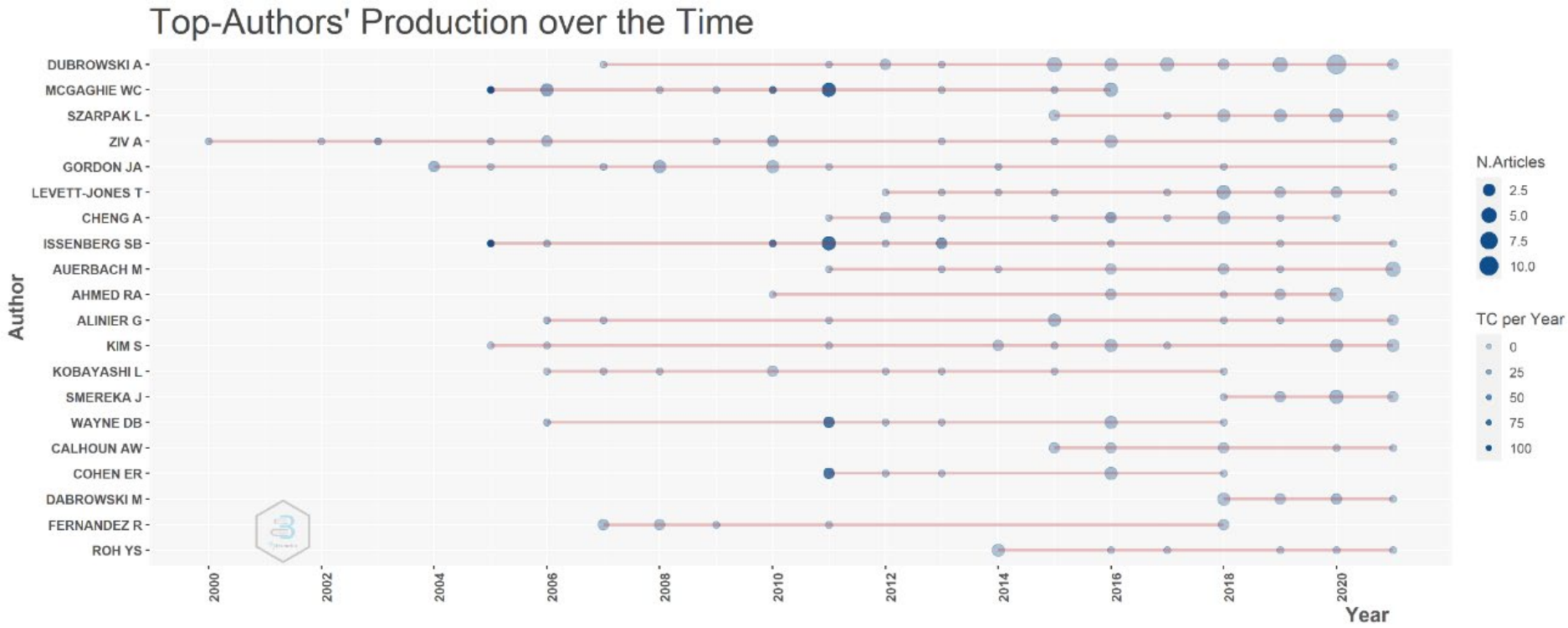
Figure 10: Most Relevant authors



Among the most productive authors, Dubrowski with 37 articles, Mcgaghie with 15, Szarpak with 15, Cheng with 13, and Levett with 13 articles, followed by Isenberg, Sb (US) with 12 documents. In the sequence, we have produced the productivity plot for the authors during different years, allowing us to understand their temporal patterns, including the most cited articles.

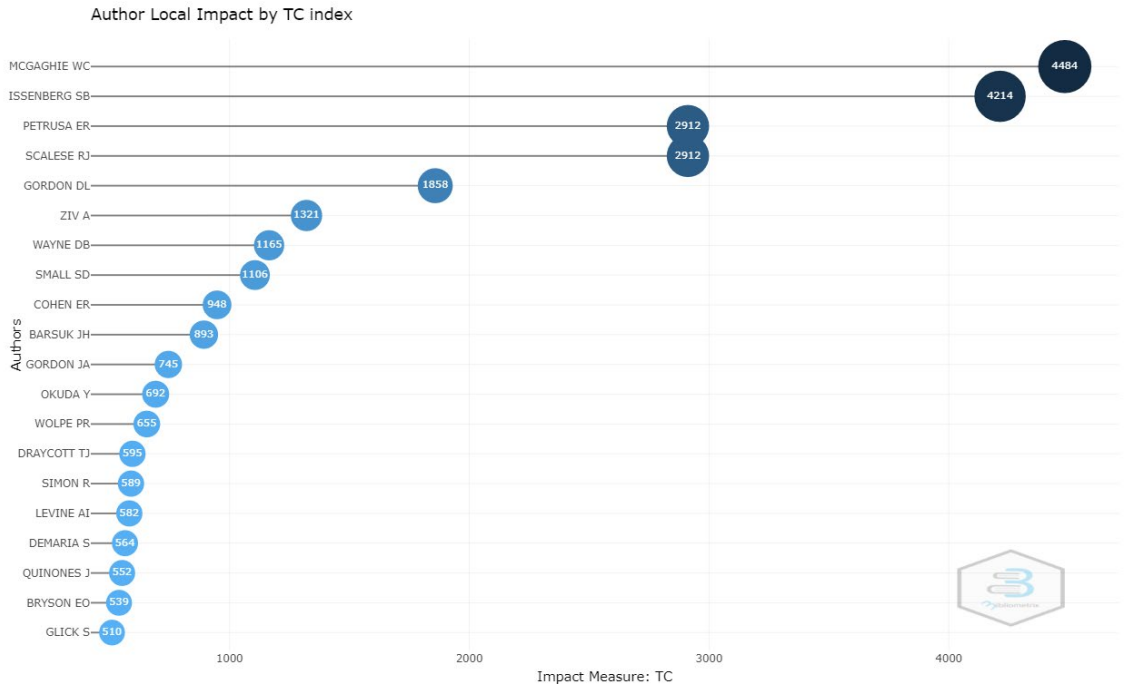
The circle's size is related to the number of publications, while the circle's color indicates the number of total citations during all years considered. The results of the top author's productivity over time are available in Figure 11:

Figure 11: Productivity of the authors over years



Following the analysis of the most productive authors, we classified the authors by the number of citations (total citation count), as displayed in Figure 12:

Figure 12: Top 20 authors – ranked by number of total citations (TC) index



The five most-cited authors carry a disproportionally high number of citations, showing some degree of concentration and leadership among those authors. Articles authored by Mcgaghie had a total of 4484 citations, while articles authored by Issenberg followed closely with 4214 citations. The following two authors, Scalese and Gordon have received 2912 citations each.

The subsequent analysis performed was the citation network among the top 30 researchers on healthcare simulation. We used the parameters numbers of nodes = 30, the minimum number of edges = 2, and the Louvain Clustering Algorithm. Results of the procedure are available in Figure 13:

Figure 13: Citation network among top 30 researchers

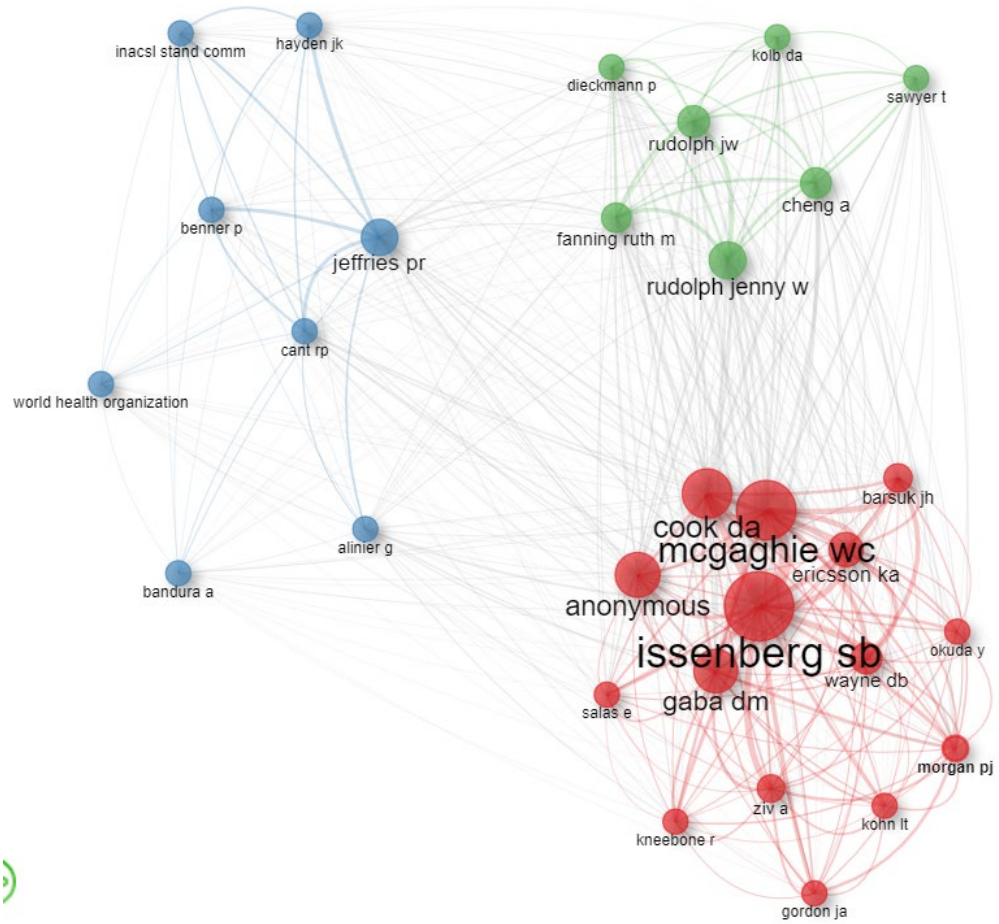


Figure 13 shows the citation network created among the top 30 researchers, where each node represents an author, and a link represents an association (co-citation). The node size represents the number of citations and, the thickness of lines, the number of citations between authors.

We have generated three clusters. “Issenberg”, “Mcgaghie”, “Cook”, and “Gaba” are significant nodes in the first cluster (red color), having publications with a large number of citations and with multiple co-citations among the group, representing the “core” of healthcare simulation.

Relatively less cited authors form the other two cores. The green cluster, led by Cheng, Rudolph, and Fanning, represents another strain of articles, while the third cluster (blue) has Jeffries, Cant, and Benner as its centroid. Although the cluster is more connected than the other two, it is interesting to note a considerable inter-cluster citation pattern.

Finally, we performed a similar analysis. Instead of plotting the network using co-citation as a grouping criterion, we have used co-authorship to determine the different clusters. Figure 14 illustrates the collaboration network among the top 50 researchers:

Figure 14: Collaboration network among top 50 researchers

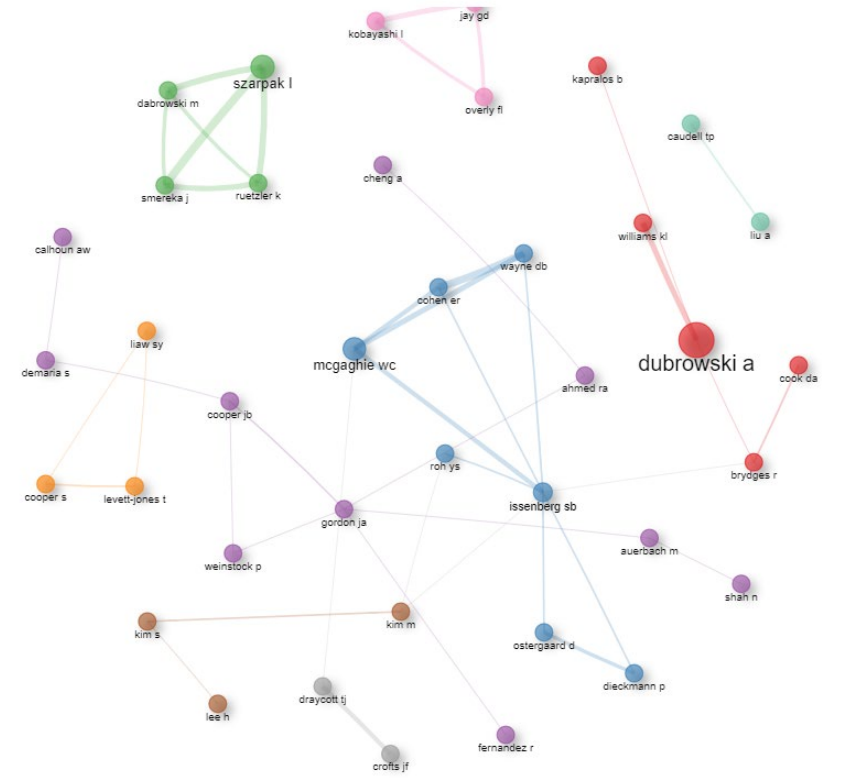


Figure 14 shows a collaboration network among the top 50 researchers, where each node represents a researcher, and the link between researchers represents a co-authorship collaboration. The size of the node represents the number of collaborations. The most productive authors such as “Dubrowski”, “Mcgaghie”, and “Issenberg” are active in collaborative work and have a significant number of collaborations.

The lines' strength demonstrates the collaboration between the authors, mainly the most cited authors and their respective teams/collaborators. We can infer the repeated collaborations in the sample studied from the width of the line.

Keyword analysis

We used two information fields for the keyword analysis to understand the central themes emerging in healthcare simulation: title and keyword plus. We then selected three-word combinations (trigrams) after experimenting with unigrams (single word) and

bigrams (two words). One- and two-word grams usually retrieved keywords such as simulation, health, medical, and nursing without a clear or valuable keywords pattern. Therefore, we opted to generate trigrams – with the top 10 Trigrams in the title field. The results are available in Figure 15:

Figure 15: Top 10 Trigram words – Title

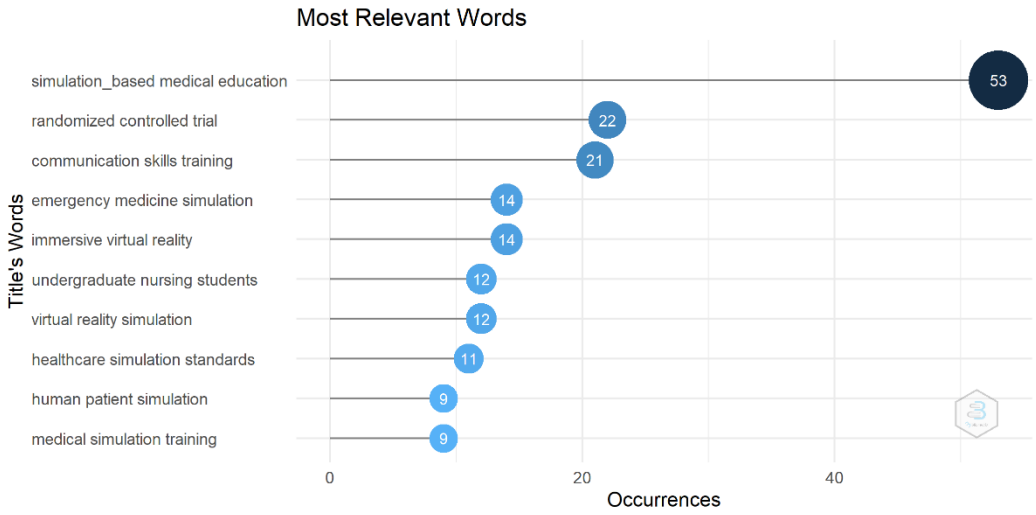


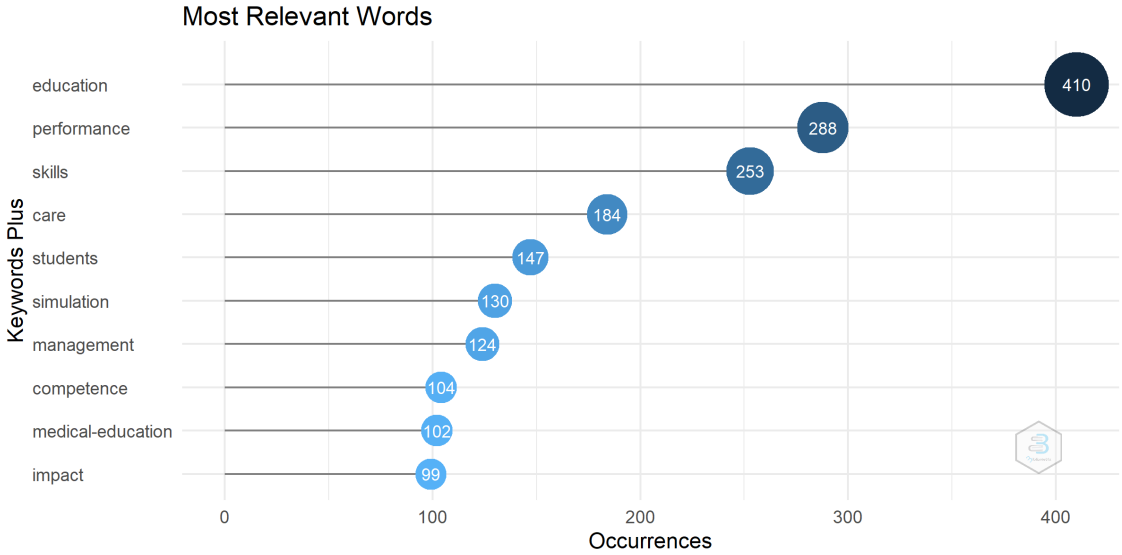
Figure 15 shows the frequency of the top 10 most relevant trigrams. The "simulation_based medical education" term appeared 53 times, making it the most frequent trigram in the top ten trigrams. "random controlled trial" appeared 22 times, which makes it the 2nd most frequent word and shows the prevalence of trials in healthcare studies. Likewise, "Communication skills training" appeared 21 times, placing it as the 3rd most frequent trigram, revealing the importance of this theme to the field. A complementary analysis was performed by producing a Word cloud, as displayed in Figure 16:

Figure 16: Top 10 Trigram words – Title



Following the process, we executed the same analysis for the keyword plus field present in the Web of Science portal. Since it is a curated keyword, we proceeded with one word at a time. Results are displayed in Figure 17:

Figure 17: Top ten most frequent words



As displayed in Figure 17, the keyword “Education” appeared 410 times, making it the most frequent word in the top ten. The word “performance” appeared 288 times, making it the 2nd most frequent word. The word “skill” appeared 253 times, making it the 3rd most frequent word. These words indicate the prevailing concern of the articles

with themes related to education, performance, and student skill. This view was confirmed by the creation of a Word Cloud using the same parameters:

Figure 18: Most frequently words – Keyword Plus Field



Conclusions of the bibliometric analysis

The bibliometric analysis allowed a general, high-level view of the field of healthcare simulation, displaying the most important countries, sources, institutions, authors, and inter-collaboration networks in terms of co-authorship and co-citation.

Our analysis clearly showed that healthcare simulation is a fast-growing field, with multidisciplinary applications, but with the establishment of healthcare simulation as a self-sustaining field, with its journals and conferences. However, we identified an intense concentration of articles in two main groups: medically and simulation specialized journals as a group and nursing and nursing education journals as the other group.

We also identified an engineering and computing component in articles and venues. Themes more related to computation and engineering seem to be virtual reality, human patient simulators (robotic manikins), and technology-enabled education.

We reviewed the profile of institutions and countries that publish the most, indicating that more established institutions, along with higher income countries, dominate the production in the field, attributable to the costs and difficulty accessing simulation centers in most areas. However, simulation education is widespread in the USA, but is not as prevalent in other countries.

2.2. Scoping Review

The usage of scoping review to aggregate research evidence is becoming popular in many scientific areas, especially in cases where the scientific field or the research question is still in its early stages (Pham et al, 2014). One of the first scoping review frameworks, still widely used, was proposed by Arksey and O'Malley (2005).

Scoping reviews are closely related to systemic reviews but are distinct in nature and purpose. Grant and Booth (2009) explain that scoping reviews present a "preliminary assessment" of the type, scope, and size of potentially available research literature in a domain.

Peters et al. (2020) clarify that scoping reviews, also known as scoping studies or mapping reviews, are used to categorize the existing literature in a specific discipline and are endowed with the capability to synthesize research evidence.

They differ significantly from systematic reviews as they provide only a comprehensive review of a particular study area without making any assumptions or assessments on the quality of research. On the other hand, systemic reviews tend to analyze and summarize the whole study to answer a particular research question (Fray and Gore, 2018). Heffernan and associates (2021) pointed out that scoping reviews aim to examine the gaps in existing research and identify the research activity's nature, extent, and range.

Rumrill, Fitzgerald, and Merchant (2010) analyzed the various literature reviews, comparing them with scoping reviews. They explained that traditional review methods such as narrative, empirical, and meta-analyses represent systemic approaches that revolve around a particular topic or research question. On the other hand, scoping reviews are non-systemic and work by utilizing research materials and non-academic materials.

They further suggested that the critical difference between scoping studies and systemic reviews is the variability of used resources. Additionally, scoping studies do not assess the quality of studies used in the review. Thus, these reviews present a viable methodological approach that can be used to assess the breadth of information on a particular topic.

Table 15: Comparison between Scoping and Systematic Review

Scoping Review	Systematic Review
A preliminary step that determines whether a systemic review is possible or not.	Uses a range of research evidence for its execution and analysis.
Identifies and clarifies critical concepts available in the literature.	Focused on a research question or a particular area of study under consideration.
It tends to identify the type of evidence in each field or domain.	It tends to uncover the evidence.
Identifies and analyzes potential gaps in existing knowledge.	Identifies conflicting findings and areas for future research.

Source: Adapted from Rumrill, Fitzgerald, and Merchant (2010)

Therefore, a scoping review is a form of knowledge synthesis that aims to answer exploratory research questions. Calquhoun (2016) explained that the method identifies conceptual underpinnings, evidence and their types, and loopholes in research related to any specific area. The review tends to search, match, select, and synthesize the existing knowledge and generates conclusions accordingly.

Additionally, Cheung and associates (2019) advise that before conducting a scoping review, it is imperative to have clear and broad research questions and the exclusion and inclusion criteria to be adopted in data extraction. While discussing the advantages of these reviews, Peter et al. (2020) defend that it is particularly beneficial if the main subject area does not have enough literature to conduct a complete systemic review. However, it still allows the adoption of inclusion and exclusion criteria with broad research objectives and questions and may organize the available data in charts or tables.

Scoping reviews are particularly beneficial for identifying research gaps in the existing literature to facilitate and set the scope of future studies. Furthermore, Tricco and associates (2018) indicate that scoping review may be essential for gathering and summarizing key findings from a target source even when the research methods or discipline is heterogeneous. They also expressed a concern that there is a general lack of consensus and standard reporting guidelines regarding the methodological framework of scoping reviews.

Again, scoping and systematic reviews employ similar approaches to locate and assess all relevant literature relative to a research question using robust and transparent procedures. The main distinctions between the two procedures are due to their different

intents and objectives. A scoping review, for example, is used to map the body of literature on a topic area. In contrast, a systematic review summarizes the best available research on a question (Pham et al., 2014). They are distinct from systemic reviews as they are not focused on precise questions but complex concepts underpinning the field of research.

A scoping review may be carried out to determine the given research question's characteristics, extent, range, and other evidence parameters. Fray and Gore (2018) emphasized that scoping reviews develop conceptual boundaries, definitions, and prospects of the topic or discipline at hand. Peters et al. (2020) mentioned some reasons to conduct a scoping review:

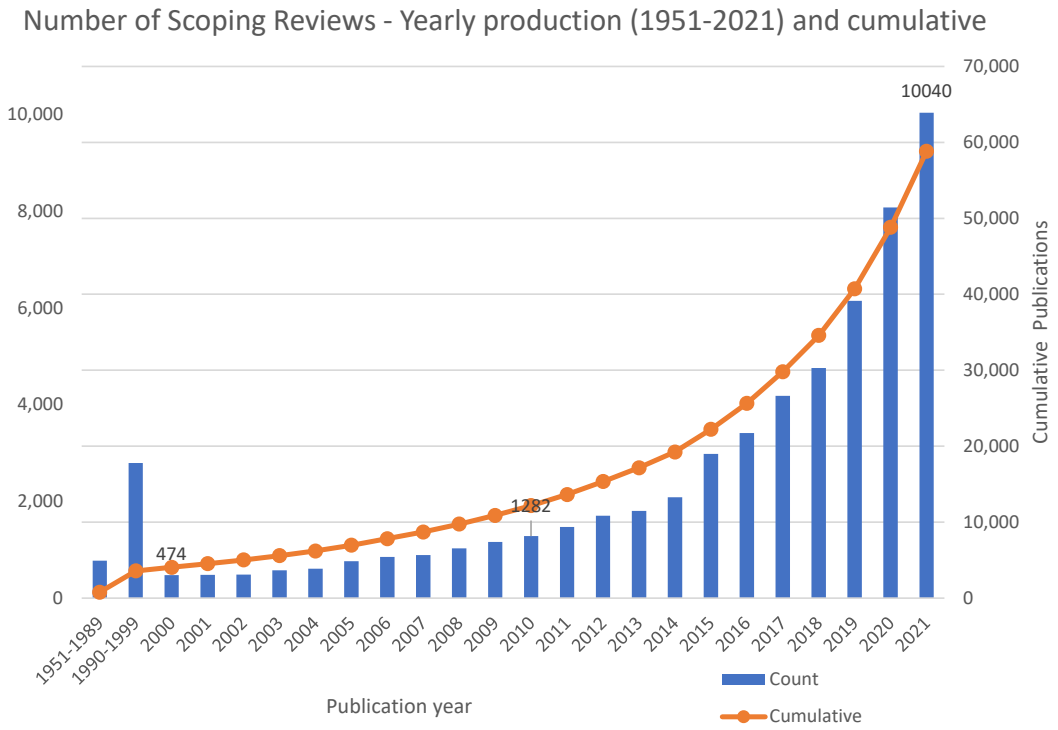
- To identify the gap in the present body of knowledge regarding a subject
- To analyze the knowledge gaps and develop subject scope
- To clarify, explain, or correct subject concepts or definitions in the literature
- To point out the type of evidence available on the given topic
- To identify various parameters related to any topis
- To study up the research methods in any topic or field

Considering the scoping review objectives, we searched the Web of Science portal for scooping reviews in all fields of knowledge during January/2022. We employed the string "scoping review" to establish the prevalence of this type of study, distributed in different research areas, countries, and years of publication.

We found 58,982 publications using the search term "scoping review". The results were distributed in 31,676 review articles, 21,361 articles, and 4244 proceeding papers, while the remaining results were distributed in categories such as meeting reviews, abstracts, editorials, and others).

In the sequence, the evolution of this type of article in recent years is shown in Figure 19:

Figure 19: Evolution in the number of scoping review studies



Source: Web of Science (WOS) – Query conducted on 3/1/2022

Of these records, most articles were published in English 56,300 (95.45%), followed by 681 articles published in Spanish (1.15%), 390 in German (0.66%), and 377 in Portuguese (0.63%), and 373 in Russian (0.63%).

The countries with highest publication counts were the USA (16,646 - 28.22%, followed by England (7,579 - 12.85%), Canada (6,162 - 10.45%), Australia (5,098 - 8.64%) and India (3,721 - 6.31%). The University of London (n=1614) , University of Toronto (n=1570), Univeristy of California System (n=1374), Harvard University (n=944) and University of Sydney (n=707) being the five most affiliations.

Regarding the research area, the three most frequent were Engineering (n=4167 - 7.06%), Chemistry (n= 3,763 - 6.38%), Public Environmental Occupational Health (n= 3,735 - 6.33 %). However, a considerable number of articles were published in Health Care Sciences Services (n = 3235 – 5.48% - 6th place in the ranking), Computer Science (n=2,085 – 3.5% - 10th place) and Educational Research (1,845 – 3.12% - 15th place).

The preliminary results have demonstrated the prevalence and growth of scoping review, including its significant usage in related fields such as Engineering, Computer Science, and Educational research, preparing us to conduct our scoping review in sequence.

2.2.1. Scoping Review Protocol

We have followed the JBI[3] Manual for synthesis evidence to undergo a scoping review. This manual is considered one of the key resources with authoritative information regarding systematic and scoping reviews. The JBI Manual for Evidence Synthesis gives authors instructions on conducting and preparing systematic reviews and evidence syntheses. Various chapters of the JBI Manual for Evidence Synthesis are devoted to the synthesis of various forms of evidence and to answering various sorts of review questions (Aromataris & Munn, 2020)

According to Aromataris & Munn (2020), a scoping review protocol must be defined a priori, as required in well-conducted systematic reviews. A scoping review protocol is essential because it pre-defines the review's objectives, methodology, and reporting and ensures that the process is transparent.

A protocol with extensions specific for scoping reviews was created: *Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews* (PRISMA-ScR), operationalized in the PRISMA-ScR checklist, as a powerful tool to support a strategy to perform and report scoping reviews, with adequate transparency and rigor.

It is beyond the scope of this chapter to detail all the processes involved. For that, we recommend that the reader follow the PRISMA-ScR references. Table 16 lists the main sections and recommended contents:

Table 16: Sections of PRISMA-ScR

Section	Items to be covered
Title	Title
ABSTRACT	Structured summary
INTRODUCTION	Rationale, Objectives
METHODS	Protocol and registration, Eligibility criteria, Information sources, Search Strategy, Selection of sources of evidence, Data charting Process, Data items, Critical appraisal of individual sources of evidence, Synthesis of results
RESULTS	Selection of sources of evidence, Characteristics of sources of evidence, Critical appraisal, within sources of evidence, Results of individual sources of evidence, Synthesis of results
DISCUSSION	Summary of evidence, Limitations, Conclusions
FUNDING	Funding

Source: Compiled from PRISMA ScR Extension Fillable Checklist (http://www.prisma-statement.org/documents/PRISMA-ScR-Fillable-Checklist_11Sept2019.pdf)

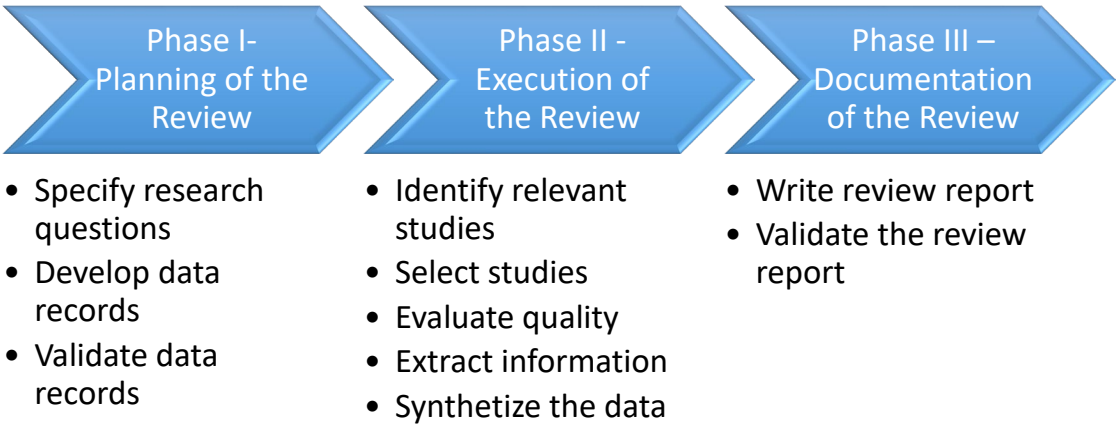
[3] JBI (Joanna Briggs Institute) is a global evidence-based healthcare research organization that collaborates with more than 70 universities and hospitals throughout the world (known as the JBI Collaboration). - https://jbi-global-wiki.refined.site/space/MANUAL

Additionally, given the similarity with the systemic review process, we applied the ten steps Brereton et al. (2007) defined for conducting a systematic review, with the due adaptations for a scoping review, considering its objectives. These ten steps are grouped into three phases: planning, implementation, and review documentation. In the planning phase, the researcher identifies the needs for revision and the review's objective and specifies and develops a data extraction form to conduct the systematic review.

The execution of the review is detailed in phase II. The search in the sources defined is executed, and the obtained studies were evaluated according to the inclusion and exclusion criteria set out in the protocol. Subsequently, the information found in the selected materials was extracted and summarized using the appropriate extraction form.

In phase III, the review results are documents. The review results were documented throughout the process and published in this section. It is important to note that this scoping review systematizes and explains the process. However, it does not involve analysis and comparison to summarize effect sizes or other statistical measures, as is common in systematic reviews (TRAVASSOS ET AL., 2008). Figure 20 illustrates the generic review process applied in this study.

Figure 20: Literature Review Process



Source: Adapted from BRERETON ET AL., (2007).

2.2.2. Scoping Review Objectives

The scoping literature conducted aimed to identify empirical studies with telesimulation as its central theme. We intended to consult the literature to encircle the

research regarding telesimulation as an upcoming field of research, trying to identify the research themes, methodologies, challenges, and future directions.

Thus, this scoping review aims to identify and summarize the status of empirical research about telesimulation, hopefully serving as a reference to researchers in the field to what has been published and what can be expected in future research agendas by highlighting the gaps.

One of the crucial steps in a scoping review (or a systematic review) is to define a question that limits and answers a fundamental question. In this work, we purposefully have set limits on the reach of the review to answer our specific questions. These limits are operationalized through the exclusion and inclusion criteria detailed in the following sections.

In summary, we filtered out research that was not empirical, purely theoretical, or systematic reviews and studies that only inquired participants about perception or satisfaction without a structured measurement or construct. Finally, we removed studies that established the feasibility of telesimulation by participants' perception instead of an objective or validated measure of a construct that could act as a proxy to feasibility.

At the same time, we have proposed a broad research question aligned with scooping review philosophies “Scoping reviews are useful for examining emerging evidence when it is still unclear what other, more specific questions can be posed for evidence syntheses and valuably addressed” (Peters et al., 2020).

The objective of this scoping review is not to provide an answer to a specific and well-defined question but to provide a path to future questions to be asked, and more specifically, to establish the importance and need of the current research to improve the understanding of telesimulation through new lenses.

2.2.3. Scoping Review Methods

This section will present the methods for executing this scoping review to ensure transparency and reproducibility. The subsections include inclusion and exclusion criteria, information source selection, and selection process.

According to Pai (2004), one must verify if one was already published before conducting a systematic review. Following their guidance, we searched PubMed, Web of Science, and ProQuest for scoping reviews. We found only one article published in Medical Education Online in 2021: *Acceptability, technological feasibility, and*

educational value of remotely facilitated simulation-based training: A scoping review, published in *Medical Education Online* (Heffernan et al. 2021).

This article's objective was to provide a narrative/iterative summary of the current literature in assessing the acceptability, educational value, and technological feasibility of telesimulation.

Although related to the present effort and methodologically sound, Heffernan and collaborator's scoping review was limited to three aspects of telesimulation (*acceptability, feasibility, and perception of education value*). At the same time, our objective was broader and tried to answer a fundamental question: What is the current body of empirical research on telesimulation?

To limit the overlap with the identified scoping review, we have excluded articles in our review that aimed to verify feasibility or perception from participants (implemented as exclusion criteria) exclusively. Therefore, after a careful analysis of the scoping review and our objective, it was deemed that another scoping review was necessary, and the process of producing a new scoping review continued.

Selection of information sources

To follow PRISMA-SCr and JBI guidelines, we have selected the information sources and will describe them after verifying the conditions established. For the selection of sources, we determined the following as conditions for digital libraries:

a) Existence of a search element that allows logical expressions or equivalent mechanisms.

b) Existence of a search element that allows to search the full text or in specific fields of publications.

c) Digital Libraries must belong to one of the publishers listed in the CAPES Portal or The University of Arizona Libraries (digital libraries accessible to the author).

d) Search engines must guarantee unique results using the same keywords.

Considering these features and the desire to obtain the largest possible selection of articles, we selected the following databases: PubMed, Web of Science, and ProQuest.

PubMed is a database hosted by the National Library of Science, part of the National Institutes of Health – U.S. Department of Health & Human Services. More than 33 million citations for biomedical literature from MEDLINE, life science journals, and online books are available in PubMed. Links to complete text information from PubMed Central. It is only of the most readily available biomedical resources on the planet. The

PubMed journal list includes all MEDLINE® titles and additional non-MEDLINE titles in PubMed, with around 30,000 journals cataloged.[4]

The Web of Science is a multidisciplinary database composed of three citation indexes: Science Citation Index Expanded (available back to 1990), Social Sciences Citation Index (with coverage going back to 1990), and Arts & Humanities Citation Index (available since 1900). Social Sciences Citation Index (archives since 199). The Web of Science claims to be the world-leading citation database, with more than 12,000 high-impact research publications globally providing authoritative and diverse coverage. [5].

Finally, ProQuest Central combines many of ProQuest's most popular databases to produce the most comprehensive, diversified, and current interdisciplinary research resource accessible. Thousands of full-text scholarly journals, trade and professional publications, dissertations, working papers, case studies, and market reports are included in the collection. In 2021, ProQuest listed 12,000 journals, more than 2 million working papers, 300,000 dissertations, and 130,000 case studies[6].

Review Question and the Search string

Following Pai et al. (2004) guidance, we constructed a straightforward, focused question to generate a search term to be utilized by the database engines. Again, using their advice, we employed the PICO (patient, intervention, comparison, and outcome) framework. It is a popular way to identify the four critical parts of a solid clinical question. The patient/population (or illness of interest), the intervention (or exposure), the comparator intervention (if relevant), and the result should be included in the protocol. To meet these goals, the references in our review should provide answers to our main review question:

Main review question:

(Q1) "What is the extent of the body of empirical studies in the telesimulation?"

Scope

Problem: Retrieve references that respond to the main question of the scoping review, more specifically with the following objective:

[4] PUBMED.GOV https://pubmed.ncbi.nlm.nih.gov/

[5] Web of Science - Statistics of access and number of indexed journals http://wokinfo.com/media/pdf/SSR1103443WoK5-2_web3.pdf

[6] ProQuest – About ProQuest https://about.proquest.com/en/products-services/ProQuest_Central/

(Q1) Find research reports describing empirical research related to telesimulation.

In the sequence, we employed the PICO framework to produce the search query with the terms available in Table 17:

Table 17: PICO components for the search question

Patient/Population	**Intervention**	**Comparison**	**Outcomes (Results)**
Nurse, Medical Doctors, and other healthcare professionals engaged in simulation education	Telesimulation	Comparison with in-person simulation	Empirical studies

According to Pai et al. (2004), the four essential elements to structure the search query should be connected by the AND logical operator[7], as well, resulting in the following structure: (population) AND (intervention) AND (comparison) AND (results).

We employed several strategies according to the PICO component to generate terms for each component. To account for the **population,** we have defined a list of healthcare professions and introduced them to the query in a preliminary manner. However, as we performed the initial queries, we determined that this component could be omitted since it would not significantly impact the results' specificity. We could risk not including a study that involved a health profession non-pre-determined.

Likewise, for the **comparison** component, we discovered that although virtually all telesimulation studies employ comparisons with the in-person method, sometimes the comparison is implicit and therefore not captured in the search terms; therefore, we have not included any terms for this component.

However, we have initially included the term “telesimulation” for the intervention term in subsequent iterations. We included synonymous and alternative forms by which telesimulation has been defined in the literature, such as remotely facilitated or remote simulation.

Finally, the **outcomes** component was defined as empirical studies. However, including this keyword did not retrieve enough articles. After consideration of the best approach, we have included a comprehensive, but a not exhaustive, list of terms that are likely present in empirical studies, such as participants (and its variation, for instance:

[7] The logical operator **AND** or is used when it conditions the presence of two or more terms so that the result is obtained. Ex. "SIMULATION" AND "SURVEY" - only returns results in which both terms appear.

subject, sample), statistical tests (e.g., ANOVA, correlation, p-value), and quantitative (e.g., survey, Likert, scale, pretest) and qualitative (e.g., focus group, Delphi, evaluation") research methods to account for a variety of research.

Therefore, we produced a search string for the three data sources selected: Web of Science, ProQuest, and PubMed. For each of the two elements of the PICO framework, the respective keywords were combined with the logical operator OR[8], and this combination resulted in the following search strings found in Table 18:

Table 18: Search strings used for the different databases

Database	Intervention	Results/Outcomes
Web of Science	(Telesimulation OR tele-simulation OR "remote simulation" OR "remotely facilitated")	(subject OR participant OR RCT OR trial OR experiment OR survey OR sample OR interview OR " n=" OR ANOVA or "test" OR "p-value" OR "scale" OR Likert OR "assessed" OR study OR correlation OR regression OR cluster OR means OR treatment OR evaluation OR Delphi OR "focus group" OR "pre-test" OR "post-test" OR "pretest" OR "posttest")
PubMed	(Telesimulation OR tele-simulation OR "remote simulation" OR "remotely facilitated")	(subject OR participant OR RCT OR trial OR experiment OR survey OR sample OR interview OR " n=" OR ANOVA or "test" OR "p-value" OR "scale" OR Likert OR "assessed" OR study OR correlation OR regression OR cluster OR means OR treatment OR evaluation OR Delphi OR "focus group" OR "pre-test" OR "post-test" OR "pretest" OR "posttest")
ProQuest	(Telesimulation OR tele-simulation OR "remote simulation" OR "remotely facilitated")	(subject OR participant OR RCT OR trial OR experiment OR survey OR sample OR interview OR " n=" OR ANOVA or "test" OR "p-value" OR "scale" OR Likert OR "assessed" OR study OR correlation OR regression OR cluster OR means OR treatment OR evaluation OR Delphi OR "focus group" OR "pre-test" OR "post-test" OR "pretest" OR "posttest")

Inclusion and exclusion criteria

We identified the eligibility criteria following the JBI Manual recommendations by establishing inclusion and exclusion criteria. All the documents retrieved by the execution of the search query were included in the appraisal process, being selected if they met all the following inclusion criteria **(IC):**

[8] The **OR** operator is used the presence of one of the search terms is enough for the result to be obtained. Ex. "Government" OR "open" - returns results in which at least one of the two terms

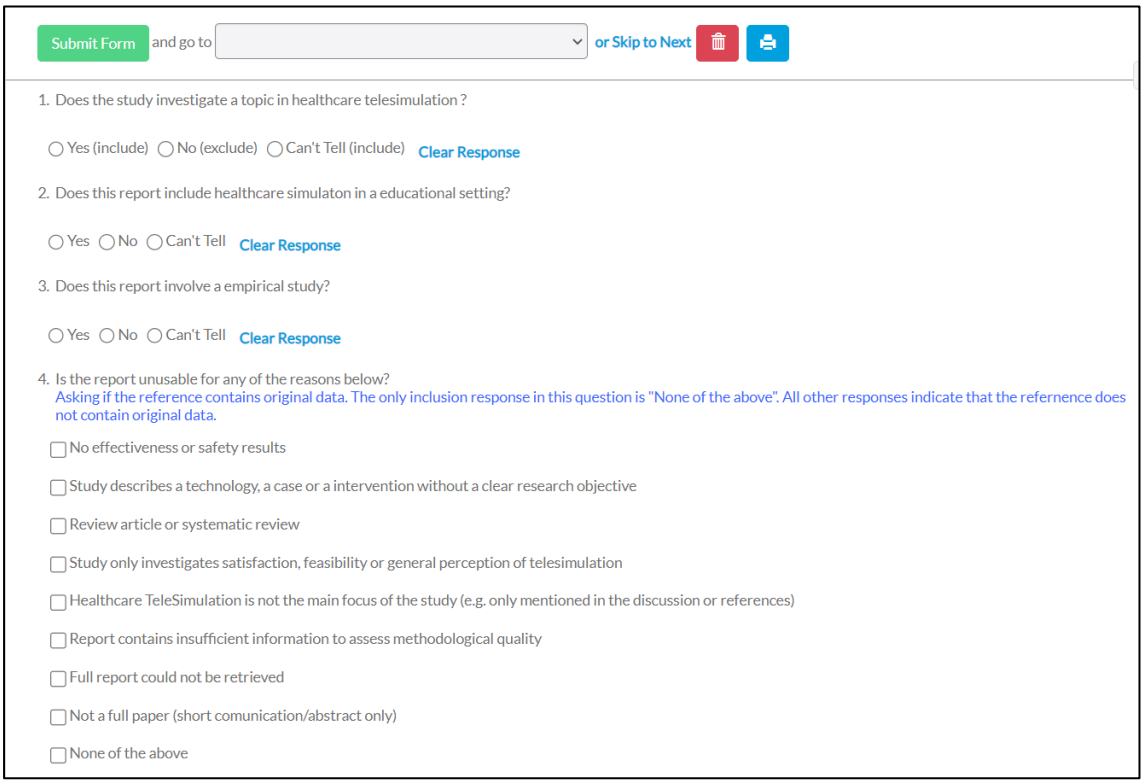
- **IC1** - Publications that describe studies that investigate a topic in healthcare telesimulation?
- **IC2** - Publications that include telesimulation in an educational setting
- **IC2** - Publications that present an empirical study

Likewise, all publications that satisfied at least one of the following exclusion criteria **(EC)** were excluded**:**

- **EC1** - Publications in which telesimulation is not the focus of the study (e.g., only mentioned in the discussion or references)
- **EC2** –Publications that are duplicated (from different data sources)
- **EC3** - Publications that feature studies that describe a technology, a case or an intervention without a defined research objective
- **EC4** - Publications that exclusively review articles or systematic reviews (without a telesimulation intervention/experience)
- **EC5 -** The study only investigates satisfaction, feasibility, or general perception of telesimulation from participants or instructors
- **EC6 –** Publications that are not considered a full paper – for instance, short communication/abstract

The inclusion and exclusion criteria were also implemented in the software DistillerSR to facilitate the selection process during the Level 2 screening (full-text review). The form utilized is available in Figure 21:

Figure 21: Custom Form to Screen References – Level 2



Source: DistillerSR

Scoping review supported by computer-supported tool

We have selected the DistillerSR software (on the cloud – available at https://www.evidencepartners.com) to support the scoping review process and generate the diagrams (PRISMA) and statistics such as inclusion and exclusion criteria. DistillerSR uses AI and intelligent processes to automate literature collection, triage, and evaluation. It has templates and customized forms to facilitate the data extraction process. The researcher can extract data and record repeated, complicated data sets with in-form validations and computations. It also allows to create reference lists, outcomes, and summary tables in seconds and know where each data item came from (DistillerSR, 2021). Thanks to a generous trial time granted to academic students, the student edition was used for this research, using the Google Academic Account available to UNIRIO graduate students.

Evaluation of the quality

There was no assessment of the articles' quality in this study. We assumed the journals and conferences indexed in the prestigious database were reliable. The articles

had undergone rigorous peer review processes, indicating the adequate quality of a selected to contribute to the present review.

Information Extraction strategy

The information extraction strategy phase was designed to collect the data needed to answer the research questions and assess the quality of the study. According to the JBI Manual for evidence synthesis (Aromataris and Munn, 2020), scoping reviews should extract the following study information: Author(s), publication year, study's country of origin (where it was published or carried out), aims/purpose, study population and sample size (if appropriate), methodology/methods, type/duration of intervention, comparator, and outcome measures (if applicable), outcomes and specifics (e.g., how they were measured) (if applicable).

Additionally, given the research specificity, we summarized whether there was a COVID-19-related study, whether the study conducted a comparison, and statistical tests to compare the treatments, limitations, and other fields. For each document or selected study after the implementation of the selection process, we extracted the information presented in Table 19:

Table 19: Fields used to extract information

FIELD	INFORMATION
Author(s)	
Year of publication	
Origin/country of origin	
Mentions COVID	
Direct comparison with in-person simulation	
Study type	
Educational activity	
Type of Simulation	
Aims/purpose	
Study population and sample size	
Data Source	
Methodology/methods	
Statistical Tests to compare interventions	
Intervention type/duration, comparator, outcome measures	
Outcomes and details (e.g., how measured) (if applicable)	
Limitations/GAP	
Key findings that relate to the scoping review question/s	

Source: Compiled by the author

Extracting information allows for collecting the information required to answer the research questions. This information collected from selected articles was compiled in chronological order and grouped according to research themes. Results were tabulated using the DistillerSR tool to facilitate the analysis and inserted in the designed form. Based on these results, we set strategies for quantitative and qualitative analysis of the body of articles selected.

The qualitative analysis considered the articles' context to discuss the research questions' findings. The quantitative analysis presents the literature review results comparing locations of conduction, methodologies, and years, among other relevant information.

2.2.4. Execution of the scoping review

After completing the planning phase and the definition of the scoping review protocol, we executed the search strategy. Details of the execution are available in the results section.

Execution of queries

The queries in the selected databases were executed in November 2021, using the "all fields" option available on PubMed, ProQuest, and Web of Science. We have inserted the queries into the web of science and executed the query related to Q1 through the following string:

(Telesimulation OR tele-simulation OR "remote simulation" OR "remotely facilitated")

AND

(subject OR participant OR RCT OR trial OR experiment OR survey OR sample OR interview OR " n=" OR ANOVA or "test" OR "p-value" OR "scale" OR Likert OR "assessed" OR study OR correlation OR regression OR cluster OR means OR treatment OR evaluation OR Delphi OR "focus group" OR "pre-test" OR "post-test" OR "pretest" OR "posttest")

The results returned the following number of results: ProQuest search yielded 183 results, PubMed search returned a total of 155 results, and Web of Science search produced 75 results in 413 articles. All the results were limited to full text and peer-reviewed articles (when applicable in the search database).

Record Identification and Removal of Duplicates

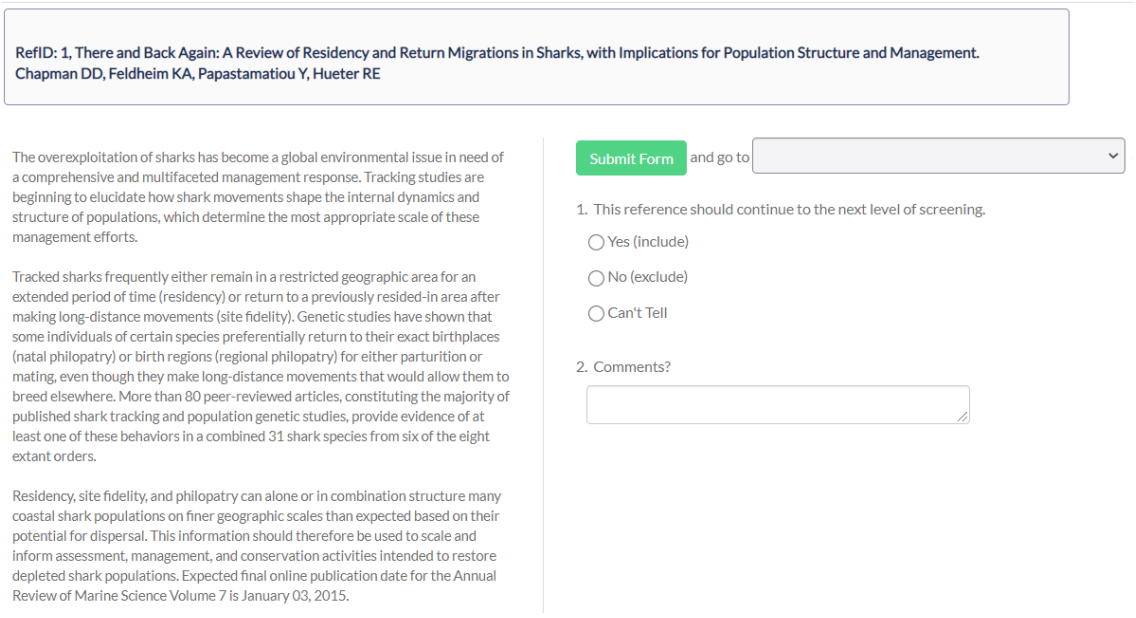
After importing the files from the various database, the first task was to remove duplicates. We employed the Duplicate Detection feature of DistillerSR to support the process with the following parameters: a) compare title, author, and abstract; b) remove words shorter than 3 and c) exact match: on. Then we ran the Run Dupe Check function and manually verified all references for duplicates. A total of 123 articles were removed, rendering the total of articles to screened to 290.

Screening of Results

The next step, defined as Level 1 analysis in the DistillerSR is the fast screening of references. This process is streamlined by reading the title and abstract and filtering out items that do not conform to the theme, not requiring a complete review or application of any inclusion or exclusion criteria.

The operationalization of this phase was made by answering the question, "This reference should continue to the next level of screening?" The possible responses are Yes or No. The form utilized in this research is available below, with a sample article included in the DistillerSR package to illustrate the process:

Figure 22: Level 1 – Preliminary screening of References



Source: DistillerSR

All the references were subject to the Level 1 screening, with a rapid screening of the publication title. The results are displayed in Table 20:

Table 20: Reference Screening Level 1

Question	Answer	% (n)
This reference should continue to the next level of screening.	Yes (include)	33.10% (n=96)
This reference should continue to the next level of screening.	No (exclude)	66.90% (n =194)

Source: DistillerSD software output

Application of Inclusion (IC) and Exclusion Criteria (EC)

In the next step of the analysis, we applied the exclusion criteria and inclusion criteria, known as Level 2 in the DistillerSR software, using the specific form developed for this research by accessing the full text. Statistics related to the application of the inclusion and exclusion criteria are available in Table 21:

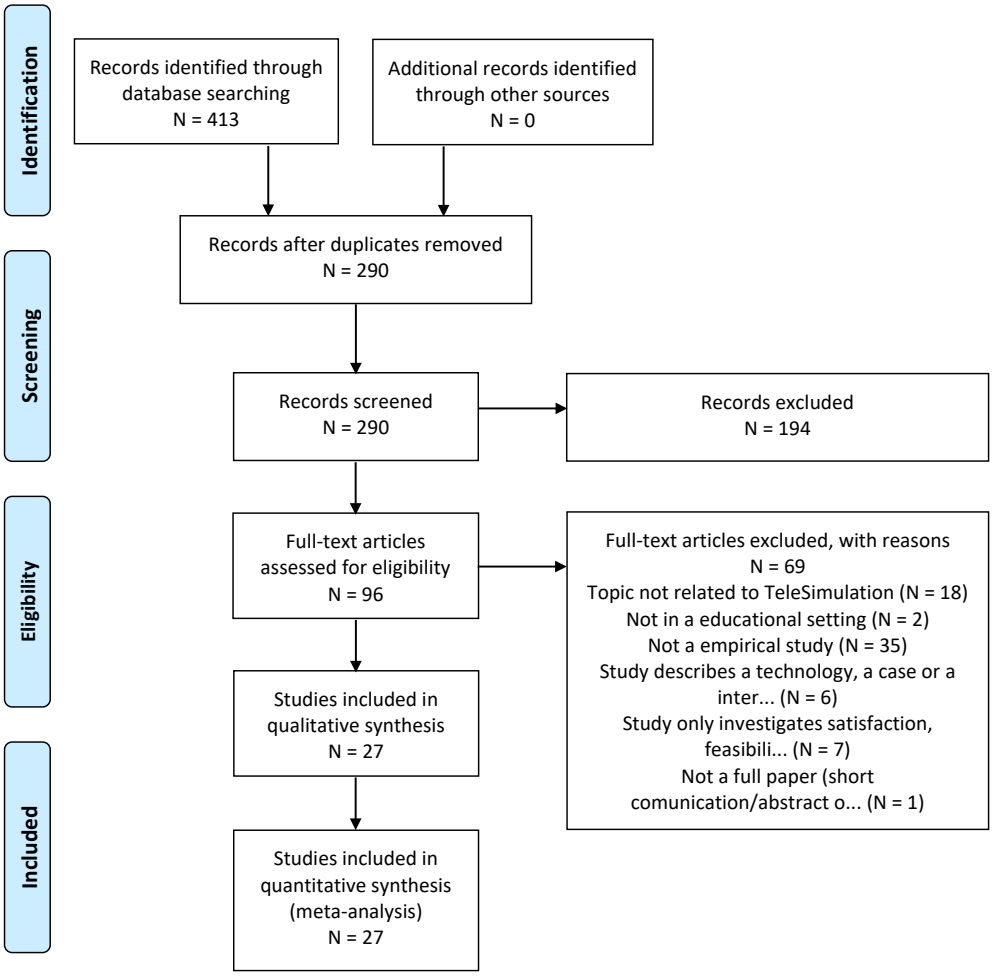
Table 21: Application of Inclusion and Exclusion Criteria

Question	Answer	Total	%
Does the study investigate a topic in healthcare telesimulation?	Yes (include)	78	81.25%
Does the study investigate a topic in healthcare telesimulation?	No (exclude)	18	18.75%
Does this report include healthcare simulation in an educational setting?	Yes (include)	83	86.46%
Does this report include healthcare simulation in an educational setting?	No (exclude)	13	13.54%
Does this report involve an empirical study?	Yes (include)	57	59.38%
Does this report involve an empirical study?	No (exclude)	39	40.63%
Is the report unusable for any of the reasons? (exclude)	Study describes a technology, a case, or an intervention without a clear research objective (exclude)	15	15.63%
	Review article or systematic review (exclude)	4	4.17%
	Study only investigates satisfaction, feasibility, or general perception of telesimulation (exclude)	12	12.50%
	Healthcare Tele Simulation is not the focus of the (exclude)	2	2.08%
	Not a full paper (exclude)	5	5.21%

Source: DistillerSD Software Output

After applying inclusion and exclusion criteria, 27 articles were selected to be examined and part of this review. The selected articles were appraised by re-reading their full text and extracting the main ideas presented using the extraction form. The process, displayed in a PRISMA diagram, is present in Figure 23:

Figure 23: PRISM Diagram for the Scoping Review



Source: Output from DISTILLERSR software – Detailed PRISMA report

After the extraction, we performed the bibliometric analysis to render an overview of the selected articles. We identified the articles in the portal Web of Science to extract citations and other information about the articles. However, out of the 27 articles, only 19 were present in the Web of Science (the other articles were exclusively on Pubmed or PROQUEST) and have not been included in the bibliometric analysis.

Therefore, the analysis presented in the sequence needs to be considered as a sample of the identified articles and not a complete evaluation of the results.

Regarding years of publication, the selected articles were mainly published in 2021 (n= 11 ; 57.90%), with only one published most other years starting in 2011. This shows the surge in interest and research in telesimulation as a response to COVID-19. In the sequence, we present Table 22 with the main publication titles:

Table 22: Publication Titles

Title	n	%
AEM EDUCATION AND TRAINING	2	10.53%
MEDICAL EDUCATION ONLINE	2	10.53%
SURGICAL ENDOSCOPY AND OTHER INTERVENTIONAL TECHNIQUES	2	10.53%
ACADEMIC EMERGENCY MEDICINE	1	5.26%
ACTA ANAESTHESIOLOGICA SCANDINAVICA	1	5.26%
BMJ SIMULATION TECHNOLOGY ENHANCED LEARNING	1	5.26%
BREASTFEEDING MEDICINE	1	5.26%
CLINICAL SIMULATION IN NURSING	1	5.26%
CUREUS	1	5.26%
FRONTIERS IN SURGERY	1	5.26%
GLOBAL PEDIATRIC HEALTH	1	5.26%
JOURNAL OF MEDICAL INTERNET RESEARCH	1	5.26%
ONLINE LEARNING	1	5.26%
PEDIATRIC EMERGENCY CARE	1	5.26%
SIMULATION IN HEALTHCARE JOURNAL OF THE SOCIETY FOR SIM. IN HEALTHCARE	1	5.26%
TEACHING AND LEARNING IN NURSING	1	5.26%

Out of our sample of 19 articles, we can identify three main types of journals: medical and nursing journals, educational journals (including online and healthcare education), and simulation journals focusing specifically on simulation.

This analysis is complemented by observing the Web of Science categories which agree with the titles of the journals in which we found the publications. Three articles were in each of the Educational Research, Health Care Sciences, and Pediatrics categories. Following Surgery, Educational Scientific Disciplines and Nursing had two articles each.

In terms of countries, most articles were published in the USA (n=10), followed by Canada (n=8) and Australia in third place (n=2). The remaining countries had one publication each and are composed of: Austria, Botswana, Denmark, Japan, Latvia, and Mexico.

The sequence will present the summarization of results for the selected articles that describe empirical studies in telesimulation.

2.2.5. Summarization of Results

This section will present the articles selected for the Scoping review, presenting a simple overview of the study, including context, authors, year, and variables of interest, along with the main findings. Whenever applicable, we also share the limitations and suggestions for future studies.

O'Rae and associates (2021) study was conducted with 101 nursing students using low fidelity skill-based simulation. The study's main objective was to assess the use of telesimulation techniques employed by nursing students to gain adequate competency in delivering family nursing care using a three-point Likert scale to collect data. The telesimulation scenario was designed in line with the standards of ASPE (Association of Standardized Patient Educators) and INACSL (International Nursing Association of Clinical Simulation and Learning). The results indicated that telesimulation is an effective strategy for students to experience and learn therapeutic communication in family nursing.

Another article by Reece and associates (2021) in Alberta, Canada, determined the utility of Virtually Facilitated Simulation (VFS) in delivering high-quality Simulation-based Education (SBE) to professionals in geographically isolated communities.

The study enrolled 200 healthcare providers in 11 rural and remote communities, distributed in 24 sessions. The data collection encompassed qualitative data collected through video reviews and standardized observation tools. After completing sessions, self-reported electronic surveys were conducted to collect demographic data and predict outcome measures related to Covid-19 preparedness. Lastly, they proposed further investigation and development of VFS methods even after the pandemic.

Min Shao et al. (2018) investigated video-enabled remote simulation training programs in multiple countries, including Asia, Europe, Central America, and South America. The study had two main objectives a) to determine the feasibility of a video-enabled remote simulation training program and b) to evaluate its role in developing a standardized approach for managing critically-ill patients.

The sample had 18 local professionals from 7 countries and eight hospitals. They found that improvements in critical task performance are attainable through this simulation training: 94% of clinicians felt more confident and prepared to handle critical cases after training. They concluded that video-based remote simulation training offers a feasible, economical, unique, and indulging framework to train critical care practitioners for critical skills in international settings.

Similarly, Harris and colleagues (2020) studied video-based simulation and determined its effectiveness in training clinicians in interprofessional late-life depression screening. The study selected medical students and nurse practitioners, comparing in-person and video-enabled technology simulations using a pretest-posttest approach. They compared scores before and after participation in standardized patient simulation to detect depression in senior citizens. The results were evaluated based on ICCAS (Interprofessional Education Collaborative Competencies Attainment Survey), indicating no significant difference between remote and in-person simulation outcomes.

Musa et al. (2021) explored another dimension of video-based simulation studies. They investigated the diminished interaction in online learning due to the COVID pandemic and conducted an interactive video simulation study. Participants were 36 undergraduate nursing students with prior knowledge and experience in health care simulation. The study was comparative, and the subjects participated in two different simulations, i.e., a video simulation (VID) and an interactive simulation (INT). Students experienced more gains in the INT condition than in the VID conditions. Thus, they concluded that interactive video simulation is more effective for remote healthcare learning.

Gross and colleagues (2020) studied telemonitoring through telesimulation to determine its effectiveness as a tool for remote instructor development. The study was longitudinally conducted over eight simulation sessions in 16 months. Sessions were facilitated by both a local instructor (LI) and a remote instructor (RI). Initially, the RI took up much of the responsibility for simulation sessions, but gradually, more responsibility was shifted to the LI. The students completed a general feedback instrument and a DASH-SV (Debriefing Assessment for Simulation in Healthcare student version) form at the end of every session.

In contrast, instructors completed the DASH-IV (DASH Instructor version). No significant change in the DASH-IV score was measured as the role of LI increased in simulation. However, the DASH-SV scores gradually declined when the LI increased.

They concluded that telemonitoring through telesimulation can ultimately deliver high-quality debriefing.

A similar article was authored by Samuel and colleagues (2021) at the National Autonomous University of Mexico, exploring telesimulation as an educational strategy and the barriers faced by using this instructional method, with the ultimate objective to create a taxonomy of the identified barriers. A relatively large study sample was used, including 26 standardized patients, 18 professors, and 407 undergraduate medical students. The simulation sessions were divided into three phases, i.e., initial briefing (10min), running the scenario (20min), and debriefing (30 mins).

The feedback was taken through a reflective debriefing, online questionnaire, and critical incidents technique. Resultantly, a taxonomy of barriers included five categories: knowledge, facilities, financing, attitude, and participants. These categories represent the factors that need to be dealt with to guarantee the quality of telesimulation for educationalists. However, the authors indicated that further studies need to be conducted to obtain different perspectives of the pre-identified barriers.

McCoy et al. (2017) further refined the previous studies on telesimulation by evaluating its comparative effectiveness against standard in-person simulation. The key focus of this study was medical students who were being taught and trained in critical care and management. A prospective nonblinded and crossover study was conducted between September 2014 to February 2015. Participants were randomly selected fourth-year students from the University of California medical school. The SIM group took part in a live, engaging simulation scenario, then debriefed with their SIM group.

Additionally, a live TV Internet connection supported the TeleSIM group. The TeleSIM group experience featured remote observation of a live simulation case from an off-site location and a shared group debriefing through a live TV Internet link. A written evaluation method assessed the subjects. The results indicated no significant difference between student preference of SIM and TeleSIM. This study was the first to utilize a direct comparison between simulation and telesimulation within our scoping review.

Another study on telesimulation was carried out by Anderson et al. (2021), focusing on hybrid telesimulation, composed of a pilot study aimed at determining the effect of hybrid telesimulation with an SP (standardized patient) in lactation support. More specifically, the study assessed student engagement and satisfaction with telesimulation. The SPs used Summative Assessment Rubric and Formative Assessment Rubric (FAR) to assess students' clinical and interpersonal skills.

Three simulation sessions were conducted, and the multiple-choice and FAR scores suggest significant improvement between sessions. Findings indicate that hybrid telesimulation was an effective strategy to enhance students' interpersonal skills, with the recommendation to incorporate telesimulation into the existing curricula to aid in developing clinical lactation competencies.

A similar article by Burckett-St.Laurent and associates (2016) focused on laparoscopic surgery using telesimulation. The focus of the study was to assess whether the skills for ultrasound-guided regional anesthesia (UGRA) could be effectively learned through telesimulation. The study platform was developed at the University of Toronto, and four to five anesthetists were recruited in each of the four selected hospitals. The training was held in four online sessions and one on-site session. The 19 participants were assessed before and after training using a GRS (Global Rating Scale) and validated checklist. Both methods delivered statistically higher post-scores after training, demonstrating the feasibility of training skills for UGRA using telesimulation.

Nikouline, Jimenez, and Okrainec (2020) explored telesimulation for laparoscopic surgery training. The study assessed the feasibility of Google Glass (GG) to develop the technical skills required for the Fundamentals of laparoscopic surgery (FLS), comparing it with Skype-conducted simulation, enrolling 28 participants using mixed methods.

The study employed participant-feedback surveys and FLS scoring data to assess performance using remote and on-site proctors. Participants were evaluated on two major FLS tasks. Two main results were found: 1) remote and onsite proctors results were strongly correlated (ICC= 0.985 for peg transfer and ICC=0.969 for suture tasks) 2) there was not a statistically significantly different between GG and Skype, therefore indicating that GG was a feasible option for remote training, even though participants preferred the use of Skype.

Furthermore, Henao and associates (2013) conducted a similar study. They evaluated the effectiveness and feasibility of telesimulation for training surgery students in the Fundamentals of Laparoscopic Surgery Program. The study lasted eight weeks in two cities in Colombia. The study recruited twenty participants, out of which ten were general surgery residents and ten were general surgeons. Pre and post-test scores were recorded through scoring parameters developed by the Society of American Gastrointestinal Endoscopic Surgeons.

The study results showed that all 20 participants passed the skills part of the FLS course, whereas the post-test scores were significantly higher than pre-test ones. Thus, it was concluded that telesimulation is an effective way to improve and optimize the laparoscopic skills of participants even if they have no previous knowledge regarding the FLS program.

Hughes et al., (2021) conducted an extensive analysis of the simulation-based communication skill curriculum, comparing resources and experiences needed to provide a simulation-based communication skill curriculum on death notification in an in-person versus a digital communication platform delivery of the same material. Three key measures were assessed: performance, confidence, and knowledge. Individual learners performed pre-and post-intervention knowledge and confidence tests. Thirty-four students participated in the study, 22 in person and 12 by the digital platform. All three measures of knowledge showed statistically significant improvements in both groups (pre and post-test), indicating that death notification conducted in-person and digital communication platform training was equivalent.

Jewer et al. (2019) discussed the educational efficacy of simulation-based training in acute care procedures, comparing a portable mobile telesimulation unit (MTU) and the conventional face-to-face approach for training. The study employed a randomized controlled design in three simulation sessions. The study sample comprised 69 medical students randomly assigned into control, intervention, and comparison groups. The results indicated that MTU simulation is optimal for teaching procedural and practical skills. Moreover, students trained through both methods showed improved learning outcomes compared to the control group.

Christensen et al. (2015) focused on establishing the effectiveness of SBT (simulation-based training) through distance education. The scenario selected addressed the management of deteriorating patients by newly graduated healthcare professionals. The study comprised 155 participants exposed to a locally facilitated format (LF) and a remotely facilitated format (RF). A 16-item questionnaire was employed, entailing interaction, comfort, learning, and quality of instruction.

The results showed that the experience of using RF (median = 76) was rated slightly less positively than the locally facilitated format (median = 78). The difference was statistically significant ($p=0.001$), but the practical significance seems negligible. Likewise, no statistically significant differences were found in learning outcomes, using the MCQ scores as indicators.

Mileder, Bereiter, and Wegscheider (2021) study examined the effectiveness of telesimulation for neonatal resuscitation training. They assessed participants' attitudes and compared the outcomes between neonatal nurses and medical students. Participants were given baseline knowledge and necessary materials for home training. Then, they received a link to participate in the one-hour telesimulation sessions broadcasted through Cisco Webex.

Overall, participants enjoyed telesimulation, perceived to have gained knowledge, and acknowledged the practical training prospects of telesimulation. Since the study did not employ a control group; thus, the evidence regarding telesimulation producing equivalent cognitive benefits to face-to-face SBT or other traditional teaching techniques could not be established.

A similar article by Ohta and associates (2017) analyzed the efficacy of pediatric simulation using remote facilitation. The study employed two groups and compared remote simulation with simulation using an on-site facilitator. They used two distinct scenarios, and one-hour simulation sessions were given to both groups, followed by debriefing. Performance and participants' perceptions were evaluated. Findings indicate that simulation through remote facilitation significantly improved students' performance. Additionally, the authors recommended that remote simulation be a credible alternative in areas with limited availability of experienced facilitators.

Patel et al. (2020) discussed the effectiveness and feasibility of telesimulation in training and education in a case-based scenario, with two cohorts of anesthesiology department residents with a gap of around two weeks. Four different roles were allocated to the simulators: simulation operator, confederate surgeon, director, and confederate anesthesiologist. All study participants were given adequate medical information before and after the simulation. The study concluded that telesimulation through Zoom is an effective way to provide scenario-based education and training to students, being a potential substitute for hands-on simulation in line with the social distancing protocols of the pandemic era.

Mikrogianakis and associates (2011) determined if pediatricians could use telesimulation to train African physicians in intraosseous (IO) insertion. It was a cross-border study where one simulator was in Botswana while the other was in Canada. Twenty-two participants engaged in telesimulation from various medical domains in Botswana. Each physician filled up a pre-and post-intervention questionnaire. Results indicated that telesimulation sessions significantly increased physicians' knowledge

scores, establishing telesimulation as a potential educational tool for teaching remote technical and procedural skills across borders.

Nadarajan and associates' (2020) study focused on assessing telesimulation for an Emergency Medicine Clerkship transformed to a remote setting as a remedy for Covid restriction imposed on the in-person training. It utilized a mixed design, with 42 responses (out of 58 students). Four clinical presentations were utilized (anaphylaxis, cardiac arrest, meningitis, and syncope), and students responded to open-ended questions about their experiences, which were analyzed. Four themes were identified: Fidelity, Realism, Engagement and Knowledge, Skills and Attitudes Outcomes.

Additionally, most respondents (90%) acknowledged that telesimulation could provide sufficient knowledge before engaging in clinical work, and 98% felt that the simulation was valuable and exciting. One important finding is that telesimulation can significantly improve non-tactile aspects of clinical training, such as teamwork, treatment algorithms, and history taking.

Diaz-Guio et al. (2020) carried out a study concerning online-synchronized clinical simulation (OSCSim) in three Latin American countries (Columbia, Mexico, and Ecuador). The key objective was to assess students' performance when faced with Covid cases. The study employed a before-and-after mixed design in three simulation centers. The sample comprised 106 participants, and six simulation sessions were designed, with Covid as its theme. A 9-point performance scale was used, and results were evaluated based on the ICAP framework. The study concluded that OSCSim is an adequate learning opportunity that can provide declarative knowledge and skills regarding the management of Covid.

Sanseau and associates' (2021) study's primary goal was to test if a remote simulation platform (TeleSimBox) could be used by faculty and if it might be an effective technique for teaching medical students.

This TeleSimBox platform was developed by a team of medical providers and simulationists in 2017 and was used to facilitate drills with emergency departments in rural areas. Students and educators from four medical schools in the United States participated in twenty-one telesimulation sessions to evaluate the educational effectiveness and likelihood of recommending the experience to colleagues and comparing the experience with other distance learning and in-person modalities.

Findings indicate that telesimulation was equivalent to other distant learning and in-person simulations. The remotely facilitated sessions were satisfactory at teaching

medical knowledge and collaboration but comparatively not as effective in teaching family communication and skills.

Riley and associates (2021) explored the impact of transferring an on-campus nursing specialty course to a remote one during COVID. The redesigned methodology included virtual case studies, remote simulations, electronic audience response systems, and recorded and broadcasted whiteboard presentations with Socratic-style questioning. Course instructors undertook a formal research study to analyze the possible impact of the remote design on student performance and satisfaction.

The results showed no significant difference between the course performance and satisfaction of students, upkeeping the learning quality, indicating that telesimulation (and related virtual technologies) have the potential to be employed in any discipline-specific curriculum in an emergency or remote locations.

Padhya et al. (2021) also evaluated the impact of remotely conducted simulation. The study aimed to determine the effects and feasibility of remote simulation training using the Checklist for Early Recognition and Treatment of Acute Illness in Pediatrics (CERTAINp). The training was conducted in 7 centers in five developing countries (China (n=3), Congo (n=1), Croatia(n=1), India(n=1), and Turkey(n=1)) through video conferencing during 2015 and 2016. Performance was evaluated according to standard care guidelines of CERTAINp, using three emergency scenarios (hypotension, respiratory issues, and status epilepticus). Results indicated significant improvements between pre and post-test in task completion, exposure assessment, review of home medications, review of allergies, and review of differential diagnosis. In summary, an overall improvement after remote training was observed.

Pennington and associates (2018) conducted a study using the CERTAIN tool to investigate the feasibility and possible impacts of long-distance, remote simulation training on team dynamics. Interdisciplinary teams underwent acute crisis scenarios before and after training using CERTAIN-ELITE at nine training sites in eight countries. Remote trainers supervised the hands-on usage of the cloud-based decision assistance tool during CERTAIN-ELITE training.

The evaluation of team dynamics was performed through long-distance virtual simulation training. Enhanced team performance during simulated patient emergencies was verified. However, language and video quality were the two most significant impediments identified throughout the review process.

2.2.6. Conclusions for the telesimulation scoping review

The scoping review and the associated queries proved successful as we found various articles presenting empirical research of various domains. One of the significant findings is that most studies were conducted in USA and Canada, representing 16 out of the 27 studies. Another important finding is the proportion of studies that mention COVID-19 as the reason for the implementation of telesimulation (48%). This finding is more pronounced if we only verify studies that were published after 2019.

Another aspect uncovered by the scoping review is that most studies were conducted to verify the feasibility, effectiveness, and learning outcomes of telesimulation. We did not find any studies that investigated emotional responses. The construct closer to emotion was satisfaction. Still, the articles selected usually employed simple measurements for satisfaction, often finding that participants were relatively content and supporting the use of remotely mediated simulation.

Pre- and post-test designs were frequent in the articles reviewed, usually divided into tests of improvement of learning and associated outcomes in telesimulation. With fewer studies comparing the two conditions directly (n=8) compared to 19 not directly performing the comparison (either using the pre-posttest or using the participant's own scores as a comparison).

In the next paragraph, we present a summarization of the main findings of the presented articles, which have implications for understanding the status of empirical telesimulation studies:

- Observable common research foci on feasibility, satisfaction, and effectiveness of telesimulation, often compared with in-person simulation.
- Smaller sample sizes range from 13 participants to 800.
- Studies were primarily conducted in developed countries (USA and Canada – 16 out of the 25 studies).
- The benefits of telesimulation were another critical thematic component, namely lower costs, access to centers, and training in underserved locations and developing countries, and rural settings
- Almost 50% of the articles (12 studies) have studied the use of telesimulation in response to COVID-19
- Considerable growth in telesimulation studies starting in 2019 as a response to COVID-19's higher usage of telesimulation.

- Comparable effectiveness compared with in-person simulation, with limitations discussed as connectivity issues, learning curves, and a limited set of scenarios can be effectively run by telesimulation.
- Lack of studies studying specifically emotional responses

Chapter 3. Literature Review

The third chapter presents the main concepts related to healthcare simulation. More specifically, it presents a brief review of healthcare simulation, the concept of affective computing, the role of emotions and emotional responses in HCI, and its importance to healthcare simulation. Then, we provide the context of COVID-19 and its impact on medical education, healthcare simulation, and the emergence of telesimulation as an alternative to clinical education when the simulation centers were closed.

Finally, we present quasi-experimental and natural experiments as a robust alternative to studying the phenomena, performing a literature review demonstrating the core concepts, applications, and limits. Finally, we will discuss the applications of non-inferiority trials and experiments used in this research to test the hypothesis that telesimulation is not an inferior approach compared to in-person simulation.

Although some key concepts were introduced during the scoping review and the bibliometric analysis, in this chapter, we discuss in greater detail the main concepts and their intertwining according to the key references in the respective fields.

3.1. Simulation

A simulation can be defined as a strategy or technique used to generate an experience without going through the actual event. Simulation is a situation in which a specific set of conditions is artificially created to provide the experience of a constructed reality. Simulation is as an artificial representation of a real-world process to accomplish educational goals through experimental learning (Flanagan et al., 2004). Simulation can be applied in different sectors to achieve diverse objectives (So et al., 2019).

Simulation enhances the active learning process and assists in achieving, exploring, communicating, and applying new information and experience to achieve effective results (Anderson & Lawton, 2004). Furthermore, numerous authors declared that instructive electronic web-based or software-based simulations compared to conventional educational apparatuses could improve learning, attitude, and behavior (Rutten et al.,2012). Dede (2007) stated that simulation could accelerate learning capacity

and empowers faster learner comprehension of complex abilities than any other learning methods permit.

Simulation assists students in learning and progressing administration capacity and improves the quality of the corporate decisions they make. As opined by Arias (2007), Simulations address the need for openings to acquire fundamental problem-solving skills by actively involving students in authentic problems and understanding aptitudes. The simulation also helps students in gaining practical skills. One of the foremost vital steps in the educational model improvement is the introduction of simulation-based medical education and learning. Ericsson (2015) opined that medical simulation allows for the purposeful practice of clinical skills rather than the apprentice learning method.

Romme (2002) specified that simulation helps students apply the learned concepts during classes or readings before engaging in a real-world, clinical setting. A developing body of evidence appears that clinical skills obtained in the medical simulation laboratory setting improve patient care practices and delivers patient outcomes (McGaghie et al., 2011). Lateef (2010) asserts that simulation-based medical education can be a platform for learning to moderate moral pressures and resolve practical concerns. Consequently, the simulated environment permits learning to rectify mistakes, permitting the students to culminate steps and fine-tune abilities to optimize clinical results.

The aerospace and aviation sectors have used simulation as a teaching resource for a long time. Simulation is a well-known technique for training personnel in aviation, the military, and industry, and it is strongly adopted in the healthcare sector.

Simulators are widely utilized to teach and train high-risk professions and sectors such as the military, commercial aviation, atomic energy, business, and medical (Issenberg et al., 2001).

Simulation is an adaptive teaching technique that may build various skills and knowledge. Simulation can be utilized in diversified fields within the domain of health sciences:

Clinical Simulation: Simulation used to research, evaluate, and recreate specific conditions, disorders, and biological processes in the human body.

Operational Simulation: Simulation used to record, evaluate, and research healthcare operations, provision of services, scheduling of healthcare business procedures, and patient flow.

Managerial Simulation: Simulation utilized in management, decision-making, implementation of policy, and strategy development.

Educational Simulation: Virtual environment and physical objects are broadly applied to supplement and improve simulation experiments used for training and educational reasons.

Healthcare Simulation: Healthcare simulation is an enthralling topic that blends advanced technology, clinical healthcare procedures, and up-to-date adult education.

3.1.1. Healthcare Simulation

Healthcare simulation is a modern method of educating and training healthcare personnel in cognitive, technical, and behavioral skills through technologically enhanced created experiences (Green et al., 2014). Simulation in healthcare can be operationalized as an instrument, device, and environment that simulates an element of clinical treatment (Lame & Dixon, 2020). According to Baily (2019), simulation training diminishes expenses, enhances performance, and lowers the likelihood of medical mistakes in the healthcare industry.

Healthcare simulation may replace real patient interaction or other clinical settings during the learning process. Some studies found that healthcare simulation has been 'superior to traditional clinical teaching, generating intense educational interventions that generate rapid and enduring effects in various disciplines (Hu et al., 2018). The advantages of simulation in medical and nursing education are widely acknowledged in the literature. Patient safety is one of the most compelling justifications for utilizing simulation in health professional education. Students can practice clinical skills without endangering actual patients by using simulation.

Schiavenato (2009) stated that simulation's enhanced experience might reduce the number of errors and malpractice. Additionally, simulation, as a method, provides a controlled environment. When students participate in the simulation, they are all exposed to the same situation in the same setting and assessed by faculty members.

Another important aspect, according to Bradley (2006), is that learners receive almost instantaneous feedback on their performance during simulations. If the simulation is video recorded, they have the chance to self-reflect and evaluate their performance during debriefing. The practical exposure of the learners to a range of complex situations

allows students to become more confident and open during their performance. A general perspective is that solid communication skills are a prerequisite when dealing with patients and handling critical situations confidently.

Hecimovich and Volet (2011) defend that simulation training provides practice, undoubtedly boosting confidence and contentment since learners will feel better prepared to face real-life circumstances. Confidence is vital for individuals but also critical for patients who have faith in healthcare practitioners.

Schiavenato (2009) indicates that digital simulation technology has opened new avenues for communicating medical information in incredibly realistic clinical settings through diverse methods. Honey and Hilton (2011) identified that these simulations allow students to see and engage with depicting natural events that would otherwise be difficult to witness. William et al. (2016) stated that the simulation experience was seen as a beneficial learning experience and that this is connected to improved academic achievement.

Al-Elq clarifies that any educational activity that employs simulation devices to imitate clinical circumstances could be classified as simulation-based medical education. According to Lopreiato and Sawyer (2015), simulation devices serve as an alternative to genuine patients. There are distinctive sorts and classifications of simulators, and their cost varies, concurring to the degree of their likeness to reality or fidelity. Although simulation-based learning is costly, however, is cost-effective in case utilized appropriately (Al-Elq, 2010).

Several simulation devices include high-tech virtual reality simulators, full-scale mannequins, standardized patients, plastic models, human cadavers, or screen-based simulators (Sorenson et al., 2017).

Traditionally, simulation training in healthcare education has been associated with high-fidelity patient simulators. Several parameters can be inputted into these high-fidelity simulators to replicate human reactions and physiology (heart and breath sounds, blood pressure, and sweating) and respond to physiologic and pharmacologic intervention (Okuda et al., 2009).

The mannequins may be disease-specific (a cardiac simulator) or capable of a wide range of disease states depending on the model, such as METI man (Adult male patient simulator). High-fidelity simulators for children and infants are also available such as ‘babysit’ (Decker et al., 2008). A task trainer is another specialized model to assist

students in practicing specific skills and lacks the comprehensive programming possibilities of high-fidelity models (Okuda et al., 2009).

Depending on the complexity of the model, these trainers are known as low-fidelity or moderate-fidelity simulators. According to Bradley (2006), the Accu-Touch endoscopic surgical simulator, for example, is a medium-fidelity simulator that allows students to practice inserting an endoscope into the upper and lower gastrointestinal systems.

Additionally, health sciences have a history of employing standard patients as one of the most prevalent kinds of simulation. Okuda et al. (2009) specified that Standardized patients are real people trained to depict patients by sharing a life history, projecting a personality, and displaying emotions in pre-determined moments according to a pre-defined script to attain a specific educational objective.

This type of simulation generally focuses on developing students' communication and intervention abilities because specialized individuals are trained to represent patient scenarios to instruct and evaluate the clinical skillset of medical students and other diversified professionals (Decker et al., 2008).

Screen-based simulators, often used for self-learning, serve an essential role in fundamental and clinical medicine, such as medical interviews or differential diagnosis (Akaike et al., 2012). Another common type of simulation that has been gaining traction in the last years with the maturation of the technology is virtual simulation. The virtual patient and the surroundings are developed and performed on the computer.

According to Bradley (2006) virtual reality is one of the more approachable kinds of simulation since virtual reality software runs on commodity, non-specialized computers.

Another widely employed simulation is the In-Situ simulation, which takes place in a natural clinical setting with on-duty healthcare practitioners throughout their typical workday (Patterson et al.,2008). In addition, in-situ simulation enables participants to evaluate and consolidate their abilities and face challenges in the clinical setting (Kurup et al., 2017).

Lasater (2007) specified that using high-fidelity mannequins, healthcare personnel, standardized patients, and associated equipment to reproduce a real-life clinical setting is known as full environment simulation.

Simulation-based education is becoming popular. However, it still encounters certain limitations in terms of acceptability and efficacy. Elliot, Coovert, and Schiflett

(2017) highlight that the capacity to deliver an experience by engrossing and captivating individuals in an artificial environment that catches their attention and exposes them to crucial contextual features related to their performance is the most significant task of simulation-based education.

Another potential limitation of simulation is that the objective is concentrated on specific capabilities in some instances. Students can enhance their blood pressure measuring skills using a simulation exercise utilizing a blood pressure task trainer, but it does not improve their communication or evaluation ability (Wenk et al., 2009).

One additional disadvantage of employing simulation in the education of health professionals is the high cost of modern equipment. According to (Rickless et al., 2009), simulation is a costly instructional method. The high cost of simulation is frequently connected with high- and medium-fidelity human patient simulators (HPS), which have very high initial purchase costs (Lin et al., 2011).

As demonstrated in our bibliometric review, the availability and usage of this approach tend to be mostly limited to high-income countries with the overwhelming dominance of the US, Canada, and the UK by most metrics in the production of knowledge in the field. However, collaborative arrangements and telesimulation can be alternatives to bring the opportunity of healthcare simulation to a broader spectrum of users and locations.

Realism and Fidelity

One of the few crucial concepts in healthcare simulation is realism and fidelity. Realism is depicting things or situations in a realistic or true-to-life manner. One of the final objectives of simulation is to make the experience as realistic as possible. Cioffi (2001) explains the concept of fidelity-realism as a realistic and process-based clinical scenario with ensured validity (Robinson & Dearmon, 2013).

According to Rashid and Gianduzzo (2016), fidelity is the level of precision attained. In addition, simulation fidelity is an umbrella term defined as the extent to which the simulation replicates the actual environment (Lewis et al., 2012).

In healthcare education, role-playing is often employed (Paskins & Peile, 2010), and the performance aspect of role-playing is considered one of the sources of anxiety among students. However, some factors might limit the effectiveness of eliciting emotions and learning objectives during this process.

Nestel and Tierney (2007) teach that a lack of realism might cause learners to perceive the role play as useless. The role plays must be realistic and challenging and help reduce anxiety, and clear instructions regarding the role must be provided. In addition, proper guidelines and instructions regarding the task may be beneficial for the learners in reducing anxiety and increasing their sense of medical competence.

However, Jefries (2005) argued that a clinical simulation must be authentic and incorporate as many realistic environmental aspects as possible to make the students' learning more relevant to their formative path.

In the same way, Adamson (2015) stressed that realism is more required than authenticity through simulation design. The high degree of fidelity might increase the cognitive workload where participants start feeling overwhelmed and experience reduced learning (Norman et al., 2012). According to Maran and Glavin (2003), existing information and stimuli are not required for mandated learning. Extra information, stimuli, and difficulty using an interface are called cognitive load.

In terms of types of fidelity, there are several classifications and ontologies to describe fidelity, especially in the healthcare simulation field. However, for practical purposes, they can be classified as:

- **Conceptual fidelity**: Low fidelity simulations are commonly used to investigate different concepts during the conceptual design phase. Healthcare simulation guarantees that all components of the situation practically relate to each other to make sense to the learner. To maximize conceptual fidelity, subject matter experts often review scenarios (Rudolph et al., (2007).
- **Physical fidelity**: Physical fidelity is the simulator's ability to replicate the feel and appearance of the natural system or object being modeled (Alexander et al., 2005)
- **Emotional or psychological fidelity**: Emotional or psychological fidelity is the degree to which a simulation can duplicate or capture a real task using a simulated task and make the student feel as if it were real (Munshi et al., 2016). This concept is close to the idea of immersion, often utilized in virtual reality and related fields of study.

Lewis et al. (2012) concluded that high fidelity simulation could give participants a secure and regulated learning environment to acquire non-technical abilities, allowing them to make mistakes, correct them as they participate in the session, and learn mistakes

without affecting patient safety. They define high fidelity simulation as a simulation that includes a computerized full-body manikin that may be programmed to generate realistic physiological responses to student actions. The continuum of fidelity goes from part-task trainer to full-scale simulation.

Nonetheless, a degree of fidelity is not necessarily always better. A higher degree of fidelity might increase the cognitive load to the point where participants start feeling overwhelmed and experience reduced learning (Norman et al., 2012). Sometimes, a part-task trainer is more appropriate to develop specific skills, such as surgery.

Table 23: Simulation Tools and associated fidelity

Simulation Tools (In order of Increasing Fidelity)	Description of Simulation Activity
T*ask* trainers	Models or manikins used to learn, practice & gain competence in simple procedures, usually comprised by a specific body part, instead of a full-body mannikin
Screen-based computerized simulators	Programs used to acquire knowledge, to assess the competency of knowledge attainment and to provide feedback related to clinical knowledge and critical thinking skills
Virtual reality (VR)	Combines a computer-generated environment that can include tactile. auditory and visual stimuli
'Haptic' systems	A simulator that combines real-world and virtual reality
Standardized patients	Use case studies and employ role-playing in which individuals behave as a patient in a realistic and consistent manner
Full-scale simulation	A simulation that incorporates a computerized full-body manikin that can be programmed to provide an actual physiological response to student actions

Source: Lewis et al. (2012)

We present a list of images illustrating the different simulation tools in the sequence. Since they are widely known, we will not screen-based computerized simulators and standardized patients. A list of figures for the remaining simulation tools is available in the sequence:

Figure 24: Task Trainer (Part Task)



Source: ASTEC website

Figure 25:VR Simulation



Source: ASTEC website

Figure 26: Example of a haptic feedback system



Source: The Verge – available at https://www.theverge.com/2018/8/14/17670304/virtual-reality-surgery-training-haptic-feedback-fundamentalvr

Figure 27: Example of Full-scale simulation in a simulation center



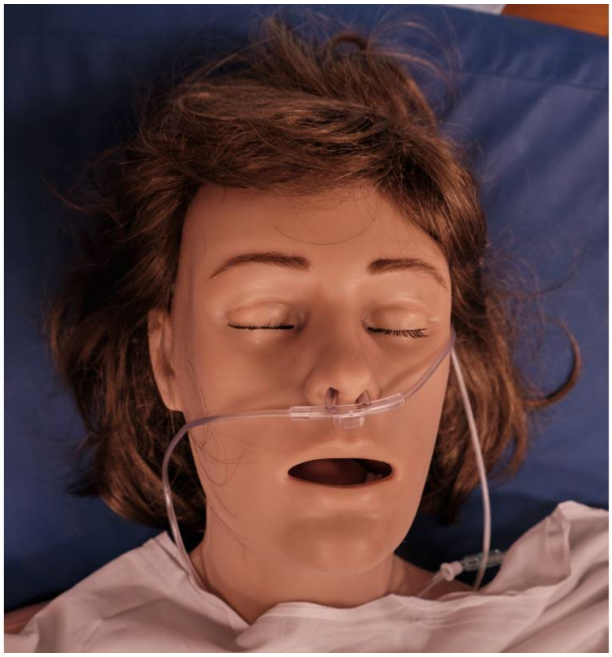
Source: ASTEC website

Figure 28: Example of full-scale simulator – newborn - in a simulation center



Source: ASTEC website

Figure 29: Example of Full-scale simulator – female mannikin - in a simulation center



Source: ASTEC website

Ferber (2002) states that a low fidelity simulation will be the least realistic. Screen-based text simulators are often simple to build and need modest computational resources, making them popular in the early stages of healthcare simulation (Doyle, 2002). Although these simulators allow learners to practice independently, they focus on specific abilities and provide little immersion due to a lack of graphical components (Datta et al., 2012).

Screen-based graphical simulators give a more accurate depiction, are portable, and are less expensive than simpler text-based interfaces. However, Kneebone (2003) teaches that these simulations could assist in understanding the intellectual framework behind the clinical practice; they rarely provide genuine practical skills. They also impede the learner from fully immersing themselves in the situation, necessitating active 'stage management of the scenario in reaction to interventions (Datta et al., 2010).

Finally, high-fidelity simulation provides the most realistic and interactive environment for learners in a setting that strongly matches reality. According to Blackstock and Jull (2007), high-fidelity simulations are computerized full-body manikins that mimic the physiology and anatomy of an actual patient. Moreover, Murray (2014) stated that many manikins have the capacity for conversation, which permits students to develop communication and problem-solving abilities. In addition, high-fidelity manikins can also perform pre-programmed scenarios (Finan et al., 2012).

3.1.2. Debriefing

Debriefing is a reflective practice that allows action-based reflection as part of the continuous learning experience (Schon, 1984). It is a planned conversation that takes place after a simulation exposure to assist participants in understanding more clearly their actions and mental processes to improve future learning outcomes and clinical performance (Sawyer et al., 2016).

During debriefing, instructors and learners reappraise the clinical interaction promoting clinical reasoning and judgment skills development through thoughtful learning processes, an essential part of healthcare education (Sawyer et al., 2016) in which learners have opportunities to link classroom content to clinical practice through experiential learning.

As a result, facilitating reflection via debriefing is critical for ensuring that students get the most out of the simulation. Apart from this, Dreifuerst (2009) stated that

a concept analysis of the debriefing process in simulated learning is critical for improving consistent, substantial student learning and facilitating faculty development.

According to Rall et al. (2000), Debriefing is "the heart and soul of the simulation experience.". Debriefing is considered a critical component of simulator training (Cantrell, 2008) and is often described as simulation's most important educational component (Barry et al., 2005).

According to Kalaila (2014), a simulated scenario or case helps learners build confidence but can also cause anxiety and stress. Therefore, the debriefing facilitator should create a welcoming environment to reduce anxiety and ensure a positive debriefing and learning experience. A successful debriefing process and learning experience are enabled by an environment where participants can share their experiences in an open, candid, and fair way to express their feelings with ease and freedom.

Debriefing serves various purposes and procedures, including assisting participants in comprehending, evaluating, and synthesizing their thoughts, feeling, and actions during the simulation to upgrade future performance in similar scenarios (Rudolph et al., 2014).

Debriefing participants should not be judged too harshly nor not being evaluated in terms of their actual performance, presenting criticism politely and sensitively. When sharing feedback during debriefing, instructors should make direct and honest comments about the students' point of view courteously and interestedly (Rudolph et al., 2008). Besides, debriefing may include altering one's state of mind, recognition, actions, and specialized abilities (Gardner, 2013)

Additionally, Kuiper and Pesut (2004) stated that debriefing is associated with critical thinking, clinical reasoning, and clinical judgment, desirable characteristics as learners advance from novice to expert practice. Debriefing guides students in identifying and exploring gaps and bridging them to improve performance (Gardner, 2013).

Debriefing offers a way to promote student thinking and help them develop their complex decision-making skills by providing opportunities to review events and make their significance visible (Dreifuerst, 2015). Participants in a simulated or practical life may experience various emotions due to their participation.

Clapper (2010) demonstrated that emotions improve learning by framing the experience positively. Emotions significantly impact a learner's ability to retain and activate information; however, the case of detracting from engagement in the experience can inhibit learning (Gardner, 2013). Students reflect on their simulated experience, have

an emotional release period, receive behavioral feedback, and integrate their observations, behavior, and feedback into a conceptual framework (Dreifuerst, 2015).

3.1.3. COVID-19 Impact on the Healthcare Simulation

The adverse effects of COVID-19 have been observed seriously in the various domains of business life, and the educational system is one of those poorly affected fields (Armitage & Nellums, 2020). As a result, numerous schools, universities, and educational institutions had to discontinue face-to-face instructional methods and delivery systems. Moreover, according to Kaur et al., (2020), the educational sectors were having a tough time dealing with the current environment, notably medical education, which was more difficult.

According to Dhawan (2002), the current COVID 19 widespread pushed universities through a profound alteration of instruction and assessment strategies by relying only on online education approaches. Moreover, this was a massive challenge for students and teachers (Houlden & Veletsianos, 2020). Medical universities have minimal e-learning facilities, and e-Learning platforms are restricted to a few fields (Cosnita et al.,2020).

The unexpected COVID 19 epidemic allowed us to investigate virtual learning environments and inspired us to develop a hybrid curriculum (Ali, 2020). Furthermore, Chin et al. (2021) stated that coordination, technical assistance, and a solid readiness to adopt various approaches were necessary to implement a hybrid curriculum.

E-learning systems enabled students to complete their qualifications during the pandemic. Although to avoid disruption in the educational process, the institutes shifted from traditional (face-to-face) to online learning management systems; this rapid shift caused several issues such as learners' engagement, satisfaction and lack of interactive teaching methodologies (Pather et al., 2020).

Nevertheless, most faculty members struggle to engage students and opt for different approaches to deliver online course materials and assessments (Mukhtar et al., 2020). The primary concern for educational institutes is to address the various challenges of online learning management systems and platforms, such as Microsoft Teams, Zoom, Canvas, Go to Meeting, and Google Classroom (Fichten et al., 2021).

Moreover, Evans (2020) argued that learning material delivery based on the internet must be customized and aligned with learning management systems to ensure

learners' effective and impactful learning outcomes. Therefore, instructors must be familiar with the different teaching methodologies to achieve effective educational outcomes.

Although several benefits have been observed regarding online educational delivery systems, certain limitations have become a challenge for learners, instructors, and facilitators (Bao et al., 2019). Out of the various challenges, most students during the COVID-19 pandemic and lockdown had to experience various levels of anxiety (Sundarasen et al., 2020);

Many researchers shared their concern about a lack of practical knowledge and expertise required for success in residency and subsequent clinical practices, specifically in healthcare education contexts (Wolff et al., 2010). Coffey et al. (2020) highlighted the obstacles to assisting students in developing clinical practice abilities generated from in-person interactions and physical exams, which are not easily transferable to remote formats such as classroom-based teaching.

Tele Simulation as a Strategy to Respond to COVID-19

Given the social distance restrictions imposed to avoid the spread of coronavirus during the COVID-19 and lockdown period, various face-to-face simulation programs have been discontinued by educational institutes (Chick et al., 2020). Since the pandemic continued for a long time, students, educational institutes, and instructors were unable to complete course work and their qualifications in time, forcing them to seek alternative options to deal with this challenge

Therefore, the concept of telesimulation has gained traction, which involves utilizing the internet to connect simulators between an instructor and a learner in distant places (Roach & Okrainec 2021). Telesimulation provides education, training, and evaluation to learners from a remote location using telecommunication and simulation resources (McCoy et al., 2017) since in-person education has been severely limited due to the Covid-19 pandemic.

Therefore, there was a growing demand for telesimulation to complement trainee teaching (Dedeilia et al., 2020). However, remote facilitation allows teams to practice together while the technology is administered remotely, which is helpful given the concentration of skilled facilitators at large, often academic organizations. McCoy et al. (2017) stated that this technology had been utilized to virtually teach procedural, teamwork, and restorative skills and evaluate students across several disciplines.

Telesimulation can be used when learners and facilitators are in different physical locations. Facilitators can exchange voice snippets, still photos, and pre-recorded films using video teleconferencing platforms like Zoom, Microsoft Teams, and Skype. According to Sanseau et al. (2021), utilizing a remote simulation platform, "Tele Sim Box," by faculty is acceptable. It was determined to be a successful approach to medical student education.

Telesimulation technology leverages telecommunication and simulation resources to provide training, education, and evaluation to individuals in distant places (McCoy (2017). Moreover, remote educational addresses the unequal distribution of medical education resources, which can substantially influence medical students' learning and access from remote locations (Christensen et al., 2015).

Diaz and Walsh (2021) declared telesimulation as a new and innovative approach that permits the expansion of distance education to many healthcare trainees and professionals across developing and low-resource areas. The COVID-19 pandemic and the necessity for social distance have transformed conventional simulation-based medical education into virtual telesimulation systems (Diaz & Walsh, 2021).

To get the full benefits of modern telesimulation options, the learners and instructors must distinguish between telesimulation and traditional simulation. The simulation has different goals, technology needs, and events that must be recognized and performed effectively to optimize learning potential (Hooper & Rieber, 1995).

Nevertheless, medical simulation is a successful educational method for addressing content and process-specific learning objectives, such as communication, problem-solving and interpersonal abilities. Interestingly, simulation is witnessing its paradigm change (Papanagnou, 2017). Telesimulation sessions facilitate the learners in educational delivery who are remotely placed and have time constraints (McCoy, 2017) since the educational sessions can be arranged off-site as per both learners and the instructors (Almousa et al., 2021).

According to several research studies, simulation-based medical training conducted remotely (telesimulation) can build the healthcare skills required in an emergency. Additionally, international simulation-based training could provide solid learning to many individuals engaged in delivering emergency services across the globe (Ypinazar & Margolis, 2006).

The potential benefits of telesimulation can be extended to remote regions, wherever learners are (McCoy et al., 2017). Telesimulation is incredibly beneficial when

the issues of distance and difficulty of access. Time pressures are eliminated, and travel time and cost are reduced with the same level of effectiveness and efficiency. This argument is supported by Sanseau et al. (2021) that the education and learning delivered through telesimulation are as effective as training conducted in a simulation center, given the sufficient resources.

Additional research indicates that most learners and instructors expressed high satisfaction with a telesimulation intervention, including all medical students, staff, and facilitators engaged remotely for pediatric emergency instruction during the COVID-19 pandemic (Wang et al., 2021).

During COVID-19, almost all the educational institutes were closed, and social distancing was imposed around the globe. The participants appreciated the learning experience through telesimulation has been appreciated by participants as a pleasant learning experience (Heffernan, 2021).

The students faced serious issues regarding their learning and qualifications, especially from health-related educational institutes, medical colleges, and universities. Consequently, face-to-face educational delivery systems have been largely shifted to virtual learning environments, especially during the most critical moments. According to a different study, most learners and instructors expressed high satisfaction with telesimulation intervention. All medical students, staff, and facilitators engaged remotely in pediatric emergency instruction during the COVID-19 pandemic (Wang et al., 2021).

Although several benefits have been discussed and can be capitalized through deploying telesimulation-based educational delivery, at the same time, a couple of challenges are yet to be addressed by both learners and instructors. For instance, the facilitators raised concerns about their inability to accurately analyze participants' emotions and facial expressions (Ikeyama et al., 2012). Moreover, a growing amount of research indicates that delayed video review has consistent interrater reliability (Thompson et al., 2012).

The actual hands-on interaction component may be diminished with telesimulation stressing the necessity to adjust goals and objectives; while telesimulation is still in its early stages of development, best practices are being developed (Corter et al., 2011). Roach and Okrainec (2021) discuss the lack of hands-on skill training, and participants will not have the chance to engage in the purposeful practice.

Additionally, the participants' lack of telesimulation expertise might be a stumbling block to the growth and accomplishment of the objectives as well as a period

must be created that allows the students to function without being hampered by a time limit or that allows the learners to avoid being disinterested (Gutierrez et al., 2021).

On the other hand, telesimulation may be used to suitably challenge learners at higher levels of competence when they are less reliant on rules and norms and can use analytical ways in novel circumstances or simulations (Papanagnou, 2017).

Another challenge of telesimulation is debriefing which is frequently seen as the most critical component of a simulation, as it is at this period that most of the learning occurs (Renouf et al., 2017). Teledebriefing includes a faculty member at a distant location providing feedback while observing the simulation in real-time using video conferencing technology (Ahmed et al., 2016). Tele debriefing in simulation extends simulation advantages outside the borders of simulation facilities.

Furthermore, McCoy (2017) argued that telesimulation, which uses cameras, microphones, and video conferencing software, allows students to communicate with instructors in a different location. Understanding the experiences and perspectives of participants involved in the communication process is critical to recognize impediments in telesimulation (Isman & Dabai, 2004). Therefore, Heffernan et al. (2021) suggested proper investigation of facilitators' attitudes and learners' perceptions.

3.2. Emotions

The reactions to the environment or humans' changes in the environment are called significant emotions. According to the American Psychological Association (APA, 2021), emotions are "a complex reaction pattern incorporating experience, physiological, and behavioral aspects." Dixon (2012) shares that the concept of emotions has a long and complex history in philosophy and psychology. Historically, it has been used as a meta-concept that combines numerous words representing feelings and attitudes.

Emotions have been described even by Aristotle (384-322 BCE) as "all those sentiments which affect judgments." However, Ekman et al. (1969) employed "The Big Six" model in their study on universal emotion identification from facial expression; happiness, fear, sadness, anger, disgust, and surprise were on the list, and they remain the most often recognized choices for basic emotions.

The term "emotion" has a variety of definitions, many of which are vague. Clore et al. (1987) examined the validity of a taxonomy of affective ideas to identify emotions

from comparable terms in the affective lexicon, which the authors define as "a subset of words connected to affect or affective situations."

Brave and Nass (2007) provided a competent introduction to emotions in the human-computer interaction domain. They defined emotions more precisely and clarified their meaning by differentiating them from two similar concepts: moods and sentiments.

According to Brave and Nass (2007), emotions indicate interactions with a specified object, while mood does not. We feel the emotion about something or someone, while a mood does not have the same intentionality or directionality towards a specific object, being more diffuse and generic. Exemplifying, when someone feels sad about something, it is a manifested emotion, while being in a depressed state represents a mood.

The other key difference presented is functional. Emotions are a direct reaction to a situation and evoke and influence an immediate action; therefore, emotions have a short time frame. On the other hand, moods are longer-term mental states that influence the cognitive process instead (Brave and Nass, 2007). Another term that is often confused with emotion is sentiments.

Unlike emotions (and moods), sentiments are attributes of an object, not states of an individual. When people say they "like" or "dislike" one object, this evaluation reveals that their association with that object is a specific emotional state. When engaging with that object, they expect to feel positive or negative emotions (Brave and Nass, 2007). So, when someone likes an object, they expect to feel positive emotions, but the sentiment is not the emotion itself.

Another similar term, “affect” has different connotations regarding the evaluative aspect of attitudes and the tripartite model composed of cognitive, affective, and connotative/behavioral components (Cohen, Pham, and Andrade, 2008).

Cohen et al. (2008) explain that an explicit or implicit “liking” for some object, person, or position is considered an evaluative judgment rather than an internal feeling state. The essential characteristic of this definition of affect is related to induced feeling states instead of a liking or a descriptive cognition. In summary, although emotions and moods are transient—emotions last seconds and moods might last hours or even days—sentiments can stay permanently and are thus responsible for the intentional decision to pursue or avoid specific objects, interactions, and circumstances.

Despite the multiple definitions, Brave and Nass (2007) summarize the multiple definitions by describing two features of emotions that most definitions share: 1) Emotion is a physical, affective, behavioral, and cognitive response to events that are significant

to an individual's wants, goals, or worries; and 2) emotion includes physiological, affective, behavioral, and cognitive components.

They illustrate the interplay between physiological, affective, behavioral, and cognitive components with two emotions: fear and joy. They defend that fear is a potent negative emotional state that evokes physiological and cognitive preparedness for action in response to a circumstance that seems threatening. On the other hand, joy is a positive, approach-oriented state that results from achieving goals.

Another popular definition of emotion is a "valenced affective reaction to perceptions of situations" (Richins, 1997, p.127), since it distinguishes emotion from other types of cognitions a component related to the valence (positiveness or negativeness), such as interest and surprise, or purely physical states such as sleepy and droopy. That concept of emotion excludes general and subjective personal judgments such as feeling self-assured or abandoned (Richins, 1997).

Even though most of us intuitively understand emotion, we lack a concrete definition of the phenomenon (Israel, 2020). While we do not rely on a deep understanding of emotions in our daily social interactions, a better understanding of effective processes and mechanisms is crucial to many research fields.

Moreover, Ekman (1999) claimed that emotions are transient, regularly enduring from seconds to minutes at most. Several dimensional models of emotions have developed, but only a few, including the circumflex model, the vector model, and the PANA (Positive-Negative Activation) model, remain the most widely accepted models today (Rubin & Talarico, 2009).

Circumplex model

The circumplex model of affect has become one of the most well-known and widely studied of the various models that have emerged over the years (Watson & Tellegen, 1999). James Russell created the circumplex model of emotion (Russell, 2003). According to this model, emotions are distributed in a two-dimensional circular space with arousal and valence dimensions.

Rubin and Talarico (2009) explain that the vertical axis represents arousal while the horizontal axis represents valence, with the circle's center representing neutral valence and a medium level of arousal. Emotional states can be defined at any valence and arousal level, including neutral levels of these factors. Furthermore, circumplex models have

been utilized to test boosts of feeling words, passionate facial expressions, and emotional states (Remington et al., 2000).

Moreover, Russell et al. (1999) stated that distinctive prototypical enthusiastic episodes, clear emotions evoked or coordinated by specific objects, can be plotted on the circumplex according to their levels of arousal and pleasantness.

Vector model

Another popular form of modeling emotions is the Vector Model, introduced in 1992 (Bradley et al., 1992). The two-dimensional model comprises vectors that point in two directions. The model has one dimension that defines the arousal (first vector), while the other component defines the direction to specify the valence(second vector). The model expects continuously underlying arousal measurement in which valence demonstrates the heading in which a specific emotion moves.

This model represents emotions with a broader combination of arousal and valence, such as combining high arousal but a neutral valence, assuming that emotions can be intense without being very positive or negative (neutral valence). High arousal states are distinguished by their valence in this model, whereas low arousal states are far more impartial and depicted near the intersection of the vectors. Vector models have been prevalent in studies testing word and picture stimuli (Remington et al., 2000).

Positive Activation – Negative Activation (PANA) model

Watson and Tellegen first proposed the positive activation – negative activation (PANA) or "Consensual" model of emotion in 1985, and they recommended that positive influence and negative influence are two isolated frameworks (Watson & Tellegen, 1985). Additionally, higher arousal states are defined by valence, similar to the vector model, while lower arousal states are more neutral in valence (Rubin & Talarico, 2009).

The vertical axis demonstrates low to high positive affect, while the horizontal axis reflects low to high negative affect in the PANA mode. In addition to this, the dimensions of arousal & valence were rotated 45 degrees around these axes (Watson & Tellegen, 1985).

PAD Emotional State Model

Albert Mehrabian and James A. Russell created the Pad emotional state model to depict emotional states (Mehrabian, 1980). Pad employments three numerical measurements to represent all emotions, including Pleasure, Arousal, and Dominance (Bales, 2017). The Pleasure-Displeasure Scale assesses the pleasantness of an emotion. Anger & fear, for example, are both unpleasant emotions that rank high on the displeasure scale, whereas joy is a pleasant emotion Russell & Mehrabian, 1977).

The Arousal-Non-Arousal scale measures how energized or calmed an individual feels. Arousal is a measure of the intensity of emotion. Low arousal intense feelings such as grief and depression are low-arousal intense feelings. In contrast, anger and rage are negative emotions, but rage is considered more intense and has a higher arousal level.

On the other hand, Boredom is an unpleasant state with a low arousal value (Mehrabian, 1980). The dominance- Submissiveness scale represents the emotion's controlling and dominant nature. According to Mehrabian and Bernath (1991), fear and anger are unpleasant emotions; anger is the dominant emotion, whereas fear is the submissive emotion.

3.2.1. Arousal as a Component of Emotions

Arousal is a state generated from certain emotions, whether positive or negative. Russel also demarcated arousal as an element of deeply felt feelings. Arousal can also cause actions generated in response to sensory stimuli (Kandel et al., 2000). The level of arousal has various degrees, from unconsciousness and sleep to alertness and activation.

Hence, Lang and Devis (2006) interpreted arousal as a significant reason for any activity. According to the James-Lange theory of emotion, our emotional experience depends on our arousal levels. According to Cannon (1927) distinct forms of arousal may produce different emotional experiences. Braidley et al. (2001) clearly stated that arousal is the level of excitement and energy resulting from emotional inputs. However, Bradley construed that arousal and emotion are not mutually exclusive but rely on arousal.

On the other hand, Walters (2020) stated that our emotional experiences are weaker without arousal. Hohmann (1966) mentions that patients with spinal injuries that limit their arousal experience also report lower emotional reactions. Emotions carry arousal and cognitive elements, which are equally crucial for a specific reason.

Jones (2003) shares that our emotional experiences are affected by various arousal levels. Jones also argued that our cognitive assessment regarding events decides our negative or positive emotions.

Valence and arousal are the two terms primarily used to describe emotional experiences. Schnall (2010) defined valence as whether an event is positive or negative and arousal as whether an individual feels a given emotion as activating or deactivating. Emotions are explored and classified as positive emotions, including hope, happiness, joy, and satisfaction, whereas negative emotions can be anxiety, shame, sadness, and disgust (Pekrun & Linnenbrink, 2014).

According to Charland (2005), valence is one of the most fundamental scientific ideas at the center of emotional experience. Lewin (1951) coined 'valence' to explain the forces that attract individuals to cut things and repel them from undesirable ones. Since then, the idea has been greatly expanded, including, but not limited to, the classification of emotions as positively or negatively valence (Colombetti, 2005).

There is widespread agreement that valence, represented by terms such as positive-negative, good-bad, or pleasure-displeasure, expresses something fundamental about emotion (Frijda, 2009).

Research from numerous fields concerned with emotions, motivation, learning, and decision-making uses valence, pleasantness, utility, or liking/wanting to distinguish between the positive and negative.

3.2.2. Emotional Appraisal Process

According to Holmstorm (2015), emotional appraisal refers to processes by which individuals' cognitions about events predict their emotional reactions to those events. Emotions are generated based on judgments and appraisals regarding situations, and Arnold (1960) was among the first to define that humans assess or appraise the significance of their surroundings. These evaluations, in turn, drive their behavior.

In simple words, an appraisal is a method for determining and evaluating the influence of the environment on one's wellbeing (Frijda, 2000). The accomplishment of the goals or obstructing concerns is the best way to consider well-being's significance. However, concerns include the individual's desires, attachments, values, current goals, and beliefs (Sherer et al., 2007).

An appraisal dimension is any attributes that an individual assesses during an interaction with the environment, which generates an appropriate emotion as a driver of emotional responses (Watson & Spence, 2007). Appraisal dimensions include an event's goal relevance or congruence, novelty, perceived pleasantness or unpleasantness, fairness, and control.

In general, pleasantness relates to an individual assessment of whether a situation is good or bad (favorable or unfavorable) regarding personal consequences (Watson & Spence, 2007). Pleasantness is a fundamental appraisal dimension and accounts for most variance explained in attempts to categorize emotions (Ruth et al., 2002). The development of positive or negative emotions depends on a scenario's classifications, whether pleasant or unpleasant.

According to Sander et al. (2005), a positive appraisal inspires people to embrace and proceed with the task, whereas negativity causes alienation from the work. Moreover, an assessment of goals congruence is also known as motivation consistency. Goals congruence influences the valence of the emotional responses and distinguishes between good and negative emotions (Roseman et al., 1990).

Furthermore, Hosany (2012) stated that a person's evaluation of goal congruence, also recognized as motivation steadiness, attractiveness, or goal relevance, leads to how favorable or unfavorable a scenario is to achieve the goal. According to Roseman and Smith (2001) the more goal-relevant a scenario is, the stronger the emotion evoked. Stein and Levin (2013) stated that pleasure denotes accomplishing a goal, whereas sadness denotes failure to reach or sustain a goal. The third fundamental appraisal component is certainty, which concerns estimating the perceived likelihood of a particular occurrence and its impact on emotional responses (Roseman, 1984).

The certainty dimension distinguishes between emotions with known outcomes (pleasure and fury) and anticipated emotions (hope and surprise) with unknown results (Macinnis & De, 2005). The experience or knowledge gained from earlier exposure provides certainty about a situation or outcome. Thus, according to Ruth et al. (2002) a high level of certainty leads to positive feelings such as love, happiness, pride, and thankfulness.

One of the fourth appraisal components is a novelty. Barto (2013) presents novelty as the process of encountering or confronting something different from what one is used to experiencing. The term "novelty" refers to how a stimulus event differs from what is expected. Unexpectedness is another term for novelty, and it relates to the

disconfirming of previously held expectations, either directly or implicitly (Reisenzein et al., 2019).

Scherer (1988) indicates that the familiarity and predictability of an event are related to the evaluation of novelty. Novelty increases attention to a targeted stimulus, promotes and encodes memory, and modifies goal-directed behavior (Skavronskaya et al., 2020). Self-compatibility is an essential appraisal factor determining how well an event confirms one's ethical and moral norms (Hosany, 2012).

Norms are concerned with standards of appropriate behavior, and these norms are the standards (Sherer, 1988). There must be consistency between internally defined and perceived standards and expectations related to the event to direct behavior that will affirm moral principles.

Additionally, disgust, a moral sensation associated with norm violation, has been linked to rejections (Fridja, 1989). Therefore, the perception of unfairness evokes unpleasant feelings, notably rage and disgust, which frequently represent rejection and refusal action tendencies (Bedious & Sherer, 2014).

3.2.3. Cognitive Workload

Cognitive workload, or mental workload, is another important concept related to emotions and cognitive processing. Its meaning is related to the disparity between a person's attention resources and the cognitive demands of a job or activity. It is based on the idea that everyone is unique and has a limited cognitive capacity (Carswell, 2005). Cognitive workload recognition is used in various settings, including education, human-computer interaction, and healthcare simulation (McCoy et al., 2008; Oviatt, 2006; Yurko et al., 2010; Fraser & McLaughlin, 2019).

Arico et al. (2016) exemplify machinery operators encountering three different workload states in real-world situations: underload, regular, and overload. According to Pryor et al. (2012), maintaining a balanced and fit workload can assist operators in working safely and effectively.

Many factors can influence learning outcomes. Kidd et al. (2018) stated that individual differences, task complexity, and social and environmental factors could affect learning. According to Blanco et al. (2016), the concept of cognitive workload and engagement are both significant in the learning process. In addition, mental overload is

an undesirable mental state that leads to sluggish and poor performance (Young et al., 2015).

Abeysekera and Dawson (2015) explain that students are required to stop and process information before moving on to the following content; active learning can help learners pace their learning which can assist manage cognitive load. In addition, the authors recommended that advanced learners fast through material they already know in an online course while struggling students can re-watch the video (Jeffery & Bauer, 2020). However, the student's active manipulation of the video's pace may have beneficial effects in reducing workload and increasing learning (Arguel & Jamet, 2009).

Similar questions are present in HCI studies that focus on cognitive load. One of the fundamental works to establish the importance of cognitive load in HCI was Oviatt's influential conference paper: "Human-centered design meets cognitive load theory: designing interfaces that help people think," in 2006 (Oviatt, 2006).

This paper summarized the characteristics of various instructional interfaces designed using user-centered design concepts with the underlying premise that Human-centered design reduces users' cognitive burden and frees up mental resources for improved performance while remaining more alert to the outside environment.

They conclude that cognitive load elicited by using a computer interface should be separated into two components: extraneous and intrinsic complexity. The "extraneous complexity" is the cognitive load associated with the difficulty of using an interface, while the "intrinsic complexity" is associated with the user's learning or work activity (Oviatt, 2006).

From an HCI perspective, the objective is to create user interfaces that reduce unnecessary cognitive load so that people's available intellectual resources may be committed to their primary job, such as walking, driving, providing medical services, or studying a new subject (Oviatt, 2006). The Cognitive Load Theory has been widely applied in education to anticipate students' performance when utilizing new educational materials and interfaces and create educational interfaces that successfully reduce students' load to focus on their core learning task (Oviatt, 2006).

Cognitive load theory is employed in human-computer interaction and health care simulation studies to limit users' cognitive load as they participate in a simulation or utilize an interface. The main concern is how to design educational, computational, and simulation experiences that do not overload the user with cognitive processing. Another

factor that positions cognitive load as an essential construct is that it is associated with higher fatigue and errors (Yurko et al., 2010);

Several techniques for measuring mental workloads have been developed and implemented over time. The NASA – TLX is one of the most extensively used devices for assessing total subjective workload. Also, it has been used to measure operator effort in complicated contexts (Honakker, 2011).

The usage of NASA-TLX is established in various fields, including aviation, healthcare, and other complex socio-technical areas (Colligan et al., 2015). Hill et al. (1992) stated that NASA – TLX has a six-dimensional rating scale extensively used in human factors research. This scale includes physical, temporal, mental, performance, and frustration categories.

Each category has a description to be read before the ranking process, graded on a scale of 100 points with 5-point increments for each activity. The workload index is created by combining these ratings, and participants must select a description for each measurement to be obtained (Schuff et al., 2011). These categories are mental demand - which outlines the level of mental and perceptual efforts; physical demand - which summarizes the kind of physical exertion; and temporal demand - which represents the time pressure individuals feel throughout a specific assignment.

The performance category denotes the percentage of the completed task and the level of satisfaction linked with the performance. The effort category refers to efforts exerted to achieve a level of performance. The frustration category shows how agitated, pressured, and annoyed learners feel instead of being pleased, calm, and contended.

The second portion of TLX aims to generate an individual weighting of these subscales by allowing participants to compare them pairwise based on perceived relevance. It asks the users to decide which metric is most important to workload. The weighted score is determined by the number of times each option is selected (Rubio et al., 2004).

Examples of usage in the healthcare simulation and HCI intersection abound. For instance, Yurko et al. (2010) used the NASA-TLX instrument to understand the relationship between workload and performance during simulator training on complex laparoscopic procedures. Participants with greater workload levels inadvertently injured nearby structures more and experienced more unintended injuries. They confirmed the instrument validity to reflect workload variations during simulator training properly.

Moreover, Barré et al. (2019) investigated the mental workload of novice surgeons using virtual reality medical simulation for gastrectomy training. They found improvement in the mental workload of novice surgeons when trained in VR. They experienced a decrease in cognitive load between subsequent training, indicating virtual reality simulation's viability and potential to train new surgeons.

Fraser & McLaughlin (2019) also investigated the temporal cognitive workload of participants during simulation and debriefing. They found that cognitive scores increased after scenario and debriefing to considerably high levels, also associated with the emotional measurements.

3.3. Emotional Responses in Learning

Learners experience a broad range of emotions in an academic setting, and such emotions increase learners' enthusiasm, which accelerates the learning process. Learning is an active process of integrating existing information and contemporary evidence, establishing networks and links across various knowledge domains.

The role of emotions is crucial in learning, given that emotions develop reactions and attitudes, which lead to behaviors. Moreover, Tyng et al. (2017) clearly stated the significance of emotions in the learning domains of different fields, including neuroscience, education, and psychology. According to Cleveland and Campbell (2012), emotions are phases of a person's intuitive appraisals of their bodily states, urges to act, and the succession of environmental circumstances in which they are embedded.

Moreover, emotions cannot be isolated from the learning environment (You & Kang, 2014). Undoubtedly new research on the impact of emotion and cognition provides substantial evidence that the association between emotion and learning needs to be investigated further (Dirkx, 2008). Emotional experiences are universal and imperative and may be essential in educational settings, as emotions balance each angle of cognition.

According to Mazer et al. (2014), educational activities and milestones such as tests, assessments, homework, and due dates are linked to various negative emotions such as frustration, anxiety, and boredom. Apart from this, Hung (2011) stated that these factors are crucial in learning environments because when learners encounter such challenges, it eliminates the possibility of education and may represent it inconsequential.

Despite the potential challenges and negative impacts, the role of emotions has been recognized as one of the primary learning principles (Pekrun & Perry, 2014).

However, a few researchers have focused on the relationship between learning and emotions (Duffy et al., 2016).

According to Fraser and McLaugin (2019), emotions affect learning in various ways, with evidence of their impact on inspiration, level of handling and solving problems, and excessive cognitive load in simulation settings. However, the connection between emotional responses and learning is complicated, and negative emotions such as anxiety may impede learning. In contrast, positive emotions, such as joy, may aid learning by encouraging collaboration and enhancing problem-solving abilities (Valiente et al., 2020).

Additionally, emotions have been shown in several research studies to influence human cognitive processes such as attention, learning, reasoning, and problem-solving skills (Tyng et al., 2017). Most imperatively, emotional stimuli show up to expand more attentional assets than non-emotional stimuli (Schupp et al., 2007).

Furthermore, emotion's attentional and motivational elements have been linked to improved cognitive performance (Dolcos et al., 2020). Apart from this, the emergence of technological advancement has changed the traditional (face-to-face) learning environment into modern computer-based educational systems and platforms, which produce different emotional responses, either positive or negative (Ananga, 2020)

The main goal of the learning environment is to engage learners in their academic learning, enhance subject understanding, and increase retention of knowledge content for a more extended period, which in turn, emotions affect humans' cognitive functions, particularly attention (Shen et al., 2009).

Learning techniques are defined by learners' attitudes, and developed based on their emotional states (Pekrun & Perry, 2014). Therefore, it is imperative to investigate the critical role of emotional response in the educational domains. Such investigation must be based on particular reasons, for instance, the effects of emotions on learning quality and students' wellbeing (Goetz et al., 2003).

Positive emotions like joy, optimism, and pride help internal and external motivation, flexible learning techniques, and self-regulation (Dewaele & Alfawzan, 2018). Positive emotions influence the students' learning, curiosity, and creativity, helping individuals become more productive and better performers in day-to-day activities (Norman, 2003).

Positive emotions are undoubtedly connected to students' perceptions of self-regulation, whereas negative emotions are adversely linked to students' perception of external regulations (Pekrun et al., 2002).

On the other hand, negative emotions may trigger stress responses, dampening the students' performance capacity. Emotion is, therefore, a double-sided aspect, with the ability to facilitate or obstruct learning. Rudd (2012) argued that emotions affect students' mental and physical states in the learning environment; hence emotions that influence one's mental state can lead to various behavioral changes, frequently reflected through communication (Rafaeli & Sutton, 1989). It demonstrates an important link between emotional arousal and behavioral outcomes.

3.3.1. Emotions in Human-Computer Interaction

Although the study and the importance of emotions are well established in fields such as Learning Sciences and Education, the same is not true of informatics and computer science, even in the discipline of human-computer interaction. According to Brave and Nass (2007), the mainstream HCI premise was that users' emotional aspects needed to be ignored so users could operate computers in the most rational, efficient, and impersonal way. The few mentions of emotions presented it as contradictory or even unimportant.

Brave and Nass (ibid) defend that research from merging psychology and technology created a new perspective on the link between humans, computers, and emotion, resulting in research on the psychology of emotion. This growth can be attributed to the development of technology to understand emotions, with affordable devices that can assess physiological indicators of emotions, coupled with advancements in the capacity to do signal and data processing. This new capability allows ordinary computers to evaluate the user's emotional state (Picard, 1997a).

Furthermore, Brave and Nass (2007) stated that multimodal interfaces with voices, faces, and bodies could now express a much wider and more subtle variety of emotions than primarily textual interfaces could. Indeed, any interface that disregards or fails to describe a user's emotional state can significantly impede performance and risk being conceived as cold, socially awkward, and ineffective (Sears & Jacko, 2002).

Picard coined the concept of Affective Computing in 1997. The notion of Affective Computing is progressively becoming important because it extends the

possibilities of computing technology integrating human emotions. Picard defines affective computing as "computing that relates to, arises from, or intentionally impacts emotion or other emotive phenomena (Wu et al., 2016).

The recognition of human emotions is one of the crucial elements of Affective Computing. Sentiments, feelings, and emotions play a vital role in one's day-to-day life. Emotions facilitate decision-making, communication, negotiations, learning, and situational awareness in human-oriented environments. Poria et al. (2017) stated that affective computing researchers aim to make machines with cognitive abilities that sense, express, and explain the detected emotions. Affective computing has emerged as a discipline (a field of study) to develop systems that can automatically detect emotions.

Affective computing is an interdisciplinary field based on multiple disciplines such as computer science, human psychology, and biomedical engineering (Picard, 1997). According to Poria et al. (2017), emotional recognition is different from sentiment analysis; thus, there are two types of recognition: coarse-grained and fine-grained affect recognition.

Emotion recognition is about fine-grained affect recognition, while sentiment analysis is related to coarse-grained affect recognition. The existing literature posited that several scholars focused on discovering a reliable method or approach to identify and acknowledge subjects' emotional states through machine learning algorithms (Picard, 1997).

Although Affective Computing has been applied in different fields such as healthcare, marketing, and education, its real potential is yet to be attained (Jerritta et al., 2011). The main goal of Affective Computing is to deliver a natural and effective interface (interaction) between computers & humans and enable machines (computers) to acknowledge and recognize users' emotions and deliver customized responses precisely (Cen et al., 2016).

Ahn and Clegg (2018) mention that plenty of initiatives are aimed at understanding the use of computers in educational settings from the perspective of human-computer interaction. However, there is a greater need to investigate how users interact with various computerized devices, particularly affective computing in healthcare (Hudlicka, 2017).

The importance of emotions in understanding user behavior when confronted with various application programs is already well established. However, emotions, on the other

hand, have a much broader impact, affecting our attention, perceptions, memory, and decision-making abilities (Sui & Humphreys, 2015).

According to Picard (1997), the concept of Affective Computing generally refers to the computer's proficiency or capability to sense human emotions and respond to those emotions. Researchers in Affective Computing have been working for decades on developing new devices that can advance understanding of basic human emotions and their role in their experience. MIT – Affective Computing Research Group is an example of the research group.

This group has studied how humans communicate affective and cognitive states, developing new methods to detect stress, moods, and frustration. How can computers be emotionally intelligent and other similar domains, including Affective Computing (Picard, 1997). Dr. Rosalind Picard introduced the term Affective Computing as the author of a book titled "Affective Computing" published in 1997.

Picard (1997) argued that HCI needs to consider emotional reasoning, and developers must consider affective aspects while producing software that interacts with human beings. Emotions, also known as affects, are identifiable conditions that can be structured to enable interactions between machines and humans as human-like as possible.

According to Tao and Tan (2005), the customary practices of affective computing include capturing affect information and modeling, affect comprehension, and expression. Tao and Tan (2005) argued that humans express their feeling (affects) through reactions in facial expressions, body language, voice tones, gestures, and other physical cues like pulse, sweat, and heart rate.

Facial expressions like smiling or frowning are used as a semantic function to disseminate emotions (innate feelings) or informal signals. It works on various coding systems for recognition and creation, a now-established approach to determining user emotions (Tao & Tan, 2005). According to Etcoff (1992) the structure of human facial main parts is parameterized with the help of 37 lines that roughly explain the facial effects' status, and Ekman (1978) developed a coding system for facial actions and expressions.

Likewise, classical works have explored speech's impact on understanding emotions. Scherer (1986) stated that human emotional speech varies concerning the auditory features of individuals; for instance, inflection features and pitch, variable jitter, range, contour, F0 level, and speaking rate have been investigated by some scholars

(Petrushin, 2000). Some experiments by Tato et al. (2002) revealed how quality features are used other than inflection features.

Another prolific set of emotion-related technologies is based on body gestures and movement. Body language can be defined by the place or position of the body and its changes or variations with time (Tan & Tan, 2005). In gesture processing, the focus remains more on hand tracking because it communicates dissimilar meanings to express the behaviors and sign language.

The precision of Affective Computing artifacts is growing with advancements in new algorithms and developments in AI (such as neural networks and deep learning) to detect human emotions with the help of advanced sources such as facial electrography and blood volume pulse (Dilmegani, 2021).

Dirkx (2008) stated whether it is to discover how emotions impact learning or decision-making or their role in human-computer interaction (Beale & Peter, 2008). However, because emotions influence decision-making, affect attention, and generate meaning, design based on emotions can significantly impact the overall user experience (Van et al., 2017).

There are two main approaches to applied emotional design that can be distinguished. The first is concerned with changing an object's aesthetic appearance or interface, while the second promotes fluid and engaging interactions (Triberti et al., 2017).

Moreover, embedding emotional stimuli into interfaces elicited positive emotions in learners and improved learning outcomes, according to studies in multimedia learning (Plass & Kaplan, 2016). In addition, flow experience is quite common in technology usage (Triberti et al., 2017). As a result, flow-inspired design models were developed and applied to interactive digital technologies like educational games and virtual reality (Hsu, 2019).

Besides, gamification or the inclusion of game mechanics in interfaces such as prizes, achievements, interactive storytelling by framing with emotional scenarios and compelling characters, and motives could be the approaches for promoting emotions through engagement (Liu et al., 2017).

Further stated by Douglas (2008) that effective design has demonstrated that: emotional design can be conceived not only by including pleasant and engaging elements

in the interface to enhance gratification but also as the acknowledgment and measurement of emotions supply to technology and revise its functioning

Apart from this, users learn to manage and control their emotional states, for example, by watching virtual environment functionalities (such as a burning fire) change in response to their psychophysiological stimulation (Zou & Cao, 2017).

3.3.2. Healthcare simulation and emotions

Simulation has acquired much consideration in healthcare education and can be considered a safe and reliable teaching method. As a result, Helle and Saljo (2012) opined that simulation-based learning environments are increasingly being employed in medical education. The simulated situations are designed to provoke emotional responses from students (Bogo et al., 2013).

Previous research studies have primarily focused on the relationship between stress and performance (Schlairet et al., 2015). Postareff et al. (2017) study aimed to understand students' emotional reactions to simulation-based learning, finding that learning is linked to various emotions.

Artino (2009) indicated that contemporary research studies on emotional experience in online courses have shown that boredom and frustration negatively influence students' satisfaction and motivation to continue.

Noteborn et al. (2012) stated that learners may experience a high level of boredom yet still achieve high academic performance on a practical assignment. Surprisingly, boredom had a detrimental impact on metacognitive learning techniques, but frustration, on the other hand, can be helpful.

According to Wosnitza and Volet (2005) frustration can pay more attention to the task or the technology in online learning contexts. In contrast, Schlairet et al. (2015) indicated that simulation-based healthcare education elicits primarily pleasant feelings, which persist beyond the course timeframe.

Simulation-based learning environments provide effective teaching and learning circumstances, particularly since they elicit mostly pleasant emotions in simulation, leading to improved learning and performance (Damasio, 2001). According to Ainley (2011) interested students are more likely to demonstrate more significant levels of learning and performance, and motivational difficulties are less likely to exit.

Lower-order cognitive and metacognitive processes frequently preceded negative emotions (anxiety & stress) and are frequently preceded by lower-order cognitive and metacognitive processes (Duffy et al., 2016), impairing team performance during simulated emergency training.

The simulation is a method that also guides professionals in learning how to deal with negative emotions like stress. Keskitalo and Ruokamo (2016) findings reveal that students' emotional responses are reviewed in a debriefing session following the simulation exercise, providing constructive feedback for students, which ultimately guides them to learn how to handle challenging situations to cope with unpleasant emotions.

In conclusion, simulation-based teaching is well accepted and elicits mostly pleasant emotions in participants, consistent with prior research findings (Schlairet et al., 2015). According to Keskitalo and Ruokamo (2021), simulation-based learning may be defined as 'hard fun' since it generates positive experiences through challenges and outcomes. McConnel and Eva (2012) highlight the opportunities of the challenging and positive experiences that may turn into improved learning and performance.

Taylor and Reyes (2012) suggested that better understanding students' emotional intelligence, psychological empowerment, resilience, and spiritual wellbeing will be beneficial in developing curriculum, teaching methodologies, and techniques. Furthermore, it will undoubtedly improve higher education practices since the environmental factors trigger an emotional or reactive behavior based on that appraisal.

3.4. Quasi-Experiments and Natural Experiments

The use of experimental research has become more prominent in recent years. True experiments are the most conventional type of experimental design (Jhangiani et al., 2015). Baker (2017) teaches that true experiments include random distribution of study subjects into respective groups to determine a cause-and-effect relationship.

True experiments or experimental designs are characterized by randomization as their foundational principle. Baker (2017) explained that for an experiment to be categorized as a true experiment, it must have the following attributes:

i. ***A modification, treatment, or intervention*** –represents the independent variable in a study

ii. ***A Control for Interventions*** – includes control techniques such as a control group and randomization. It provides significant room for comparison and contrasts to identify the research outcomes.

iii. ***Randomization -*** encompasses the random selection of study participants where every individual has an equal chance of being selected for either of the groups.

Random clinical trials (RCTs) are sometimes called efficacy studies, intended to test hypotheses under ideal and controlled conditions. They are the gold standard for assessing the effectiveness of a treatment or intervention. Baker (2017) clarified that participants are randomly allocated to either the intervention or control group in such studies and then tracked in time (prospective) to compare outcomes.

When done correctly, randomization provides equivalent groups, allowing differences between the groups to be attributed to the independent variable. To avoid selection bias, random allocation of participants into groups entails establishing a random sequence to hide it from anybody who recruits subjects to a study (Mangold and Adler, 2019).

However, other experimental designs, such as quasi-experiments and natural experiments, do not include randomization. Dean (2016) clarified that one distinguishing feature of a quasi or natural experiment is that it does not involve the random application of an intervention or group allocation. The study subjects are intentionally chosen according to a pre-determined research criterion (quasi-experiment) or are part of a natural assignment beyond the researcher's control (natural experiment).

Thus, quasi-experiments are like true experiments, except they are not randomized based on specific conditions of the study subjects (Baker, 2017). According to Harris et al. (2006), randomization is advantageous because it distributes known and unknown confounding variables evenly between the intervention and control groups on average.

Harris and collaborators also highlight that quasi-experiments do not employ random sampling; thus, there will be the problem of confounding variables. However, they also explain that there will always be differences in experimental conditions and results for natural and quasi-experiments.

In a quasi-experiment study, an independent and a dependent variable are selected, and the independent variable is manipulated, followed by the determination of the dependent variable (White & Sabarwal, 2014; Jhangiani et al., 2015). Harris et al.

(2006) explained that in this way, quasi-experimental studies reduce the directionality problem present in other research designs.

Leatherdale (2019) mentions that quasi or natural experiments are employed when it is not feasible to apply randomization in interventions when considering ethical reasons. In those cases, the participant is selected after a particular intervention has taken place (or will take place), with the circumstances for its selection and allocation not being in the control of the researchers.

Studies with a quasi-experimental design are frequently used in medical informatics. Informatics interventions cannot generally be randomized to specific locations or populations, not strictly because of the feasibility but practical implications.

Harris et al. (2006) share the example of a pharmacy order-entry system to explain that selecting only limited hospital units to implement a new system is usually not feasible. In similar cases, a natural experiment may be a better alternative.

In this case, although the randomization could be implemented, effects such as the interaction between groups in both wards (in which the software was implemented and non-implemented) could change the effect of the implementation. Also, there may be more confounding variables in different units. For instance, the first unit to serve as a pilot implementation location might have more technological resources or smaller than average team size than other units.

Mangold and Adler (2019) clarified that quasi-experimental research could be used to measure post and pre-intervention in non-random control groups. However, even in quasi-experimental designs, Baker (2017) indicates that pretest data, also known as baseline data, can be used with demographic data to determine if randomization produced equivalent groups. This baseline data comparison is usually mentioned directly at the start of the Results section and displayed in one or more tables.

Therefore, quasi-experiments are most likely to be employed when random assignment is difficult or unfeasible. They are frequently used to assess the efficacy of a treatment, such as psychotherapy or educational intervention, with abounding examples.

A quasi-experiment was conducted on an online formative assessment and tutoring system to check whether the system could improve the scores of middle school students. The acquired data was pretested and post-tested to check treatment efficiency (Koedinger, McLaughlin, and Heffernan, 2010). The results indicated a significant improvement in students who have experienced treatment.

Another example in IT and social management is a study examining whether information disclosure could improve Internet security. The study investigated European and North American countries and outgoing spam of 1718 autonomous systems (Linden, Quarterman, and Whinston, 2013). The results showed that social comparison and social information could encourage desirable behaviors effectively.

Quasi-experiments have increased internal validity compared with non-experimental research. Furthermore, they are increasingly being employed in researching various treatments, public health challenges, and policy changes. They are beneficial in social sciences, political science, psychology, biological sciences, and epidemiological investigations (Jhangiani et al., 2015).

These experiments are purposed to determine the causality between the given interventions and respective outcomes. Additionally, Cook and Campbell (1979) indicate that quasi-experiments can be used in non-random control groups to measure pre-intervention and post-intervention.

However, Jhangiani, Chiang, and Price (2015) warned that there should be variations in conditions since quasi-experiments do not employ random sampling, therefore not preventing the issue of confounding variables. According to Baker (2017), quasi-experiments fall between true experiments and correlational studies regarding their internal validity. According to (Jhangiani, 2015), there are numerous quasi-experiment designs, the most common being non-equivalent groups design, pre-posttest, time-series, and combination designs.

A Non-equivalent group nomenclature indicates the dissimilarity and non-equivalence between the study subjects. It results from the non-random selection of the quasi-experiment participants (Harris et al., 2006). Jhangiani et al. (2015) explained that researchers usually take steps to ensure that their study group is as similar as possible in this design. Such steps eliminate the threat of confounding variables and may significantly enhance the study's internal validity. Nevertheless, since non-equivalent design employs non-random sampling, there is a chance that multiple confounding variables are present.

Regarding the pre-posttest design, Jhangiani et al. (2015) explain that this designation comes from the measurements of the study dependent variables. The dependent variable is first measured before the intervention's start and after the intervention has been completed.

The logic of this design is that if the average post-test score is higher than the average pretest score, it is reasonable to believe that the effect could be attributed to the intervention. However, one cannot always conclude this with certainty because there could be other reasons the post-test results are higher than the pre-test.

Jhangiani et al. (2015) mention that potential reasons must be initially ruled out as the multiple other factors that may cause a change in the dependent variable. They point out two factors, regression to the mean and spontaneous remission (the natural tendency of medical and psychological issues to improve over time even without specific treatment).

An interrupted time-series design is similar to the pretest-posttest design as it measures outcomes before and after the intervention. However, one of its distinct features is that it takes multiple measurements at different time series to determine the causality (Jhangiani et al., 2015). This design can be viewed as a series of measurements taken over specific time intervals. Finally, to describe the "interrupted" component of the name of the design, it implies that the time series was interrupted by an intervention or treatment.

Combination designs are generally considered better than non-equivalent group designs and the pretest-posttest designs as it combines features of both.

In this design, the treatment group is given a pretest, then a treatment, followed by a posttest. In contrast, the control group is administered a pretest, is not given the treatment, and then a post-test. The results indicate whether the intervention has improved the conditions of participants receiving treatment more than those who did not (Jhangiani et al., 2015)

Natural Experiments

A natural experiment denotes a quasi-experiment in which the selection of the research participants into treatment or control groups without the researcher's intervention. This method has been widely employed by many researchers and scholars in multiple scientific fields.

Dean (2015) explained that the independent variable varies inherently through some natural factors independent of the dependent variable in a natural experiment. Such natural independent variables can be understood by an example of an epidemic or a hurricane. History has conducted multiple studies on storms, floods, and hurricanes. Craig et al., (2012) added that natural experiments could be used to examine labor markets,

commodity prices, food availability, markets, and wages in these incidents, among other factors.

Natural experiments employ two separate groups of subjects with characteristics that make them broadly comparable. As the name indicates, these experiments are usually not conducted in artificial settings or a laboratory, as they require natural circumstances and outcomes.

Sekhon and Titiunik (2021) pointed out that the subjects or recipients of treatment for natural experiments are chosen explicitly following a pre-set criterion. The first group receives the intervention while the second group does not and is thus called the control group. Dean (2015) pointed out that the control group in a quasi-experiment is the comparison group, unlike true experiments.

Natural experiments differ significantly from true experiments. Khullar and Jena (2021) described that the difference between natural and true experiments lies in the researcher's lack of control of the assignment of participants. However, natural experiments typically use quasi-randomization. It designates the allocation of study groups that is not purely random but not assigned by a researcher.

However, one crucial feature of those studies is that the similarity of groups can be measured. They should be as close as possible regarding sociodemographic characteristics, health status, medical issues, or other factors that may influence the outcome of the dependent variable (Khullar & Jena, 2021).

Although the researcher has no control over the treatment or exposure allocation, one assumption is that the allocation is unrelated to other factors that influence the dependent variable, especially in complicated real-world contexts, allowing strong causal information and providing impact sizes that are near to RCT (De Vocht et al., 2021)

De Vocht et al. (2021) call for natural experiments by describing that such experiments can be more valuable and informative than traditional experimental methods. Furthermore, the more closely a study design resembles RCTs (in terms of sample randomization or equivalence) in natural experimental methods, the more valid its findings tend to be.

Furthermore, natural designs have significant benefits over traditional research designs and thus are widely employed in various disciplines such as public health research, epidemiology, and public health, among many others (De Vocht et al., 2021).

Various analyses in natural experiments identify the potential causation and impact relationships (Sekhon and Titiunik, 2012). Leatherdale (2018) added that some

include interrupted time series analyses, event studies, and DID (difference in differences). Such designs are limited and cannot wholly control the unmeasurable variations between study groups.

Dean (2016) further explained that the variety of methods available for natural experiments depicts their versatility in research studies. Leatherdale (2019) added that this versatility portrays that the study design can be adapted to the characteristics or nature of the policy intervention. Thus, these experiments are more suitable for different policy and educational evaluations.

3.5. Non-Inferiority Studies

Although frequently used to test if the means of two groups are equivalent, the traditional t-test is not adequate to conduct and establish equivalence between two groups. When a traditional t-test is non-significative (p-value more significant than the established significance level), it simply indicates that it is not possible to reject a null hypothesis. Consequently, it does not allow the researcher to conclude that the means-tested are equal. The formal outcome of a non-significative test is to say that there is not enough evidence to reject the null hypothesis (Limentani et al., 2005).

Lewis and Lewis (2005) explain the limitation in detail in a response article justifying their usage of the t-test to establish the equivalence in cognitive measures between two groups in a previous article. They argued that when a p-value is obtained, it is compared to a predefined alpha value (usually 0.05). The null hypothesis is rejected if the p-value is less than 0.05, accepting the alternative hypothesis. In their case, they accepted the null hypothesis based on a non-significative p-value and wrongly concluded that the group's means were equivalent.

The response article recognized that this method for establishing equivalence, although frequently used, is not robust. When the t-test provides no evidence to reject the hypothesis difference, it does not automatically establish the equivalence since failing to reject the null hypothesis does not accept the null hypothesis. The authors re-evaluated the analysis and concluded that when using a properly designed equivalence testing, only two out of five measures were considered statistically equivalent, therefore calling the need for the conduction of more statistically correct equivalence, superiority, or non-inferiority studies (Lewis and Lewis, 2005)

More recently, in the Natural sciences and Medical fields, numerous studies are being conducted to test whether new a treatment or intervention has equal or non-inferior efficacy to treatments presently utilized (Walker & Nowacki, 2011).

Noninferiority studies are conducted when alternative treatments are not expected to be more effective than conventional treatments but offer additional benefits. Potential benefits of the alternative treatment that can support its use despite a non-superior effectivity may include a safer intervention, fewer side effects, quicker administration, less need for laboratory testing, or a reduced cost.

A non-inferiority approach aims to show that the experimental treatment retains at least a portion of the established control impact or effect while having significant practical or clinical benefits that compensate for the loss of efficacy (Hung et al., 2003).

In recent years, equivalence and non-inferiority studies have grown in popularity in medical education. Usually, new cost-effective methods that produce equivalent outcomes to established approaches in educational research are frequently pursued (Klasen and Sopka, 2021).

According to Klasen and Sopka (2021), when a comparison study is performed in an educational setting, the goal is to demonstrate that one instructional method provides a superior learning outcome to the other.

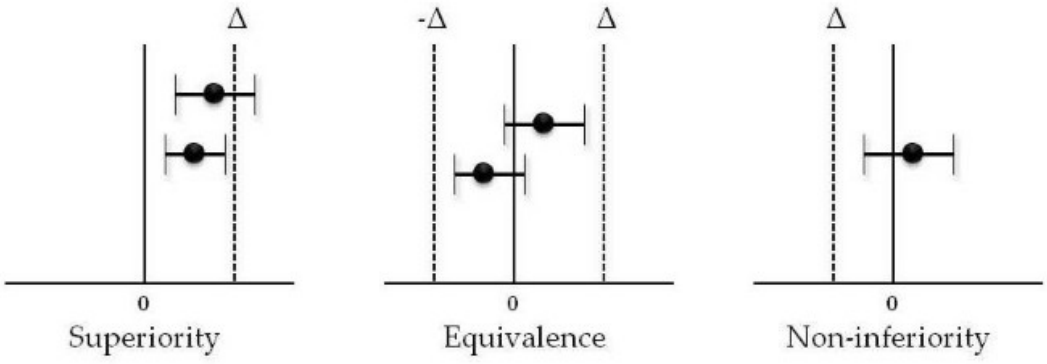
However, to adopt the superiority test to compare a novel approach with limited knowledge about its outcomes with a conventional approach is not adequate for two reasons: 1) it is established that the standard produces satisfactory learning outcomes, and therefore it is difficult to provide a better alternative if the standard does represent the best in terms of performance; and consequently, 2) when the current standard already produces valuable learning outcomes, the superiority of a novel approach is not required. Instead, a similar performance is sufficient, considering the advantageous characteristics of the new approach.

Another important application of a non-inferiority approach is to test two treatments when comparing a new treatment with a non-treatment (or placebo) is not viable or ethically possible. D'Agostino Sr et al. (2003) indicate that placebo-controlled studies help test the efficacy of treatments in the medical field, being the most efficient in detecting a beneficial effect with the smallest number of participants. This method is advised to be used in cases where the standard treatment has not been demonstrated successful and there are no risks associated with postponing treatment.

However, D'Agostino Sr et al. (2003) establish that placebo-controlled (non-treatment) studies are frequently unethical, given the numerous proven successful treatments in most cases. Therefore, active-controlled trials are usually the best option in these cases, comparing a standard and a new treatment. Noninferiority approaches focus on proving that an alternative treatment is not inferior to the established standard. In contrast, superiority or traditional experiments aim to determine whether there is a difference in efficacies or one of them is superior.

This difference in approach has important implications for the types of the study questions, statistical tests, and potential responses (Ferreira-González, 2014). Schumi and Wittes (2011) instruct that a non-inferiority study aims to demonstrate that the tested alternative is not 'unacceptably worse' than the control (or comparison). Figure 31 illustrates the difference in the margin and its implications for testing.

Figure 30: Margins in superiority, equivalence, and non-inferiority studies.



Source: Schumi, J., & Wittes, J. T. (2011). Through the looking glass: understanding non-inferiority. *Trials*, *12*(1), 1-12.

The purpose of testing for superiority, equivalence, and non-inferiority are distinct, and therefore the traditional approach to test hypotheses for superiority must be adapted in each case. In the scientific method, independently of the type of testing, the research or alternative hypothesis represents what the investigation attempts to prove in hypotheses testing by refuting the null hypothesis. The null hypothesis is the logical oppositive of the research hypothesis (Walker & Nowacki, 2011).

Walker & Nowacki (2011) mention that the three approaches differ in their premises and the practical aspects of statistical testing. The classic (two-sided) comparison studies approach tests if there are differences between the alternatives.

Equality cannot be ruled out if the evidence for a difference is insufficient. Equivalence studies, on the other hand, aim to prove statistical equality.

The null and research hypotheses in equivalency testing are the opposite of those in a standard comparison study. Finally, the premise of non-inferiority approaches is that the alternative treatment is either equivalent to or superior to the standard treatment. Table 24 presents a summary of the different types of studies and their hypothesis statement:

Table 24: Key Differences between traditional, equivalence, and non-inferiority studies

Type of study	**Null hypotheses**	**Research hypothesis**
Traditional approach (comparative)	There is no difference between the treatments	There is a difference between the treatments
Equivalence	The treatments are not equivalent	The new therapy is equivalent to standard treatment
Noninferiority	The new therapy is inferior to the current treatment	The new therapy is not inferior to the current treatment

Source: Adapted from Walker & Nowacki (2011)

According to Klasen and Sopka (2021), two treatments will never generate identical results in practice. The fundamental purpose of the non-inferiority approach is to establish which differences across treatments may still be regarded as unimportant when determining equivalence. Any research that aims to compare two treatments and test whether one is not inferior must define this difference.

Ferreira-González (2014) reveals that the definition of what constitutes "at least as effective as" or "not inferior to" the standard treatment in non-inferiority studies must be determined a priori by establishing a non-inferiority margin or threshold.

In other words, the goal of non-inferiority is to establish that the experimental treatment's impact is not inferior to the standard treatment to "a certain amount," which is known as the non-inferiority threshold, non-inferiority margin, or delta (δ). The non-inferiority margin (δ) must be chosen based on the best available evidence on the efficacy of the standard treatment vs. a control condition, considering the degree of certainty or

uncertainty ascribed to the impact of the standard treatment is often conservative (Ferreira-González, 2014).

The research hypothesis is that the new treatment efficacy is just δ units lower than the existing therapy's (when higher is better). If the lower limit of a (1–2α) x 100 % confidence interval for the difference (new – current) is greater than -δ (non-inferiority margin), non-inferiority is demonstrated (Walker & Nowacki, 2011).

Establishing the non-inferiority margin is critical when testing for equivalence or non-inferiority. Suppose a small value of δ is selected. In that case, the equivalence test is more stringent, being more difficult to establish if there is an equivalence or non-inferiority between the two treatments. The margin directly impacts the statistical test but the overall study reliability. The margin must be justified in evidence and practical considerations, including applicable regulatory recommendations (Walker & Nowacki, 2011).

Several alternative methods can complete the process of defining a non-inferiority margin. One common approach is to define the margin based on previous research. Another frequent approach is to base the margin on the existing treatment margin of superiority over the placebo, typically by setting the margin to be a fraction of the lower limit of a confidence interval of the difference between the current treatment and the non-treatment (Walker & Nowacki, 2011).

Examples of non-inferiority cases abound in medical research but also in other fields. The approach is used to investigate new drugs and therapeutical treatments. A study by Yang et al. (2014) compared *acarbose* and *metformin* for diabetic patients. Another study compared *prostate cancer patients' ultra-hypofractionated and conventionally fractionated radiotherapy* (Widmark et al., 2019). Likewise, Koster and collaborators (2017) employed the approach to evaluating the safety of mechanical chest compression devices against established gold standards.

However, the method's usage in other fields is expanding. In a behavioral intervention study, Millof and collaborators (2019) conducted a non-randomized trial to test the non-inferiority of a virtual reality exposure therapy against an in-vivo treatment.

More examples are found in the intersection of informatics, education, and medicine, such as the study conducted by Meltzer et al. (2018) to compare telerehabilitation treatment to in-person to treat post-stroke language disorders.

Recently Veldhuijzen et al. (2021) tested the non-inferiority of computer-based patient education with nurse counseling before colonoscopy sessions, finding that the

computer-based approach was non-inferior to the nurse counseling. The alternative educational approach displayed lower patient sick leave while being non-inferior, indicating that the new approach can serve as an alternative educational tool.

Likewise, Oh and collaborators (2021) investigated the effectiveness of remote simulation for periodontal instrumentation during the COVID-19 pandemic, finding that the remote instruction produced equivalent student learning outcomes with reduced staff, time, and financial investment.

Klasen and Sopka (2021) comprehensively discuss the application of non-inferiority tests, specifically in medical education, sharing that various didactic approaches are available for teaching and developing medical skills.

However, using a non-inferiority approach needs to have pragmatic considerations. Testing a new method has an outcome comparable to a gold standard is relevant when benefits are using the new approach, such as lower costs, decreased use of material or staff, or is feasible given the circumstances.

Klasen and Sopka (2021) suggest that a critical application of non-inferiority studies is to compare different teaching approaches, especially in-person formats like lectures and seminars, with online/virtual educational approaches. They call attention to the fact that the substantive question does not aim to determine if the online format is better than the in-person alternative. Instead, the question is to verify if an equivalent amount of knowledge might be gained digitally for less costly and using less infrastructure.

Regardless of the type or complexity, statistical comparison tests cannot determine if an instructional approach produces 'good' or 'sufficient' results. Being sufficient is a matter of definition rather than a statistical analysis. Since two methods with statistically equivalent outcomes can be equivalently great or poor, several factors must be evaluated to define an approach's quality (Klasen and Sopka, 2021).

However, equivalence and non-inferiority experiments help determine viable educational approaches, especially when one of the approaches compared is established as effective and a new approach with diverse benefits or fewer constraints is proposed (Klasen and Sopka, 2021)

Chapter 4. Research Methodology and Methods

The objective of the present study was to determine whether telesimulation is a viable comparable instructional alternative using as the criteria emotional and cognitive responses induced in participants during healthcare simulation. The study was conducted in two leading simulation centers in the United States that resorted to telesimulation as a teaching strategy during COVID-19 when simulation centers all around the country (and most of the world) were closed.

The present study employed a natural-experiment design with a mixed design, with a combination of pretest-posttest design and comparison group design to test the hypothesis regarding the non-inferiority of the telesimulation as an alternative to in-person simulation, focusing on the emotional and cognitive aspects.

We employed surveys with widely psychometrically validated scales to measure emotions, workload, sense of realism, and presence and debriefing feedback (dependent variables) in three distinct temporal points, using pre-posttests whenever applicable. However, this chapter will discuss the methods and methodological approaches in detail.

The research design – natural experiments - was chosen to generate a robust research design coupled with the feasibility of studying the phenomena and collecting data in an unaltered, instructor-led simulation session, considering the lack of control of the course schedules and content by the researcher (Craig et al., 2012).

Still, regarding the research design, the study employed a *non-inferiority* approach to testing the non-inferiority of telesimulation compared with in-person simulation, given the specifics of our research hypothesis and research question.

The study had the in-person simulation as the control variable (independent variable) and other control variables such as the scenario complexity and other demographic variables to account for confounding factors while increasing the internal validity of the research.

During this chapter, we will present the methodological aspects of the study, including the inspiration, the formal definition of the research as a natural experiment, including the treatments and control conditions, the description of the place of conduction of the research, procedures, instruments employed, sample sizes and power analysis and other methodological aspects of this research.

4.1. Methodological Considerations

Science is a methodological and systematic approach to the acquisition of new knowledge. The National Academy of Sciences defines science as "the use of evidence to construct testable explanations and predictions of natural phenomena and the knowledge generated through this process."[9]. Instead of merely casual observations and the use of an informal way of knowing the world, scientists seek to generate new knowledge through careful observations of reality, using systematic, controlled, and methodical approaches to generate valid and reliable conclusions about the subject studied (Marczyk et al., 2005)

Kothari (2004) defines research as "an inquiry into the nature of, the reasons for, and the consequences of any particular set of circumstances.". The Merriam-Webster.com dictionary lists research as *"investigation or experimentation aimed at the discovery and interpretation of facts, revision of accepted theories or laws in the light of new facts, or practical application of such new or revised theories or laws"*[10].

What makes scientific research unique is the scientific method, described by Roger Bacon in detail in his masterpiece *Opus Majus* in 1267. The scientific method is the foundation for all scientific research, an approach to knowledge acquisition not limited by a single method but composed of the following elements: an empirical approach, observations, questions, hypotheses, experiments, analyses, conclusions, and replication (Marczyk et al., 2005).

Research methodology is a science of studying how research is conducted scientifically. Researchers generally adopt the various steps in studying a research problem. Researchers need to understand the assumptions underlying various techniques and the criteria by which they can decide which ones are relevant and not.

In that regard, Kothari (2004) distinguishes the concept of *research methods* and *research methodology*. Research methods are the collection of the procedures/techniques used to investigate any research topic. On the other hand, research methodology is a method for systematically solving a research problem. A research methodology employs a scientific approach to solve a research question concerning the various stages a researcher takes to investigate a research topic and its reasoning.

[9] Science - https://www.nationalacademies.org/evolution/definitions
[10] Research – Definition https://www.merriam-webster.com/dictionary/research

Therefore, any researcher must comprehend the research methods and procedures. It is insufficient to know how to perform specific statistical or data collection procedures. However, they must understand the assumptions underlying chosen techniques and their aptness to solve problems. Additionally, the researcher must reveal and explain her/his research methodological approaches and clearly and precisely share their selected decisions. The rationale for their selection is to be evaluated by their academic peers (Kothari, 2004).

In conclusion, a research methodology encompasses many components, including research methods. Therefore, besides discussing the research methods, a methodology section should include the research's justification, a clear problem statement, a testable study hypothesis, and data analysis approaches. It should also present the choice of methods and the criteria to exclude alternative methods so the choice can be evaluated (Kothari, 2004).

The following sections will describe our research design, approach, instruments and measurement, data analysis, and other research methodological choices for our study. We will assess our study against its objectives and be better comprehended.

4.2.Research Design

Kothari (2004) defines research design as the challenge following the definition of a research problem, involving all decisions regarding the research and how the problem will be addressed. It involves a series of decisions that determine the study's objective, any geographical or site collection decisions, data periods for collection, and the available resources.

Therefore, Kothari (p.31, 2004) quotes Selltiz et al. (1962) p. 50 “A research design is the arrangement of conditions for collection and analysis of data in a manner that aims to combine relevance to the research purpose with economy in procedure”.

In other words, the research design specifies the structure for the conduction of research, with an indication of procedures for collection, measurement, and analysis of data, including the rationale for conduction of the study, sample information, data analysis procedures (and tools) and objectives of the research.

One of the advantages of a research design is that it facilitates the research activities' conduction, improves the study efficiency, and makes it easier to understand and justify the choices employed during its planning phase (Kothari, 2004).

The methodological choices, or the research design, need to consider the feasibility of data collection, the research availability and skills, the specific problem being studied, and the associated resources available.

Kothari (2004) presents the different components related to a research design, organizing them into a) dependent and independent variables; b) extraneous variables; c) control; d) confounding relationships, e) research hypothesis; f) experimental and non-experimental hypotheses testing; g) experimental and control groups; h) treatments; i)experiments, and j) experimental units. In the following paragraphs, we will define each component briefly.

The first component explained is the dependent and independent variables. Kothari (2004) explains that the dependent variable is a consequence of another variable in the study, while the variable that causes or influences the former is known as an independent variable. On the other hand, an extraneous variable is a variable that is not related to the objective of the study (e.g., not included in the hypothesis) but may affect the value of the dependent variable.

Considering extraneous variables is crucial since any study should be designed so the effect on the dependent variables can be attributed to the independent variable and not affected by the extraneous variable influence or impact (Kothari, 2004).

The next component described by Kothari (2004) is the definition of controls to minimize the influence of extraneous variables on the dependent variables. In the context of experiments, controls are usually put in place to limit the conditions of an experiment to provide higher study validity.

Confounding relationships, on the contrary, happen when the influence of extraneous variables influences the dependent and independent variables are present in the study, despite efforts to control the influence. It is not always possible to eliminate confounding relationships. The impact of cofounding relationships can be attained by matching characteristics of groups, for instance, so all the groups are affected in a similar form by the confounding variables.

Kothari (2004) presented two other components: the research hypothesis and experimental and non-experimental hypothesis-testing research. The research hypothesis is a formal statement tested and hopefully rejected about a research problem. Therefore, the hypothesis must be testable and falsifiable (capable of being proven false) to be considered a research hypothesis.

In the sequence, Kothari presents a series of components related to experiments, making first a distinction based on the presence of an experimental design for hypothesis testing. The research can be either experimental or non-experimental. In Kothari's (2004) definition, an experimental design is one in which the independent variable is manipulated, while designs in which the independent are not manipulated are called non-experimental.

In the theme of experiments, experimental and control groups are mechanisms to improve the methodology and attribute effects to treatment. A control group is exposed to normal or no treatment. In contrast, the experimental group (or treatment) undergoes a condition that is either new or is being tested against the established condition.

A treatment is a set of conditions that differentiates both groups, and it could be a new educational approach, a new drug, or a different intervention. An experiment examines a hypothesis to answer a research problem. Kothari (2004) defines two types: absolute and comparative.

While an absolute experiment often tests the hypothesis against a metric or of the presence of an effect (an increase of a value by more than a prespecified value), a comparative experiment usually evaluates two groups in which one receives the treatment, and the other does not, is the most common type of experiment. Finally, the experimental units are the blocks or units in which treatment and control are applied.

The Research Methodology and Methods chapter will describe the instruments, variables, treatments, statistical procedures, and other research characteristics. For the reader's convenience, we have compiled the main features of the research design, described in more detail in the upcoming sections. Table 25 presents a summary of the research design:

Table 25: Summary of research design

Research Design Component	Description
Dependent Variables	Emotional and Cognitive responses, operationalized by the following scales: • Emotion Scale – Feldman, Barret and Russell (2012) • FAS - Felt Arousal Scale (FAS) (Svebak & Murgatroyd, 1985) • NASA Task Load Index (NASA-TLX) – (Hart, 2006; Hoonakker et al., 2011) • Emotion Appraisal – custom instrument largely based on Scherer and Moors' (2019) categorization of appraisal dimensions of emotion process. •Fidelity/Realism dimension of Simulation Design Scale (SDS student version; Jeffries, 2016) • Dinh's Concept of Presence (COP) tool (Dinh, Walker, Song, Kobayashi, & Hodges, 1999)
Independent Variables	Treatment (telesimulation vs in-person) and scenario complexity (low and medium complexity)
Treatments	Telesimulation (compared to in-person simulation as the control treatment)
Extraneous variables & Cofounding factors	Class standing (nursing and medical students, professional in healthcare professions), gender, age, prior experience in simulation and in the scenario being depicted, stress levels, COVID-19 pandemic effect on emotional and cognitive aspects of learners.
Type of Experiment	Comparative experiment (telesimulation vs in-person) with a non-inferiority approach
Experimental and control groups	Participants engaged in telesimulation are the treatment group (participants engaged in in-person scenarios are the control group)
Experimental unit(s)	Simulation participants (students/medical professionals)

Source: Compiled by the author using Kothari (2004) categorizations of research

The present research employed a quantitative approach with a quasi-experimental design that primarily examined whether telesimulation was a non-inferior alternative to in-person healthcare simulation. We analyzed the remotely facilitated and in-person simulation experiments with a carefully established process to ensure anonymity and

protect participants' data, information, and consent. The Internal IRB submitted and approved the research, as determined by the policies of the centers where data was collected.

The availability of instruments took our choice for a quantitative approach to measure the constructs of interests and the feasibility of collecting data during the experiment, with minimum time elapsed from the event and data collection, to provide a more precise recall of the emotional and cognitive states.

The variables were operationalized using validated instruments with established psychometric proprieties in three different moments established as necessary within the healthcare simulation literature and potentially generating cognitive and emotional responses.

Additionally, we performed confirmatory factorial analysis and reliability analysis to determine the adequacy of the instruments for our study. We collected measures before the simulation sessions, after the simulation, and after the debriefing process.

To minimize the effects of extraneous and confounding variables, we collected information on the participants' class standing, prior experience with simulation, age, gender, and other information to ensure the equivalence of our groups. Additionally, we balanced our samples in terms of scenario complexity. We have compared the groups to ensure their equivalence regarding those variables during our data analysis.

We employed a combination or mixed design, depending on the variables and the comparisons performed. The study can be classified as a between-subjects (or non-equal group design) when comparing cognitive response (mental workload), sense of presence and fidelity, and perception of feedback quality in terms of participation telesimulation or in-person simulation.

These variables were not measured in a pre-posttest design since they could only be measured after the exposure to the simulation scenario. A baseline level does not make practical or academic sense (such as sense of presence, task difficulty, or perception of an event, post-fact). These measurements are taken to be the actual effect. In contrast, in a pre-posttest approach, the effect is the difference between the measurement after the event and the measurement taken at a baseline moment before being exposed to a treatment.

The study can also be classified as a within-subjects design in aspects that were measured in distinct temporal points, such as the emotional responses, which had three temporal points, therefore utilizing a pre-posttest approach.

Shaughness et al. (2000) indicate that using a combination design is better than using a pre-test and post-test or a non-equivalent group design. It provides additional control and combines elements of both approaches, making the design more robust.

Finally, in terms of analysis, we employed a non-inferiority approach to compare the in-person and telesimulation. The non-inferiority approach was determined by the importance of testing for a meaningful difference between the two simulation methods and not only if they are statistically different.

The implications and advantages of using non-inferiority will be discussed in detail during the analysis, results, and conclusions. We employed parametric and non-parametric versions of Two Sample T-tests and Repeated Measures ANOVA (and paired t-tests for pairwise comparisons) for the statistical tests to perform the comparisons. The information presented in this item will be described in detail in the following subitems.

4.3.Research Approach

According to Kothari (2004) there are fundamentally two approaches to research: quantitative and qualitative research. A quantitative approach involves collecting data in quantitative form, usually subject to rigorous quantitative analysis using formal and well-defined procedures and instruments.

On the other hand, a qualitative approach deals with the subjective assessment of attitudes, views, and behavior. In this approach, the research is based on the researcher's observations and impressions, yielding either non-quantitative data or data not submitted to rigorous quantitative assessment. Examples of such techniques are focus groups, projective techniques, and in-depth interviews.

Research is usually quantitative or qualitative, depending on the type of data collected and the analyses performed. The researcher mainly uses the post-positivist paradigm to create a quantitative research approach. They study cause and effect relations between variables, reducing the problem and its analysis to variables, hypotheses, and precise questions, using measures and observation using hypothesis tests. Typically, the quantitative research strategy employs experiments and surveys to collect such data through predetermined instruments that provide statistical data (Creswell, 1998).

Quantitative research is usually appropriate when quantifiable measures of variables and inferences from population samples are possible. This type of research uses numerical measures to test scientific constructs and hypotheses or seeks numerical patterns related to everyday concepts. In contrast, qualitative research is mainly characterized by the absence of numerical measures and statistical analyses, examining deeper and more subjective aspects of the subject under study. (Dias, 1999).

Qualitative research denotes any research that produces conclusions without statistical procedures or other means of quantification (Strauss & Crobin, 1998 apud CRESWELL, 1998). The most common methods of collecting qualitative data are observation, interviews, documentary analysis, and audiovisual materials (Creswell, 1998).

Qualitative research is inductive, meaning that the researcher develops concepts, ideas, and understandings from patterns found in the data rather than collecting them to prove preconceived theories, hypotheses, and models. Qualitative methods generally use interpretative procedures, relativistic assumptions, and verbal data representation instead of a numerical representation (Sutton, 1993). Usually, qualitative methods are less structured; but provide a longer and more flexible relationship between the researcher and respondents and use more subjective, broader, and richer data than quantitative methods (Dias, 1999).

While there are benefits and many applications of qualitative methodologies, we employed a quantitative approach for the current research due to several factors. First, we had a well-defined, closed-end research question with an objective answer. In comparison, some of the recent literature in the field aims to establish whether telesimulation is feasible (O'Rae et al., 2021; Reece et al., 2021; Min Shao et al.,2018), based on the perception of participants or instructors. However, in this research, we proposed to use emotions and mental workload as objectively measured constructs as proxies for the effectiveness and feasibility of telesimulation.

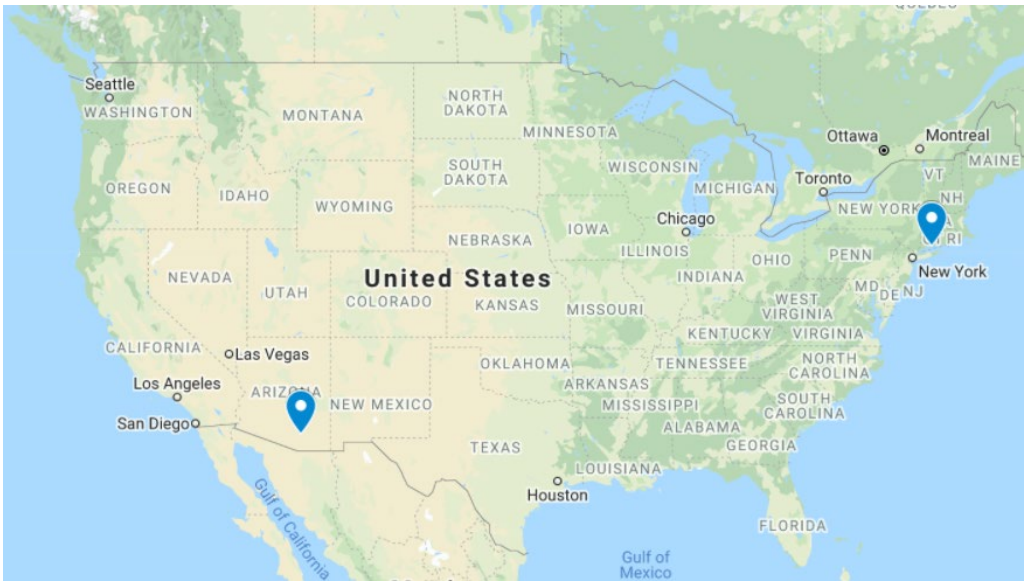
Another aspect that led us to define this approach was the availability of validated instruments to measure the research's constructs of interest (emotions, mental workload, and debriefing feedback). A further factor supporting using a quantitative approach was the ease and feasibility of applying surveys during the simulation and telesimulation compared to interviews and focus groups, which would be much more intrusive and time-intensive for participants.

Finally, the primary constructs of this research – emotions and mental workload – are rarely recalled reliably after a considerable amount of time after the events that evoke emotions or impose a cognitive workload (Chanel, 2009). Therefore, surveys were a viable approach to collect data shortly after (and during) the exposure to the treatments. Using approaches associated with qualitative analysis, like interviews, open-ended questions, or focal groups, could represent a long gap between the event and the data collection.

4.4.Research Location

The present research was conducted on two of the leading centers of medical education in the United States: Arizona Simulation Technology and Education Center (ASTEC) and the Yale Center for Medical Simulation (YCMS), which are located respectively in Tucson, Arizona, and New Haven, Connecticut. Figure **32** locates both centers in the United States map:

Figure 31: Location of ASTEC and YCMS in the Continental US



Source: Google Maps rendering by the author

4.4.1. ASTEC

The Arizona Simulation Technology & Education Center (ASTEC) is situated in the Health Sciences Innovation Building on the University of Arizona Health Sciences

(UAHS) campus. ASTEC is one of the country's most diverse medical simulation centers, with medical simulation training available throughout the University of Arizona Health Sciences Center. Medical flight crews, fire departments, military, and federal rescue employees, community outreach workers, and students in pre-college programs benefit from simulation training delivered by ASTEC (https://astec.arizona.edu/).

Figure 32: Technologies Available in ASTEC



Source: ASTEC – Retrieved on Dec, 17 - 2021 (https://astec.arizona.edu/)

Hands-on instruction comprises a wide variety of low to high-fidelity simulation technology that allows learners to achieve extensive real-time exposure to the anatomy and physiology of the human body and the symptoms of diseases, disorders, and dysfunction.

Figure 33: Simulation involving Robotic Mannekins in ASTEC



Source: ASTEC – Retrieved on Dec, 17 - 2021 (https://astec.arizona.edu/)

According to the official website (http://astec.arizona.edu) ASTEC is a 30,000-square-foot (approx. 2780 square meters) facility that offers interprofessional training in a high-tech, accurately reproduced setting. ASTEC uses high-fidelity simulation technology combined with unique experiential learning theory to engage learners at all levels of healthcare education.

The facility has a 6,000-square-foot (approx. 560 square meters) Sim Deck with six patient rooms, three of which can be simulated to be any type of hospital environment, and three rooms that mimic an operating room, intensive care unit, and labor and delivery suite. These rooms surround a 3,000 sq. ft., two-story multipurpose area that may be used for any large-scale training event, containing a range of procedural-based teaching stations and a simulated pharmacy.

The entire floor, including all human patient simulators and audio-visual controls for the learner management system, is controlled from a centralized command center. A synthetic cadaver and 3D immersive anatomy lab, two debriefing rooms, and a spacious, flexible classroom are also on the floor.

ASTEC also has a 2,000-square-foot (approx. 185 square meters) workshop that is an innovative place for the development of simulation technology. Fully equipped casting and molding technologies, high-end 3D printing/scanning for patient-specific anatomic models, and rapid fabrication of procedural models for training events are available. It also has space for virtual, augmented, and mixed reality technologies, motion analysis research, and electrical and mechanical engineering demands, such as optic technology testing and design.

4.4.2. YCMS

The other site used for the conduction of the research was the Yale Center for Medical Simulation Mission. According to the YCMS website (https://medicine.yale.edu/emergencymed/ycms/) their mission is to provide excellent patient care through innovative medical education, assessment, research, and outreach. The state-of-the-art center promotes the acquisition of basic and advanced clinical skills that are essential for medical students and residents to master and are integral to the educational mission of Yale School of Medicine.

The Center is available to all Yale School of Medicine community members for educational research opportunities. Yale Center for Medical Simulation uses mannequin-based, fully interactive simulators in emergency medicine, surgery, internal medicine, pediatrics, obstetrics and gynecology, neurology, anesthesia, and other clinical settings.

Simulation scenarios include critical actions in both clinical management and communication skills. Dedicated simulation faculty experts in scenario design and debriefing skills guide students through scenarios. Each simulation room allows for sophisticated video recording to enable participants to have the opportunity to reflect on their performance in an interactive debriefing session following the simulation scenario.

Figure 34: Medical Students



Source: https://medicine.yale.edu/emergencymed/ycms/education/

4.5.Ethical considerations and IRB evaluation

In the United States, an Institutional Review Board (IRB) is an administrative body established to protect the rights and welfare of human research subjects participating in research activities in each institution[11]. The website of the Human Subjects Protection Program (HSPP) of The University of Arizona states:

> *"The Human Subjects Protection Program (HSPP), as the administrative and regulatory support program to the Institutional Review Boards (IRBs), works in collaboration with the research community to maintain an ethical and compliant research program. The IRBs are the independent review committee charged with the protection of human research subjects. An IRB must review all research and related activities involving human subjects conducted at the University of Arizona or by in which the University is a responsible participant "Source: https://research.arizona.edu/compliance/human-subjects-protection-program*

The IRB is responsible for reviewing, prior to its initiation, all research (whether funded or not) involving human participants. Additionally, each IRB establishes criteria for conducting research, namely that the investigator has the appropriate training. According to the HSPP website (2021):

> *"All members of a research team are required to complete training to conduct human subjects research. Training is required prior to submission to the Institutional Review Board (IRB), before IRB approval will be granted, and/or before new personnel may begin work on a human research protocol." – UA HSPP (2021)*[12]

More specifically, UA Policies instructs researchers about the training that needs to be completed before the IRB submission:

[11] What is the Institutional Review Board (IRB)? https://research.oregonstate.edu/irb/frequently-asked-questions/what-institutional-review-board-irb

[12] Human Subjects Protection Program (HSPP) https://research.arizona.edu/research-compliance/training

> *"All study team members engaged in human subjects research are required to complete the University of Arizona human subjects component of the Collaborative Institutional Training Initiative (CITI) online training program … For social/behavioral projects, researchers must complete the CITI Social Behavioral module. Note that both the Social Behavioral and Biomedical courses also include the Native American Research module, which is required training for all researchers. All researchers are required to complete Conflict of Interest (COI) training.*" *UA HSPP* (2021)[13]

According to Common Rule definitions, the proposed study represents a minimal risk, a candidate for Exemption 1 of the 45 CFR 46.104(d). The (U.S.) Policy for the Protection of Human Subjects, 45 CFR part 46, also known as Common Rule, define the criteria for exemption as:

> *"Research, conducted in established or commonly accepted educational settings, that specifically involves normal educational practices that are not likely to adversely impact students' opportunity to learn required educational content or the assessment of educators who provide instruction. This includes most research on regular and special education instructional strategies, and research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods."*

The current research was conducted in an educational setting, involving standard educational practices, not likely to adversely affect (a) students' opportunity to learn required content or (b) teachers' assessment. Students were not interrupted during their participation in the simulation scenario or debriefing. Likewise, the content of scenarios was unaltered, and the moments to fill out the survey were chosen to minimize any interruption or change to the instruction format.

Participants were approached before participating in the simulation, responding to short surveys with non-identifiable data about the educational practice and associated

[13]Getting Started – HSPP https://research.arizona.edu/compliance/human-subjects-protection-program/getting-started

feelings. No sensitive information such as birthdate, medical conditions, medical histories, health attitudes, or behavior data was collected.

Likewise, during the simulation lab, the instructors were not required to change their practice, engage with the researcher, or be distracted from their role. Therefore, no negative impact of the research was expected in assessing the instructors or capacity to perform during the research.

Another requirement for the exempt status is that the research does not involve access to student education records under the Family Educational Rights and Privacy Act (FERPA). No record information such as individual student's performance written or electronic records, prior grades/courses/GPA/test scores, or student as tests and assignments were recorded, analyzed, or produced.

Likewise, the study did not plan to collect sensitive information about subjects that could potentially place them at risk if inadvertently disclosed outside the research. No record of illegal activities, genetic or medical information, sexual behaviors, or negative attitudes/opinions about an employer, colleagues, instructors, or the university were collected or attempted to be collected at any stage.

The informed consent was online, as the participant clicked the survey link implemented on the survey's first page before any data collection. The unsigned/online version of the consent form prevented any link between the participant's identities and their responses, further protecting their privacy and identification.

The researchers or instructors could not verify the completion of a survey by any given individual. We did not have a uniquely assigned code to each participant, e-mail, or other information to verify whether a person completed the survey or not. Also, the surveys were voluntary and not used for grading or evaluation.

Participants were informed about their right to withdraw from the study without repercussion. Since the data was collected during a short period (during a 30-minute simulation session), if a subject desired to be removed from the study, they could simply close their survey window and not continue the research. According to the established protocol, it was determined that if a participant left the survey before completing the phases of pre-simulation, post-simulation, and debriefing, data would be removed from the study responses.

The instructors did not know whether a participant consented to participate or the number of responses during a simulation session. Participation in the research was not a

condition for participation in the simulation scenario, nor the decision to participate was inquired about during the simulation session.

Before contacting the instructors and participants, the simulation center leadership (ASTEC and YMCS) was contacted to present research objectives, risks, and benefits and discuss any study issues.

The submission for UA-IRB was submitted electronically on July 29, 2020 to the Human Subjects Protection Program, and the authorization was granted on November 04, 2020 with the protocol number 2010177164.

Although the study carried minimal risks, the investigator monitored the safety parameters and events. There was no suggestion that the current research could place subjects or others at greater risk of harm (including physical, psychological, economic, or social). No adverse event was observed during the execution of the current research.

4.6.Research Instruments

Participants filled out standardized instruments to provide information about socio-demographic and educational background, emotions, appraisal of the emotional process and workload, perceptions of realism and fidelity, and feedback regarding the debriefing. We selected validated scales, widely employed in the simulation and emotion field, whenever available, with one scale adapted to understand the emotion appraisal (Scherer and Moors, 2019) process due to the specificity of the research.

In the final survey, items related to demographic and educational profiles were collected. Additionally, an emotion situation scale – Feldman, Barret, and Russell (2012), the Felt Arousal Scale (FAS) (Svebak & Murgatroyd, 1985), the NASA Task Load Index (NASA-TLX) – (Hart, 2006; Hoonakker et al., 2011), a custom scale about emotion Appraisal –based on Scherer and Moors' (2019) categorization of appraisal dimensions of emotion process and the Simulation Design Scale (SDS student version; Jeffries, 2016) in the subscales of Fidelity/Realism.

The final instrument distributed to participants had 67 items, most of them on a Likert-type scale, on three separate screens, counting all the research moments described earlier. The instrument is available in Appendix 1: **Research Instrument**. The following sections present the employed instruments in detail, including references to validation studies and some examples of their usage in healthcare simulation and other fields.

Participant information

To collect background information and the participants' educational profile, questions about their gender, age, race/ethnicity, class standing/professional profile, and prior exposure to simulation-based training. This information was collected to compare the different groups' emotional response and serve as a control for cofounding variables to enhance the study validity and comparison between the different samples.

Emotional State Scale

To measure participants' emotional states, we utilized a scale with bipolar opposing descriptions of emotions generated from Feldman-Barrett and Russell (2012) semantic structure of emotions to measure emotion states before, during, and after simulation sessions. The underlying theoretical support of this instrument is the Circumplex Model of Emotions (composed of two-axis a) valence - pleasant-unpleasant and b) arousal - activation and deactivation) (Young et al., 2021).

We employed Fraser & McLaughlin's (2018) implementation, in which 16 descriptions are combined into eight pairs of bipolar opposites and a five-point Likert scale. Participants were asked to score their feelings based on these descriptions with anchors at -2 and +2, denoting the bipolar opposites since emotions are often dichotomized into positive and negative states (Fraser et al. 2012).

We initially adopted the model proposed by Fraser and collaborators (2012), which classifies emotions in two components: tranquility and invigoration, as a theoretical and analytical framework. These two key emotion components were proposed during their factorial analysis of simulation training and accounted for 62% of the overall emotion variance. We tested their model in a confirmatory factorial analysis, and the results will be discussed in the respective section.

In this context, positive deactivated emotions are represented by tranquility, whereas positive activated emotions are represented by invigoration (Fraser et al. 2012). The vectors "invigoration" and "tranquility," according to Young et al. (2021), are the two emotional aspects most frequently observed in medical education research focused on emotions and workload to date. Table **26** contains the pair of emotions utilized:

Table 26: 8 Pair of emotions - Feldman-Barrett and Russell (2012)

Pair of emotions descriptions – Circumplex Model		**Factor (Feldman-Barrett and Russell, 2012))**
Unpleasant (Negative Activation and Deactivation)	**Pleasant (Positive Activation and Deactivation)**	
Bored	Alert	Invigoration (pleasant activation)
Lethargic	Excited	
Depressed	Elated	
Sad	Happy	
Upset	Contended	
Tense	Calm	Tranquility (pleasant deactivation)
Stressed	Serene	
Nervous	Relaxed	

Source: Compiled by the author based on Fraser & McLaughlin's (2018)

FAS - Felt Arousal Scale (FAS)

Participant activation (arousal) during this research was measured using the single item Felt Arousal Scale (FAS) (Svebak & Murgatroyd, 1985), in which participants responded to the following statement: "Estimate here how aroused you actually feel. By "arousal" we meant how "worked-up" you feel. The item was composed of a Likert-type question ranging from 1 (Low arousal) to 6 (High arousal). The FAS has been utilized in extensive prior research and has high convergent validity (Van Landuyt, Ekkekakis, Hall, & Petruzzello, 2000; Murray et al., 2016; Niedermeier et al., 2017).

This item served as a baseline of arousal in participants, pre and post-simulation control. Participants responded to this item before engaging in the simulation scenario, just after, and following the debriefing process.

Emotional Appraisal Scale

The emotional appraisal scale employed in this study contains 19 statements in a 7-point semantic differential scale. Although the study compiled items to suit the present research particularities, the sub-scales came from Ellsworth and Smith (1988) and related work, with the dimensions: Pleasantness, Goal Congruence, Certainty, Novelty, Predictability, Situational Control, Importance, and Fairness.

Other significant sources for validated items for this scale were: Ruth, Brunel, and Otnes (2002), Yanchus (2006), and Hosany (2012). The wording was maintained as close as possible to the original scales. We introduced minor changes to render the statements

specific to simulation to retain the validity of the instruments. The adaptation of this scale for specific contexts indicated that it maintains the instrument validity (Hosany, 2012)

Table 27: Emotional Appraisal Scale – Items, Sources, and Dimensions

Source for scale items	Appraisal Dimensions	Wording of the item
Ellsworth & Smith, (1988) Ruth, Brunel, andOtnes (2002) Yanchus (2006); Hosany (2012)	Pleasantness	How enjoyable was your overall experience?
		How pleasant or unpleasant was your overall experience?
Hosany (2012)	Goal congruence	To what extent was the overall experience consistent to what you wanted to achieve during this session?
		In comparison to what you desired, how consistent was your overall experience?
		To what extent do you see the overall experience in this session important or not important in achieving your learning goals or needs?
Ellsworth & Smith, (1988) Ruth, Brunel, andOtnes (2002) Yanchus (2006); Hosany (2012)	Certainty	How well did you understand what was happening around you?
		How sure were you about what was happening?
Hosany (2012)	Novelty	To what extent did the overall experience was something you expected or unexpected?
		How usual or unusual was your overall experience?
		To what extent was the overall experience you had in the simulation session familiar or unfamiliar?
Ellsworth & Smith, (1988)	Predictability	How well could you predict what was going to happen?
Yanchus(2006) Ruth, Brunel, and Otnes (2002)	Situational agency (Other)	To what extent did you feel that circumstances beyond anyone's control determined the outcomes/results of this simulation? (Situational responsibility)
		How responsible did you think someone or something other than yourself was for the outcomes/results of the simulation session? (Situational Control)
Yanchus (2006); Ellsworth & Smith, (1988)	Self-agency	When you were in this session, how powerful (dominant) did you feel? (power)
Yanchus(2006)		When you were in this simulation session, how helpless did you feel? (helplessness)
Ellsworth & Smith, (1988) Ruth, Brunel, and Otnes (2002) Yanchus (2006)		To what extent did you feel that you could influence what was happening in the simulation session? (Self-control)
Ellsworth & Smith, (1988) Yanchus (2006)	Importance	How important was this simulation session to you?
Ellsworth & Smith, (1988) Ruth, Brunel, and Otnes (2002) Yanchus (2006)	Fairness	To what extent did you feel cheated or wronged in this simulation session?
		How fair did you think what happened to you in this simulation session was?

Source: Compiled by the Author

NASA Task Load Index (NASA-TLX)

We selected the widely employed NASA-TLX (Task Load Index) instrument to measure cognitive workload during healthcare simulations (Hart, & Staveland, 1988; Hart, 2006; Yurko et al., 2010)). Initially, the instrument resulted from a multi-year study to determine the elements that influence perceived workload fluctuations within and across different activities (Hart, & Staveland, 1988).

The multi-dimensional rating scale was developed to collect data on the size and sources of six workload-related elements, integrated to create an assessment of the multidimensional workload. The NASA – TLX is one of the most extensively used instruments for assessing total subjective workload (Honakker, 2011).

This instrument has been employed in various fields, including aviation, healthcare, and other complicated socio-technical areas (Colligan et al., 2015; Yurko et al., 2010). This scale includes Mental, Physical, and Temporal Demands, Frustration, Effort, and Performance. The scale was implemented with a 7-point scale with endpoints "Very Low" to "Very High" or "Good" to "Poor".

According to Said et al. (2020) NASA's TLX is a multidimensional instrument with six dimensions that can be further integrated into a second-order dimension. Three items measure the different demands placed on the subject (mental, physical, and temporal). In comparison, the other three items track how the subject responds to the imposed task (self-rated performance, effort, and frustration level). The dimensions, midpoints, and descriptions are presented below:

Table 28: NASA-TLX – Dimension, endpoints, and descriptions

Dimension	Endpoints	Description	Second order dimensions
Mental Demand	Very Low to Very High	How mentally demanding was the simulation?	Demand on the subject
Physical Demand	Very Low to Very High	How physically demanding was the task?	
Temporal Demand	Very Low to Very High	How hurried or rushed was the pace of the simulation?	
Performance	Good to Poor	How successful were you in accomplishing what you were asked to do?	Subject response to the task
Effort	Very Low to Very High	How hard did you have to work to accomplish your level of performance?	
Frustration	Very Low to Very High	How insecure, discouraged, irritated, stressed, and annoyed were you?	

Source: Adapted from NASA-TLX (Hart, & Staveland, 1988) and Said et al. (2020)

We employed the Raw NASA-TLX approach. According to Hart (2006), the Raw TLX (RTLX) is the most popular modification, in which the weighting process between the different dimensions is completely bypassed. The RTLX is appealing for obvious reasons: estimating the overall workload is as simple as adding the sub-scale results. In other words, no other calculations than a simple sum are required.

Considering the lack of physical demand for telesimulation, we also computed an alternative version of the "Demand on the Subject" second-order dimension, removing the physical component task demand to compare the two methods.

This decision supported Virtanen et al. (2021) findings that the physical demand dimension (PD) was the lowest rated in a study investigating the weighting process with a sample of 20 qualified F/A-18 fighter pilots. They found that the relative weight of PD was 0.01 when using the NASA-TLX weighting process, being the least important demand, in a similar setting, while the mental demand weighted 0.27, therefore deemed much more important.

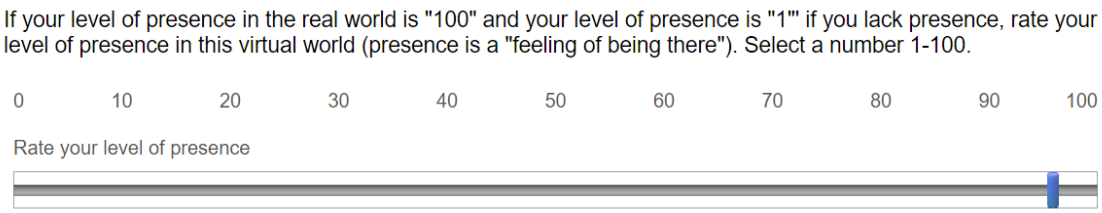
Fidelity/Realism

We employed two validated subscales to measure fidelity (realism) during the simulation session. This subscale is part of the Simulation Design Scale – Student Version (Jeffries, 2016) with a 2-item, 5-point Likert-type agreement scale. The items were "The scenario resembled a real-life situation," and "Real-life factors, situation, and variables were built into the simulation scenario" with a 6-point Likert-type scale, ranging from "Strongly disagree" to "Strongly agree".

Sense of Presence

We employed a single item measurement from Dinh's Concept of Presence (COP) tool (Dinh, Walker, Song, Kobayashi, & Hodges, 1999). This combination (fidelity and sense of presence scales) was proposed and validated by MacLean et al. (2019) in a similar setting. We implemented this item in Qualtrics as a "Slider" control with values ranging from 1 (no presence) to 100 ("feeling of being there"). The implemented control is available in Figure **34**:

Figure 35: Concept of Presence Tool – Qualtrics



Source: Qualtrics Survey

Feedback/Guided Reflection

To measure the feedback after the simulation session, we employed the Simulation Design Scale (SDS student version; Jeffries, 2016) in the feedback dimension. The SDS feedback dimension is a 4-item instrument with a 5-point agreement scale to evaluate the Feedback/Guided Reflection dimension. The items were "Feedback provided has constructive", "Feedback was provided in a timely manner", "The simulation allowed me to analyze my own behavior and actions" and "There was an opportunity after the simulation to obtain guidance/feedback from the teacher in order to build knowledge to another level", with a 6-point Likert-type scale, ranging from "Strongly disagree" to "Strongly agree".

4.7. Research Procedures

Data collection for this study was conducted in simulation sessions that usually last approximately 45 minutes in remotely facilitated or in-person formats. In the simulation-based education approach, typically, students participate in a briefing session in which they discuss central concepts and previously distributed reading material before participating in a simulation session. Figure **35** illustrates a briefing session before the start of an in-person simulation:

Figure 36: Briefing Session before the start of a simulation scenario

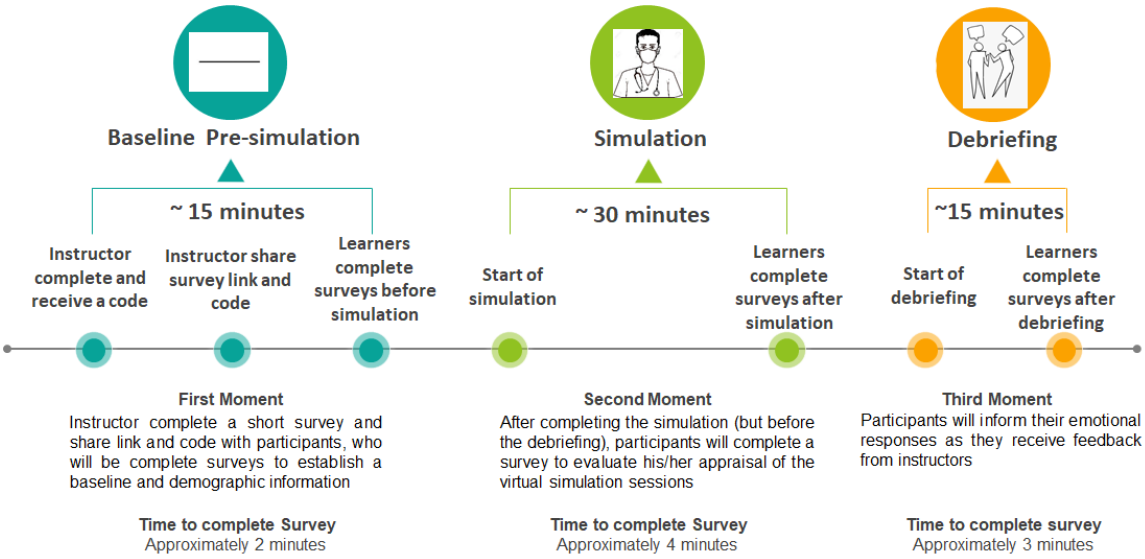


Source: Washington Square News (2017) - https://nyunews.com/2017/11/27/nursing-students-feel-neglected/

The present research has three key temporal points for data collection during simulation sessions (in-person and telesimulation). The three temporal points design was based on Fraser & McLaughlin (2018) – and prior work - which investigated medical students' temporal patterns during simulation training and debriefing using scales to measure cognitive load and emotional response. Furthermore, the division of simulation classes into three temporal points is widely prevalent in simulation research (Fraser & McLaughlin, 2018; Lavoie et al., 2019; Pawar et al., 2018)

The times shown in the Figure below are for the simulation, not the time necessary to complete the surveys. The timeline of a typical simulation session and the corresponding data collection are summarized in Figure **37**:

Figure 37: Research Moments and associated activities



Source: Compiled by the author

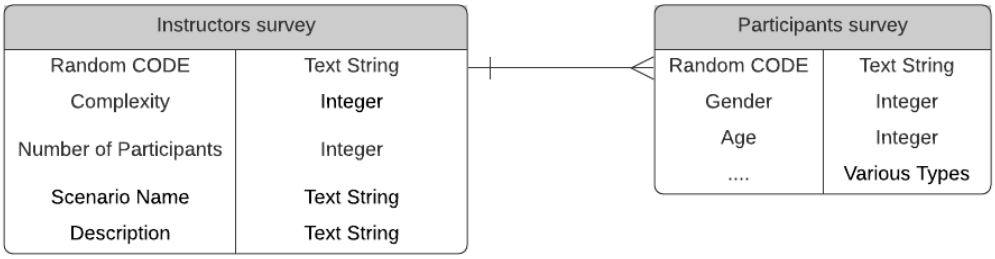
Each research moment and the respective details are presented in the following items in sequence:

Baseline-presimulation (Moment 1)

Before each simulation session (in-person and telesimulation), instructors were informed about the research goals, voluntary nature, and a link to be filled with meta-information about the simulation scenario they would conduct. The data collected was distributed in four questions: an anonymous instructor's perception of the simulation scenario complexity (low, medium, or high complexity), the number of participants expected for that session, a simple description of the scenario, and an automatically generated location marker corresponding to the U.S. State where the simulation center was located.

After completing the instructor survey, Qualtrics generated an automated template with a description of the research, including a random code with two letters from the US state and a number (from 1 to 9999 – e.g., AZ1234). This code collected information about the simulation scenario and linked responses from individual and session-level surveys. Figure 38 presents the linkage between the instructor's survey and the participants' survey:

Figure 38: Relationship between the instructor's survey and the participants' survey



Source: Compiled by the author

The automatic text was available to be copied by the instructors. It could be shared by e-mail, chat, or in the telesimulation platform according to the instructor's preferences, with the following wording:

Table 29: Sample invitation text automatically generated by Qualtrics

"Dear Participant, This is an invitation to participate in an online research. This study aims to understand Emotional Responses in Virtual Healthcare Simulation. You will be presented with information relevant to your experience during virtual healthcare simulation and asked to answer some questions about it. Please be assured that your responses will be kept completely confidential. The study should take you around 10 minutes to complete. Your participation in this research is voluntary. You have the right to withdraw at any point during the study, for any reason, and without any prejudice. There are no follow up questions, interviews or any contact information required for your participation. An Institutional Review Board responsible for human subjects research at The University of Arizona reviewed this research project and found it to be acceptable, according to applicable state and federal regulations and University policies designed to protect the rights and welfare of participants in research. If you would like to contact the Principal Investigator in the study to discuss this research, please e-mail Gustavo Almeida gustavo@arizona.edu postdoctoral associate, advised by Professor Win Burleson win@arizona.edu. The purpose of this research is to understand learners' emotional responses during virtual simulation sessions, using questionnaires to assess emotions, perception of task difficulty and presence. This questionnaire, should you want to participate, will be administered to you today in three parts, with a total estimated time of about 10 minutes:

- **First moment- To be completed shortly before the simulation (Approx. 2 minutes)**
- **Second part - To be completed after the simulation, but before the debriefing (Approx. 4 minutes)**
- **Third part - To be completed after the debriefing session (Approx. 3 minutes)**

If you would like to participate, please go to this link:

FINAL_URL_LINK ---> and type the CODE: AZ779 [**This code is automatically and randomly generated and will be different for each survey]**

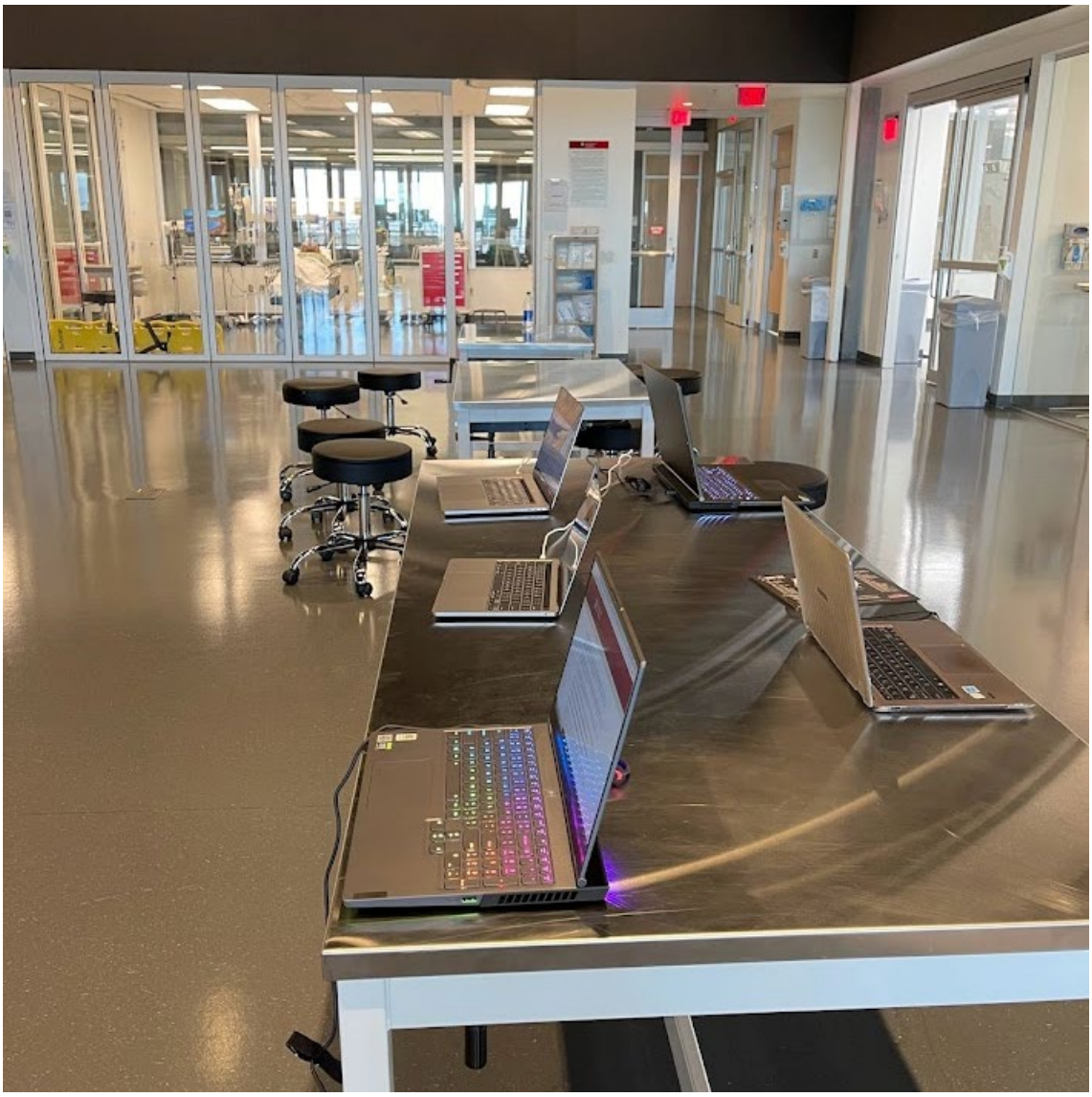
Very Best,

Gustavo de Oliveira Almeida
gustavo@arizona.edu

For the in-person simulation, the procedure was similar. The only difference was that the laptop computers were ready for use. The Qualtrics survey was open before the simulation session to minimize interference in the learning procedures and optimize time.

Likewise, in-person participants were informed about the research, its voluntary nature, and their rights, along with the exact text available to the participants in telesimulation explaining the research in more detail. Figure **38** displays a sample setup created for data collection during the in-person simulation.

Figure 39: Sample setup for in-person data collection



For both in-person and telesimulation, the questionnaires were available only in an electronic format as surveys hosted in UA Qualtrics servers in a secure environment. In both cases, the simulation center leadership determined the selection of participants according to the class schedule and the instructor's consent. The researcher did not determine this selection process or allocation to either form of instruction, classifying the current research as a natural experiment.

After showing up for the simulation session (or joining the virtual room), participants were approached just before the simulation and invited to participate in the research. In an affirmative case, subjects responded to a short demographic survey with non-identifiable data, read and completed the consent form, and responded to questionnaires to furnish a baseline for their emotional state.

The first part of the questionnaire had questions regarding gender, current year/program, prior experience in health care, in-person and telesimulation, and a field to add the code shared by the instructor. In addition, participants filled out two specific scales for establishing emotional response baselines:

- Emotion Scale – Feldman, Barrett, and Russell (2012)
- FAS - Felt Arousal Scale (FAS) (Svebak & Murgatroyd, 1985)

Immediately after filling out the consent form and the questionnaires, participants started participating in the simulation session conducted by the simulation center instructor. There was no change to their practice and no interference in the educational content of the simulation.

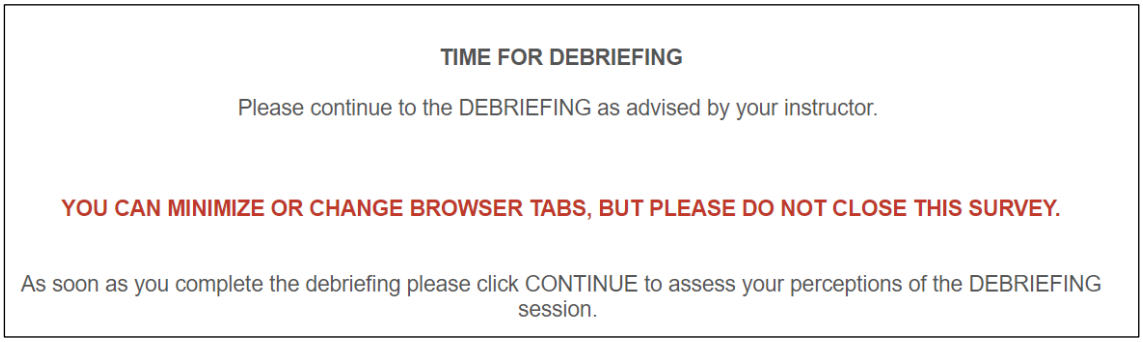
Simulation (Moment 2)

As participants started the simulation sessions (either telesimulation or in-person), no additional engagement with the research were required from them. After completing the simulation sessions, participants returned to the questionnaire to respond to their appraisal of the simulation session. The specific questionnaires administered at this point were:

- Emotion Scale – Feldman, Barret, and Russell (2012)
- FAS - Felt Arousal Scale (FAS) (Svebak & Murgatroyd, 1985)
- NASA Task Load Index (NASA-TLX) – (Hart, 2006; Hoonakker et al., 2011)
- Emotion Appraisal – custom scale primarily based on Scherer and Moors' (2019) categorization of appraisal dimensions of the emotion process.
- Fidelity/Realism dimension of Simulation Design Scale (SDS student version; Jeffries, 2016)

The estimated time to complete all scales at this moment (after the simulation) is 3 minutes. Then, after completing the mentioned scales, subjects received a notice that the debriefing section would be starting and they should wait until completing the next part of the survey, as illustrated in Figure 39:

Figure 40: Prompt for user to engage in the debriefing



Source: Qualtrics Survey

Debriefing (Moment 3)

The debriefing section of a simulation is considered a critical component of the learning process (Lavoie, 2019). Subjects provided their perceptions after receiving feedback from instructors and fellow learners. Immediately after the completion of the debriefing session, participants completed the following instruments:

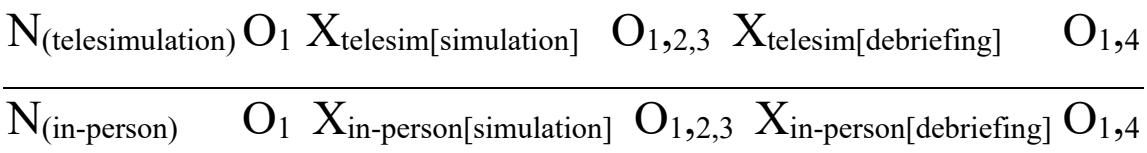
- Emotion Scale – Feldman, Barrett and Russell (2012)
- FAS - Felt Arousal Scale (FAS) (Svebak & Murgatroyd, 1985)
- Feedback dimension of Simulation Design Scale (SDS student version; Jeffries, 2016).

The estimated time to complete the four administered instruments at this moment (after debriefing) was 3 minutes. The total time for the learner's engagement for the three moments was about 10 minutes.

Experiment Notation and summary of moments and associated scales

We have modeled the current research using experimental notation to facilitate the understanding of the research procedures, the moments of research, and associated scales according to the quasi-experimental (natural experiment) design:

Figure 41: Design of the Natural Experiment



N – Natural experiment

O – Observation or measurement

X - Refers to experimental manipulation, intervention, or treatment. In our case, it can refer to the in-person simulation and telesimulation

O_1 – emotional response measurement (ESS and FAS Scale)- Pretest/posttest design

O_2 –NASA-TLX mental workload – (post-test – after simulation)

O_3 – Fidelity/Realism Dimension – (post-test – after simulation)

O_4 - Simulation Design Scale (can only be applied as a post-test – after simulation)

Table 30 presents in more detail the scales employed, including a brief description and their usage in the corresponding moment: pre simulation (1), simulation (2), and debriefing (3). Table 30 applies identically for telesimulation and in-person groups:

Table 30: Scales, Description and Usage in Pre Simulation (1), Simulation (2), and Debriefing (3)

Scales	Description	Moment		
		1	2	3
Participant Information	Participant information regarding educational profile, gender, age, race/ethnicity, class standing/professional profile, prior exposure to simulation-based training among others items	**X**		
Emotion Scale	Eight item scale - semantic differential between pairs of emotions based on Feldman, Barrett and Russell (2012) model	**X**	**X**	**X**
FAS - Felt Arousal Scale (FAS)	Single item scale – 6 point Likert – ranging from "low arousal" to "high arousal" - (Svebak & Murgatroyd, 1985)	**X**	**X**	**X**
NASA Task Load Index (NASA-TLX)	Six item scale – 7 points Likert scale used for workload measurement composed by the dimensions: Mental Demand, Physical Demand, Temporal Demand, Overall Performance, Effort, Frustration Level (Hart, 2006; Hoonakker et al., 2011)		**X**	
Emotion Appraisal Process	Custom scale largely based on Scherer and Moors' (2019) categorization of appraisal dimensions of emotion process. The scale contains 19 statements in the form of a 7 point semantic differential scale. Sub-scales were adapted from Ellsworth and Smith (1988) and related work, with the dimensions: Pleasantness, Goal Congruence, Certainty, Novelty, Predictability, Situational Control, Importance and Fairness.		**X**	
Fidelity/Realism dimension (SDS)	Simulation Design Scale (SDS student version; Jeffries, 2016) with a 2-item, 5-point agreement scale combined with a single from Dinh's Concept of Presence (COP) tool (Dinh, Walker, Song, Kobayashi, & Hodges, 1999). This combination was proposed and validated by MacLean et al. (2019) in a similar setting.		**X**	
Feedback dimension (SDS)	Simulation Design Scale (SDS student version; Jeffries, 2016). The SDS feedback dimension is a 4-item instrument with a 5-point agreement scale to evaluate the Feedback/Guided Reflection dimension.			**X**

Source: Compiled by the author

4.7.1. Participants and Selection Procedure

The present research did not intend to actively triage or recruit students with a specific trait, condition, or background before being invited to participate in the research. They were participants scheduled to participate in the simulation sessions. Participants were recruited briefly before the simulation session and were selected during the research *in locus*, meeting the definition of a natural experiment. Once participants were referred to the research, the inclusion criteria were as follows:

Inclusion Criteria - Any participant in an in-person or telesimulation session conducted in ASTEC or YMCS, including undergraduate or graduate students (medical students, for instance), nurses, medical or nursing assistants, paramedics, physicians, and other healthcare professionals. Participants needed to be 18 or older.

Exclusion criteria - Anyone not presently participating in an in-person or telesimulation could not participate in the study since the instruments needed were completed during the simulation session.

Therefore, our target population was over 18 years, students/professionals participating in healthcare simulation sessions (either in-person or telesimulation). All genders were part of the research population. The research was not limited or specific to any single ethnicity, nor were those variables hypothesized to cause differences in responses of the variables of interest.

4.7.2. Software for Data Analysis

For the data analysis, we employed various software, according to their specific functions and capabilities, in five packages: SPSS, NCSS, JASP, PASS, and SmartPLS. We employed the commercial statistical software SPSS (Armonk, NY: IBM Corp), version 27, to produce: i) descriptive statistics; analysis of variance, inter and intragroup ii) factor analysis ii) missing value analysis; vi) normality analysis iv). All two-tailed analyses involved a significance level of 5%.

We employed the commercial software Number Cruncher Statistical Systems (NCSS) 2021 to perform two-sample t-tests for non-inferiority to produce graphics and exploratory data analysis. To calculate the sample size and associated power, we used the

PASS 2021 software to calculate the n and power for t-tests employed in the study. To conduct confirmatory factor (CFA) analysis, we opted to use SmartPLS 3.3.7 (SmartPLS GmbH) to produce models and statistical tests related to reliability, convergent and divergent validity of the study's constructs.

Finally, we used JASP, version 0.1.16, free and open-source software for statistical analysis developed at the University of Amsterdam, Netherlands. JASP was used to conduct ANOVA Repeat-Measures, Independent Samples t-test, and other graphics.

4.7.3. Statistical Power Analysis

This section will present a power analysis conducted before the data collection to estimate feasible sample sizes and a post hoc analysis with the number of samples effectively collected and their associated power to detect effects and reject the hypothesis.

Several factors determine the size of a statistical sample. The parameters essentially differ in terms of the statistical tests performed. However, it is possible to group minimum sample size recommendations into two categories: one that considers statistical test error and power, and the other that considers a minimum number of observations for a specific test.

According to Rudestam & Newton (2007), the best technique to determine the number of participants necessary is to undertake a power analysis. A power analysis informs the researcher about the number of subjects needed to detect any effects resulting from the independent variables, considering (a) the size of the effect of these variables in the population, (b) the statistical tests to be used, and (c) the study's level of significance (or alpha level).

Statistical power is the ability of a test to detect an effect in a sample based on its sensitivity or power. The amount of error present limits statistical power, and there are two types of error: type I and type II. When the null hypothesis (h0) is rejected despite being true, it implies an effect or relationship when there is not one. A false positive is a term for this situation. The significance level for basic hypothesis testing is also known as a type I error, and the letter α is often used.

A Type II error arises when the null hypothesis is not rejected, even though false. Failure to reject a false null hypothesis indicates that an effect exists but was missed by the investigation, leading to a false negative. Type II error is often represented by the

letter, and it is proportional to the statistical power of a test, which is 1 – β. Table 31 presents the interrelationship between the errors:

Table 31: Types of Error: Type I, II, and hypothesis rejection

Decision	**H0 is true**	**H0 is False**
Rejects null hypothesis (H0)	Type I Error	Correct (H0 false)
Null hypothesis is not rejected (H0)	Correct (H0 true)	Type II error

Typically, a type I error of 5% is considered acceptable. This level signifies that the null hypothesis will be rejected five times out of 100 times when it was not rejected. For the error II (β) the value of 0.20 indicates that a 20% possibility of type II error is acceptable, a test's power must be at least 80% (1 - 0, 2), admitting that a true effect may be missed in 20% of cases (Rudestam & Newton, 2007). Therefore, unless specified otherwise, we used a significance level of 5% and power of 80% for tests in this research.

We present three tests performed to calculate the minimum sample sizes in the sequence, assuming equal group sizes and normally distributed data.

We employed the software PASS version 2021 for sample size and power analysis. We used the Equal Variance T-test since it was one of the main tests proposed to be used, and a non-inferiority test Mann-Whitney U or Wilcoxon Rank-Sum assuming normal distributions. The third was a power test to detect different effect sizes using a one-sample t-test.

The first power analysis, solving for sample size (with varying degrees of power) for an Equal-Variance T-test, is displayed in Table 32:

Table 32: Sample Power calculation: Numeric Results for an Equal-Variance T-Test

Hypotheses:	H0: $\delta = 0$ vs. H1: $\delta \neq 0$							
Target	Actual							
Power	Power	N1	N2	N	μ1	μ2	Std dev For both groups	Alpha
0.8	0.80704	17	17	34	1	0	1	0.05
0.85	0.85061	19	19	38	1	0	1	0.05
0.9	0.9125	23	23	46	1	0	1	0.05
0.95	0.95008	27	27	54	1	0	1	0.05

According to the PASS software output, group sample sizes of 17 and 17 achieve 80.704% power to reject the null hypothesis of equal means when the population means difference is μ1 - μ2 = 1 - 0 = 1 with a standard deviation for both groups of 1 and with a significance level (alpha) of 0.05 using a two-sided two-sample equal-variance t-test.

On the other end of the power analysis (with a larger sample), sample sizes of 27 for each group achieve 95.008% power to reject the null hypothesis of equal means when the population means the difference is μ1 - μ2 = 1 - 0 = 1 with a standard deviation for both groups of 1 and with a significance level (alpha) of 0.05 using a two-sided two-sample equal-variance t-test.

In the sequence, we have performed a power analysis of Mann-Whitney U or Wilcoxon Rank-Sum, solving for sample size, using the default power options: 0.8, alpha:0.05, equal group sizes, standard deviation=1, actual difference=0, and non-inferiority margin=1.

Group sample sizes of 17 for each group achieve 81% power to detect non-inferiority using a one-sided Mann-Whitney U or Wilcoxon Rank-Sum test, assuming the actual data distribution is normal. The margin of non-inferiority is -1. The actual difference between the means is assumed to be 0. The significance level (alpha) of the test is 0.025. The parameters were populations with a standard deviation of 1 in both groups. The results are displayed in Table 33:

Table 33: Sample analysis requirement – non-inferiority tests

Hypotheses: H0: $\delta \leq$ -NIM vs. H1: $\delta >$ -NIM

Target	Actual							
Power	Power	N1	N2	N	-NIM	δ	σ	Alpha
0.8	0.80704	18	18	36	-1	0	1	0.025

Note: Higher Means are Better - $\delta = \mu 1 - \mu 2 = \mu T - \mu R$

According to PASS 2021 output, a group sample sizes of 17 for each group achieve 81% power to detect non-inferiority using a one-sided Mann-Whitney U or Wilcoxon Rank-Sum test, assuming that the actual data distribution is normal. The margin of non-inferiority is -1. The actual difference between the means is assumed to be 0. The significance level (alpha) of the test is 0.025. The data were inputted for populations with a standard deviation of 1 in both groups.

Therefore, we performed a power analysis of a paired t-test solving for sample size (total), using as parameters power level =0.8 and effect sizes of 0.2 (small), 0.5 (medium), and 0.8 (large). The result of the analysis is displayed in Table 34:

Table 34: Sample analysis requirement – paired t-test sample vs effect size

Target	Actual		Effect Size	
Power	Power	N	D	Alpha
0.8	0.80017	156	0.2	0.05
0.8	0.81183	27	0.5	0.05
0.8	0.82898	12	0.8	0.05

The PASS 2021 software output indicates that a sample size of 156 data pairs achieves 80% power to reject the null hypothesis of zero effect size when the population effect size is 0.2, and the significance level (alpha) is 0.05 using a one-sided paired t-test. Conversely, a sample of 12 observations can detect a large effect (0.8) with a 0.82 power. Therefore, the larger the effect size, the smaller the sample required.

Post Hoc Analysis

After the data collection, we performed post hoc data analysis to calculate the actual statistical power of this study, using the same tests. However, we inputted the sample size and solved for adequate statistical power after the data collection instead of solving for sample size. We inputted information extracted from the data, including

means, standard deviation, and associated distributional parameters. We used the total number of useable and complete responses (n=134), accounting for dropout rates and incomplete data. The results of the analysis are in Table 35:

Table 35: Post Hoc statistical power– one-sided t-test sample

– solve for power (n=134)

Power	N	d	Alpha
0.7449	134	0.2	0.05
0.99998	134	0.5	0.05
1	134	0.8	0.05

The one-sided t-test sample power test executed in PASS indicated that with a sample of 134, we have 0.75 power to detect a small effect and 0.99 power to detect an effect of a medium magnitude of 0.5.

For the paired t-test, we have used data from the observations, namely the FAS Arousal Scale paired difference between baseline and simulation with a mean value of 0.81 and a standard deviation of 1.45. We have solved for power, with the calculated effect being 0.5586 (medium effect size). Results are available in Table 36:

Table 36: Post-Hoc statistical power– paired t-test sample

				Effect		
Power	N	δ1	σ	Size	Alpha	Beta
1	134	0.81	1.45	0.55862	0.05	0

According to PASS 2021 output, a sample size of 134 achieves 100% power to detect a mean of paired differences of 0.81 with an estimated standard deviation of paired differences of 1.45 and with a significance level (alpha) of 0.05 using a one-sided paired t-test. Finally, we analyzed a non-inferiority test assuming equal variances, using the NASA-TLX RAW score as a parameter, one of the central constructs and the most widely used instrument in this research, using the descriptive statistics to populate the procedure in PASS.

The analysis indicated that our group sample sizes of 62 (n-telesimulation) and 72 (n=in=person) achieved 84% power to detect non-inferiority using a one-sided, two-sample unequal-variance t-test. The margin of non-inferiority is -2.768. The actual difference between the means is assumed to be 0. The significance level (alpha) of the

test is 0.05. The data are drawn from populations with standard deviations of 5.65 for group 1 and 6.52 for group 2.

After the initial power analysis and post hoc analysis with the study data, we can infer that the final sample size is sufficient for the analysis. The collected sample provides enough power to detect small (<0.2) and medium effect sizes (<0.5) in most instances and nearly 100% power to detect large effects.

Chapter 5. Data Analysis

The data analysis result chapter presents the results of the empirical analysis of this thesis. It is divided into two sub-items, exploratory data analysis and hypothesis testing. The exploratory data section is composed of the following items: sample characterization, exploratory data analysis, missing data analysis, common method bias, analysis of outliers, data normality tests, confirmatory factor analysis, including scale validity and reliability tests.

The section will describe summary statistics of the variables for participants in telesimulation and in-person simulation groups. Then, we will demonstrate scale reliability tests and exploratory factor analysis to confirm that the scales used in this study are reliable and apt to measure the constructs studied.

In the sequence, we will perform the hypothesis tests to examine the effect of telesimulation and in-person simulation on arousal, mental workload, and emotional responses. The hypothesis testing was implemented by the Two samples t-test test (parametric and non-parametric), paired-t, and ANOVA Repeated Measures.

5.1.Sample Description

This section will describe our sample in terms of their demographic information and additional variables related to participants' experience in simulation, scenario complexity, team sizes, and collection sites.

Finally, we compared the characteristics of our two main groups: in-person and telesimulation, to ensure the similarity of the two groups. The comparison at this stage was performed to support the effects of either approach on the emotional and cognitive responses, even though our samples were not randomly drawn.

Our final sample consisted of 134 participants, predominantly Males (n=73); 59 participants self-identified as Females, while two selected their gender as “Other”. Regarding ethnicity and race, the most reported response was White/Caucasian by 78 respondents (48.1%), followed by 30 (18.5%) Asian/Pacific Islanders, 25 (15.4%) Hispanic and Latinos, and 14 reported “Other option,” six declined to state, five reported being black. Four self-identified as American Indian or Alaskan Native.

Participants age was more concentrated into the 26-31 years old group (n=80 ; 60.2%), followed by the group from 22-25 (n=31 ; 23.3 %) and 32-36 (n=10; 7.5%). There were 5 participants in the range of 35-40, while seven reported being over 40. No participant reported being younger than 21 in our sample.

In terms of prior experience with simulation (in-person), most participants had between 3-4 semesters (n=53; 39.6%), followed by those who had less than one semester of experience (n=34; 25.4%), other sizeable groups had between one and two semesters (n=27; 20.1%). Finally, only 9 participants had more than six semesters of experience with traditional simulation.

Participants were 109 graduate students (Master/DNP/PhD/MD), 10 Nurses, Medical Doctors, or other healthcare professionals (not enrolled as a student), and seven were residents/fellows. Regarding the site of conduction and collection of data, most responses were collected at the University of Arizona (n=121; 90.3%), while the Yale Center for Medical Simulation sample was 13 (9.7%).

A total of 33 simulation sessions were part of this study, with the mean number of participants for each session being four. The minimum for a session was one participant, while the maximum for one session was nine.

The average time of the simulation sessions from the start of the data collection to the end was 46 minutes (SD = 16.804), divided into low complexity (n=67) and medium complexity scenarios (n=59) as defined by the simulation instructors.

In the sequence, we present a list of the simulation sessions grouped by the complexity level in Table 37, with information regarding the type of simulation, the scenario code (random code shared with the participants during the questionnaire), scenario name, and a short description of the scenario, provided by the instructor.

Table 37: List of simulation scenarios and the associated meta-information

Instructor Rated Complexity / Difficulty	Type of Simulation	Random Scenario Code	Simulation scenario label/ name	Description
Low complexity (low difficulty)	Telesimulation	AZ1074	Asyphixiation	Difficulty breathing
	In-Person	AZ2543	Allergic Reaction	Patient comes in with shortness of breath
	In-Person	AZ3272	Allergic Reaction	Patient comes into the hospital with trouble breathing
	Telesimulation	AZ383	COPD	The patient has trouble breathing with prior history of COPD
	In-Person	AZ4210	Allergic Reaction	Patient comes in with trouble breathing
	Telesimulation	AZ4610	COPD	Patient comes into Emergency Room with shortness of breath
	In-Person	AZ5148	Allergic Reaction	Patient comes in with trouble breathing
	Telesimulation	AZ5260	COPD	Patient comes in with difficulty breathing
	In-Person	AZ5445	Allergic Reaction	Patient comes into the hospital after a bees sting
	Telesimulation	AZ5533	COPD	Patient comes into the emergency room with trouble breathing
	Telesimulation	AZ5705	COPD	Patient enters Emergency Room with trouble breathing
	Telesimulation	AZ7528	Allergic Reaction	Patient comes into the emergency room with trouble breathing
	Telesimulation	AZ782	GI Bleed PEA	Unresponsive patient enters the Emergency Room for treatment
	In-Person	AZ8942	Anaphylaxis	Patient comes in with difficulty breathing
	Telesimulation	AZ8994	COPD	Patient comes in with trouble breathing
Medium complexity (medium difficulty)	Telesimulation	CT2426	Endocarditis with valve rupture	This is a case of a patient with endocarditis who develops acute heart failure after a valve rupture.
	Telesimulation	AZ1104	EMCC GI Bleed	Unconscious patient comes into the Emergency room with blood in throat
	Telesimulation	AZ115	GI Bleed PEA	Patient comes into the Emergency Room unresponsive

	In-Person	AZ3507	STEMI	Unconscious patient enters the hospital
	In-Person	AZ3594	STEMI	Patient comes into ER with chest pain and shortness of breath
	In-Person	AZ3802	STEMI	Patient is in the hospital for knee replacement and complains of chest pain
	Telesimulation	AZ4407	GI Bleed	Unconscious patient enters Emergency Room with dropping vitals
	In-Person	AZ4589	STEMI	Patient is awaiting surgery while complaining of chest pain
	In-Person	AZ6291	STEMI	Patient has chest pain
	Telesimulation	AZ7091	GI Bleed	Patient comes in to the Emergency room unresponsive and dropping vitals
	In-Person	AZ7963	STEMI	Patient is getting ready for surgery and complains of chest pain
	In-Person	AZ999	STEMI	Patient is awaiting surgery with chest pain
	Telesimulation	CT201	Endocarditis with valve rupture	A valve rupture secondary to endocarditis.
	Telesimulation	CT2169	Endocarditis with valve rupture	A patient develops cardiogenic shock secondary to a valve rupture in the setting of endocarditis.
	Telesimulation	CT6851	Endocarditis with valve rupture	Adult patient presents with heart failure and is found to have endocarditis.
	Telesimulation	CT8208	Hypovolemic shock	A young female with a history of drug use disorder is in a motor vehicle accident and presents with abdominal pain. Initially stable, the patient deteriorates and is found to have free fluid in the abdomen, becomes unstable, requires transfusion and is taken to the operating room.
	Telesimulation	CT8688	Postpartum hemorrhage	A patient arrives in late term pregnancy and has a precipitous delivery. The infant requires resuscitation and then the mother also has a postpartum hemorrhage requiring resuscitation.

In sequence, we summarized the main demographic characteristics of subjects in the telesimulation and in-person simulation groups. Since this study can be classified as a nonequivalent groups design, we followed Jhangiani and collaborators' guidance (2019) to make our groups as similar as possible, matching variables such as class standing, age, and experience between the groups.

Matching group characteristics improve a study's internal validity by removing some of the most critical confounding factors. Therefore, we tested the similarity of the groups for our main comparison in this study: in-person and telesimulation. Table 38 explores the demographic characteristics of the subjects in the telesimulation and in-person simulation groups and tests the differences between these two groups:

Table 38: Summary statistics of demographic characteristics of subjects in telesimulation and in-person simulation

Variable	Categories	Method of Simulation				Total (3)	
		Telesimulation (1)		In-person (2)			
		N	%	N	%	N	%
Gender	Male	33_a	53.20%	40_a	55.60%	73	54.50%
	Female	29_a	46.80%	30_a	41.70%	59	44.00%
	Other	0_a	0.00%	2_a	2.80%	2	1.50%
Age group	22 – 25	7_a	11.50%	24_b	33.30%	31	23.30%
	26 – 31	40_a	65.60%	40_a	55.60%	80	60.20%
	32 – 36	6_a	9.80%	4_a	5.60%	10	7.50%
	37 – 40	3_a	4.90%	2_a	2.80%	5	3.80%
	Over 40	5_a	8.20%	2_a	2.80%	7	5.30%
Class standing or professional situation	Graduate Student	47_a	75.80%	64_b	88.90%	111	82.80%
	Healthcare professional	3_a	4.80%	8_a	11.10%	11	8.20%
	Resident / Fellow	12_a	19.40%	0_b	0.00%	12	9.00%
Number of semesters of experience with traditional simulation	Less than 1 semester	14_a	22.60%	20_a	27.80%	34	25.40%
	1 - 2 semesters	8_a	12.90%	19_a	26.40%	27	20.10%
	3 - 4 semesters	24_a	38.70%	29_a	40.30%	53	39.60%
	5 - 6 semesters	9_a	14.50%	2_b	2.80%	11	8.20%
	Over 6 semesters	7_a	11.30%	2_b	2.80%	9	6.70%
Scenario Complexity	Low complexity	33_a	53.20%	36_a	50.00%	69	51.50%
	Medium complexity	29_a	46.80%	36_a	50.00%	65	48.50%

Note: Z-test for column proportions - Each subscript letter (a or b) denotes a subset of Method of Simulation categories whose column proportions do not differ significantly from each other at the .05 level.

We conducted z-tests to test the differences in characteristics between the two groups with the null hypothesis that the frequencies of characteristics are the same between the two groups. The tests are employed for each category in each demographic characteristic between the telesimulation and in-person simulation groups.

The Z-test results are the subscript letters next to the frequency number in each category in the table. The same subscript letters, i.e., ‘a,’ indicate that we fail to reject the null hypothesis at a 5% statistical significance level.

We found no significant differences in frequencies between the two groups, as shown by the Z-tests for gender and scenario complexity. We reject the null hypothesis at a 5% statistical significance level for one cohort (22-25 years) of age group and some categories of class standing or professional situation and two cohorts out of the five in the number of semesters of experience with traditional simulation.

The proportion of participants in medium and low complexity was the same among the two groups. Overall, the background characteristics are similar among the groups, with localized differences in some of the subgroups tested.

We also explored subjects' arousal and emotional response in the telesimulation and in-person simulation groups in the baseline period to verify if they were similar in their initial emotional status. Table **39** shows the mean and standard deviation of arousal and emotional response for subjects in the telesimulation and in-person simulation groups before the simulation.

Table 39: Summary statistics of arousal and emotional response for Telesimulation and in-person groups for baseline

Variables		Telesimulation (1)			In-person (2)			T-test (3)
		N	Mean	SD	N	Mean	SD	
Arousal	*BL*	62	3.27	1.09	72	3.37	1.16	t=-0.52, p=0.61
Emotional state variables								
Bored to alert	*BL*	62	3.39	0.82	72	3.43	0.89	t=-0.29, p=0.77
Lethargic to excited	*BL*	62	3.16	0.71	72	3.31	0.87	t=-1.06, p=0.29
Depressed to elated	*BL*	62	3.21	0.48	72	3.27	0.61	t=-0.68, p=0.5
Tense to calm	*BL*	62	3.42	1.1	72	3.19	0.97	t=1.26, p=0.21
Sad to happy	*BL*	62	3.53	0.67	72	3.5	0.77	t=0.25, p=0.8
Upset to contended	*BL*	62	3.71	0.8	72	3.51	0.84	t=1.38, p=0.171
Stressed to serene	*BL*	62	3.24	0.95	72	2.89	0.93	t=2.17, p=0.03
Nervous to relaxed	*BL*	62	3.16	1.03	72	3.06	0.96	t=0.62, p=0.54

NOTE – BL = Baseline

The statistics tests show no differences between the telesimulation and in-person simulation groups in arousal and emotional responses scales, except for the stress to serene items before simulation. The mean for stress or serene items is 3.24 (SD=0.95) in the telesimulation group, and 2.89 (SD=0.93) in the in-person simulation.

We can observe that frequencies for the study variables are similar in all the characteristics, i.e., gender, age group, class standing or professional situation, number of semesters of experience with traditional simulation, and scenario complexity.

From Table 37 and Table 38, we can infer that the subjects in the telesimulation and in-person simulation groups are similar in arousal, emotional response, and demographic characteristics, prior to the interventions, therefore making the attribution to any other differences due to the treatments more compelling and valid from a statistical standpoint.

5.2. Exploratory Data Analysis

Researchers often utilize exploratory data analysis (EDA) to study and describe data sets, frequently using data visualization techniques as the first procedures before conducting statistical tests and testing their hypothesis. The EDA can determine if statistical approaches are appropriate by understanding the underlying structure of data,

extracting variables from the dataset, verifying the presence of outliers and other errors, and testing data distribution assumptions such as normality assumptions.

The emergence of EDA is attributed to John Tukey, an American mathematician author of the seminal paper “Exploratory data analysis” during the 1970s (Tukey, 1977), and influenced many research fields beyond statistics and mathematics. Diaconis (2006) claims that EDA exploratory data analysis has been essential to scientific advancement when performed rigorously and combined with classical statistical tests.

Hair Jr. et al. (2010) instruct readers that data must be processed and checked before being submitted to statistical analysis to verify correlations between variables and test hypotheses. If the data quality is not verified before the conduction of statistical tests, biased results may emerge, threatening the reliability of the findings and introducing bias in various statistics for population estimation.

Exploratory analysis can be used to find errors in the data input process and coding, examine missing data, identify the presence of outliers, describe, and calculate distribution features like means, standard errors, kurtosis, and skewness, perform normality tests, and ultimately determine whether to employ parametric or non-parametric procedures. The exploratory analysis for this study is described in the following sections.

5.3.Missing Data Analysis

Missing data analysis is one of the first analyses conducted in a survey-based study. Missing data (or missing values) is a data value for a variable in the observation that is not saved in a dataset. Missing data is a prevalent problem in virtually all studies, and it can substantially impact the conclusions that can be taken from the data (KANT, 2013). The unwillingness or forgetting of a responder to mark a question is the most common cause of lost data in surveys that use questionnaires. Another reason is related to errors in the data input step (HAIR et al., 2010).

Missing data in a survey can cause several problems, especially if widespread. It reduces the total sample size and may reveal tendencies, such as a bias between the group with missing data and the group that answered all questions (HAIR Jr. et al., 2010).

Kang (2013) indicates four critical problems created by missing data. The first problem is that missing data may affect a study’s statistical power. Additionally, it can also lead to parameter estimation bias. Third, it may diminish the sample's

representativeness. Finally, the fourth problem is that missing data hinders the data analysis or makes it impossible since some statistical procedures are missing data.

Missing data is generally expected in any study; thus, the planned sample size will be increased to account for that fact. However, this does not rule out the possibility of bias. In the planning and execution of research and the analysis of data, researchers should pay attention to missing data, especially in the case of any underlying missing data pattern (Kang,2013)

Kant (2013) explains that when a research protocol is created and the data is collected, the best approach to avoid or minimize missing data is to optimize data collection. Complex missing-data replacement techniques should be used only after eliminating missing data in the design and preventive measures.

In the present, we balanced the participant's freedom not to answer questions with the risk of missing data. We did not make any response mandatory, allowing questions to go unanswered. On the other hand, during the pre-test stage, we aimed to simplify the survey design to limit users' exhaustion and problematic questions during the completion of the questionnaires.

For the study of the missing data pattern as recommended by Kant (2013) and Hair et al. (2010) we employed the SPSS Version 26.0.1 procedures – "Analyze Patterns – Multiple Imputation" and "Missing Value Analysis". Before executing the missing data pattern, we visually observed the responses and removed five since they were deemed incomplete, with more than 50 % of the data considered missing.

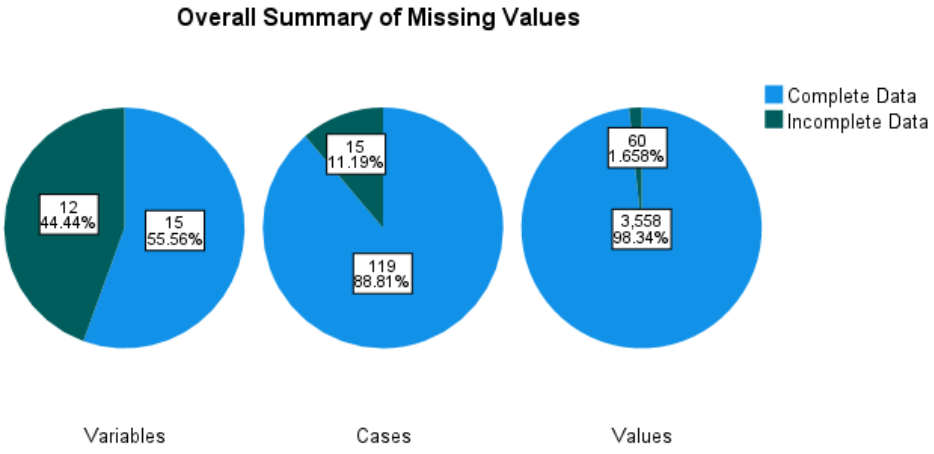
Given the considerable number of variables in this study, we grouped the items by the instrument utilized to execute the missing data analysis related to the several scales used and grouped them into similar scales and subdimensions in four blocks. These were analyzed with similar scales to understand any pattern and facilitate the imputation process. Table 39 lists the main blocks used for testing missing data:

Table 40: Blocks for testing of Missing Data

Block	Description	Moments
Block 1	Emotion Scale – Feldman, Barret and Russell (2012) FAS - Felt Arousal Scale (FAS) (Svebak & Murgatroyd, 1985)	Baseline, post simulation and post debriefing.
Block 2	Emotion Appraisal –Scherer and Moors' (2019) categorization of appraisal dimensions of the emotion process.	Post-Simulation
Block 3	NASA Task Load Index (NASA-TLX) – (Hart, 2006; Hoonakker et al., 2011)	Post-Simulation
Block 4	Fidelity/Realism and Feedback dimensions of Simulation Design Scale (SDS student version; Jeffries, 2016) –	Post simulation and post debriefing.

The first block tested was related to the Emotion Scale and the Single item Felt Arousal Scale. These instruments were collected in the three moments of the simulation session. A total of 27 variables were tested using the Missing Analysis Procedure. In the sequence, we extracted the Summary of Missing values. The results are available in Figure 41:

Figure 42: Summary of Missing Values – Block 1



The procedure results indicate that 98.84% of the data was complete, with 60 missing values out of 3,558 complete values. 88.81% of the cases had no missing value, while 15 cases (11.19%) had at least one missing value for all the variables. Among all participants, 55.56% of the variables (15) were complete, while 44.44% (n=12) had at least one missing value among all the observations.

Following the high-level analysis, we performed a frequency analysis for all the variables that compose the first Block. The frequency analysis of missing values in this block is available in Table 40:

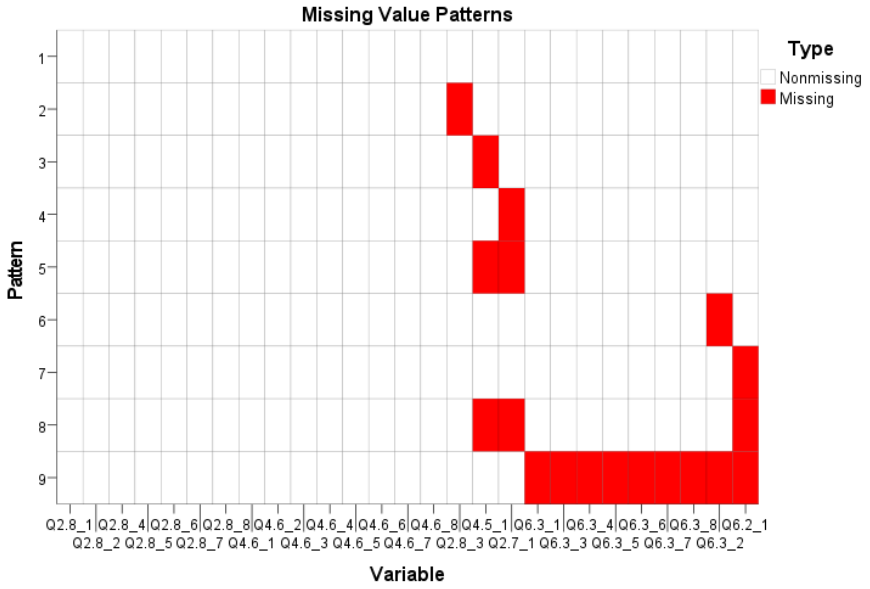
Table 41: Missing Data Summary for Block 1 – Emotional Responses

Variable Summary		Valid N	Missing	
Variable	Description		N	Percent
Q2.7_1	(baseline) Estimate here how aroused you feel. - Low Arousal:High Arousal	129	5	3.70%
Q2.8_1	(baseline) how you feel now - Bored:Alert	134	0	0.00%
Q2.8_2	(baseline) how you feel now - Lethargic:Excited	134	0	0.00%
Q2.8_3	(baseline) how you feel now - Depressed:Elated	133	1	0.70%
Q2.8_4	(baseline) how you feel now - Tense:Calm	134	0	0.00%
Q2.8_5	(baseline) how you feel now - Sad:Happy	134	0	0.00%
Q2.8_6	(baseline) how you feel now - Upset:Contended	134	0	0.00%
Q2.8_7	(baseline) how you feel now - Stressed:Serene	134	0	0.00%
Q2.8_8	(baseline) how you feel now - Nervous:Relaxed	134	0	0.00%
Q4.5_1	(PS) Estimate here how aroused you feel. - Low Arousal:High Arousal	130	4	3.00%
Q4.6_1	(PS) how you felt during the simulation - Bored:Alert	134	0	0.00%
Q4.6_2	(PS) how you felt during the simulation - Lethargic:Excited	134	0	0.00%
Q4.6_3	(PS) how you felt during the simulation - Depressed:Elated	134	0	0.00%
Q4.6_4	(PS) how you felt during the simulation - Tense:Calm	134	0	0.00%
Q4.6_5	(PS) how you felt during the simulation - Sad:Happy	134	0	0.00%
Q4.6_6	(PS) how you felt during the simulation - Stressed:Serene	134	0	0.00%
Q4.6_7	(PS) how you felt during the simulation - Upset:Contended	134	0	0.00%
Q4.6_8	(PS) how you felt during the simulation - Nervous:Relaxed	134	0	0.00%
Q6.2_1	(PD) Estimate here how aroused you feel - Low Arousal:High Arousal	125	9	6.70%
Q6.3_1	(PD) how you felt during the debriefing - Bored:Alert	129	5	3.70%
Q6.3_2	(PD) how you felt during the debriefing - Lethargic:Excited	128	6	4.50%
Q6.3_3	(PD) how you felt during the debriefing - Depressed:Elated	129	5	3.70%
Q6.3_4	(PD) how you felt during the debriefing - Tense:Calm	129	5	3.70%
Q6.3_5	(PD) how you felt during the debriefing - Sad:Happy	129	5	3.70%
Q6.3_6	(PD) how you felt during the debriefing - Stressed:Serene	129	5	3.70%
Q6.3_7	(PD) how you felt during the debriefing - Upset:Contended	129	5	3.70%
Q6.3_8	(PD) how you felt during the debriefing - Nervous:Relaxed	129	5	3.70%

The next step was to understand the different missing data patterns. This procedure categorizes the different patterns of missing data by identifying any regular patterns that might indicate problems with instruments and demonstrating if the missing data is random or an underlying pattern.

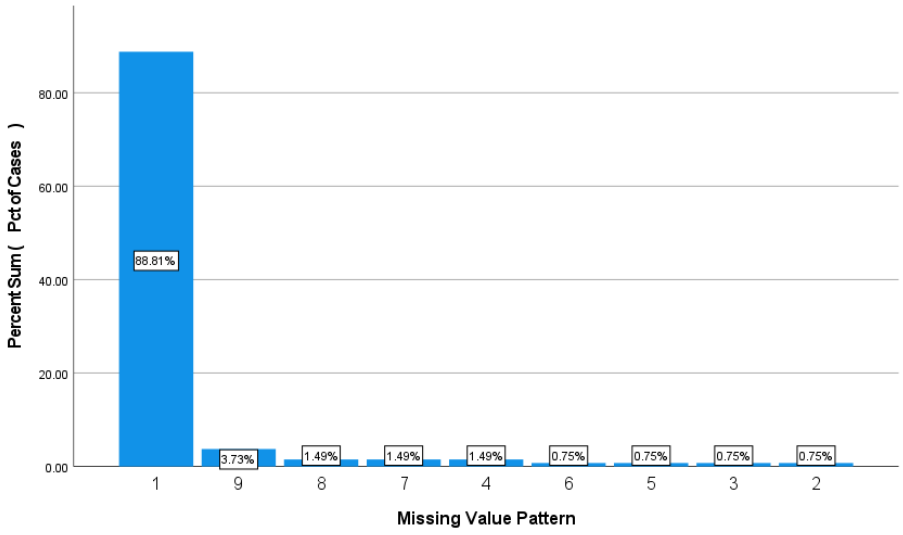
A total of nine patterns were found. The first pattern (1) represents a blank line, indicating no missing values. The remaining patterns (2 to 8) have a small number of missing variables, varying from one to three missing values. The 9th pattern indicated a missed questionnaire section not being filled out. Those were related to the debriefing section of the questionnaire. Figure 43 displays the Missing Value Pattern for the Emotional State and Arousal Scales:

Figure 43: Missing Value Pattern – Block 1



Finally, a complementary analysis was performed to understand the missing value pattern. Using the information about the nine different patterns, we performed a frequency analysis of the occurrence of each of the patterns. Results are available in Figure 44:

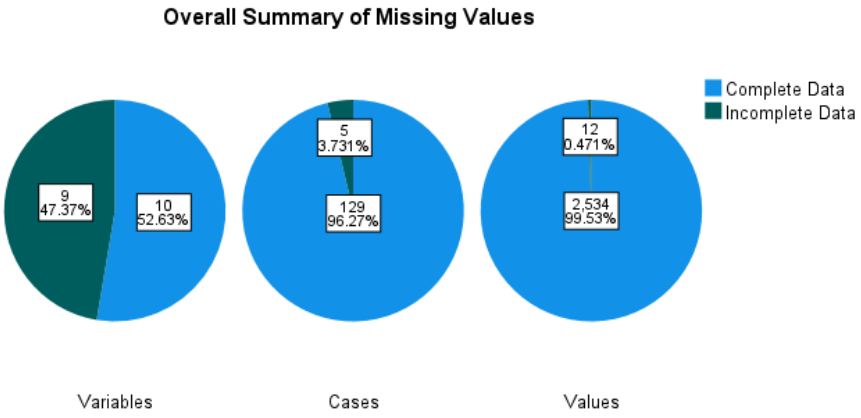
Figure 44: Missing Value Pattern – Pattern Distribution – Block 1



The distribution of patterns regarding the missing values indicates that 88.91% of the responses do not have any missing data. The number of the other missing data patterns is below 5%, indicating no problem of systematic or widespread problem with the instrument or data collection.

In the sequence, we continued our analysis, now with variables from the second block of variables, composed of 19 items related to the emotional appraisal process. Figure 45 presents the summary of missing values:

Figure 45: Summary of Missing Values – Block 2



For Block 2, 99.53% of the data was complete, with 12 missing values out of 2,534 complete values. In this group, 129 cases (96.27%) had no missing value, while 5 cases (3.73%) had at least one missing value for the variables in Block 2. Ten variables

were complete, and nine had at least one missing value for each participant. In the sequence, we performed the frequency analysis for all Block 2 variables:

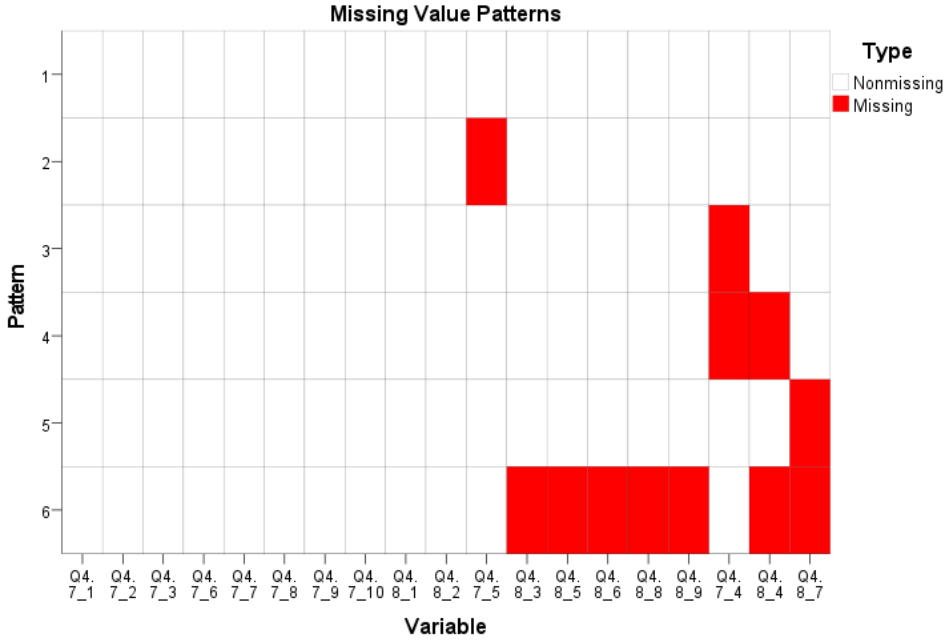
Table 42: Missing Data Summary for Block 2 – Emotional Appraisal

Variable Summary		Valid N	Missing	
Variable	Variable Description		N	Percent
Q4.7_1	How pleasant or unpleasant was your overall experience ?	134	0	0.00%
Q4.7_2	How enjoyable was your overall experience ?	134	0	0.00%
Q4.7_3	To what extent was the overall experience consistent to what you wanted to achieve during this session?	134	0	0.00%
Q4.7_4	In comparison to what you desired, how consistent was your overall experience ?	132	2	1.50%
Q4.7_5	To what extent do you see the overall experience in this session important or not important in achieving your learning goals or needs?	133	1	0.70%
Q4.7_6	How sure were you about what was happening?	134	0	0.00%
Q4.7_7	How well did you understand what was happening around you?	134	0	0.00%
Q4.7_8	To what extent did the overall experience was something you expected or unexpected?	134	0	0.00%
Q4.7_9	How usual or unusual was your overall experience ?	134	0	0.00%
Q4.7_10	To what extent was the overall experience you had in the simulation session familiar or unfamiliar ?	134	0	0.00%
Q4.8_1	How well could you predict what was going to happen?	134	0	0.00%
Q4.8_2	To what extent did you feel that circumstances beyond anyone's control determined the outcomes/results of this simulation?	134	0	0.00%
Q4.8_3	How responsible did you think someone or something other than yourself was for the outcomes/results of the simulation session?	133	1	0.70%
Q4.8_4	When you were in this session, how powerful (dominant) did you feel?	132	2	1.50%
Q4.8_5	When you were in this simulation session, how helpless did you feel?	133	1	0.70%
Q4.8_6	To what extent did you feel that you could influence what was happening in the simulation session?	133	1	0.70%
Q4.8_7	How important was this simulation session to you?	132	2	1.50%
Q4.8_8	How fair did you think what happened to you in this simulation session was?	133	1	0.70%
Q4.8_9	To what extent did you feel cheated or wronged in this simulation session ?	133	1	0.70%

Six patterns were found in Block 2. The first pattern (1) represents a blank line, indicating no missing values. The remaining patterns (2 to 5) had either one or two missing variables. The last pattern (6) represents a more significant portion of the questionnaire not being filled out. Those were related to the questionnaire's situational

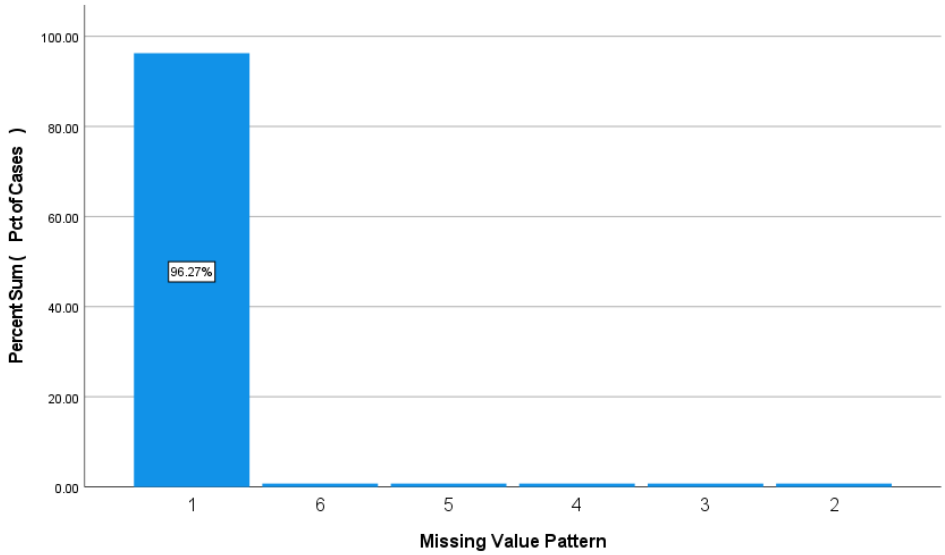
agency, fairness, and importance section. Figure 46 displays the Missing Value Pattern for the Emotional Appraisal process:

Figure 46: Missing Value Pattern – Block 2



The last analysis for this block was the distribution of the missing value pattern. Again, we performed a frequency analysis of the occurrence of each of the patterns. Results are available in Figure 47:

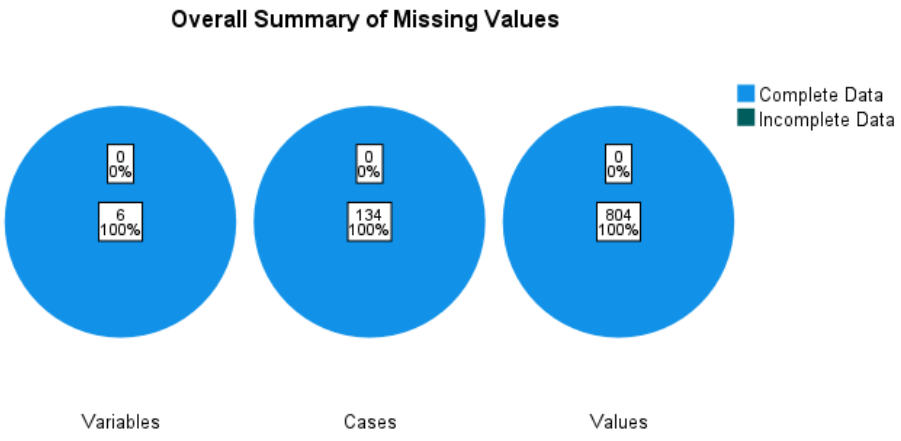
Figure 47: Missing Value Pattern – Pattern Distribution – Block 2



Once again, most of the data for Block 2 was complete, with all other missing value patterns responding together for less than 4% of the missing values. Therefore, no systematic problem of missing values was detected in this Block.

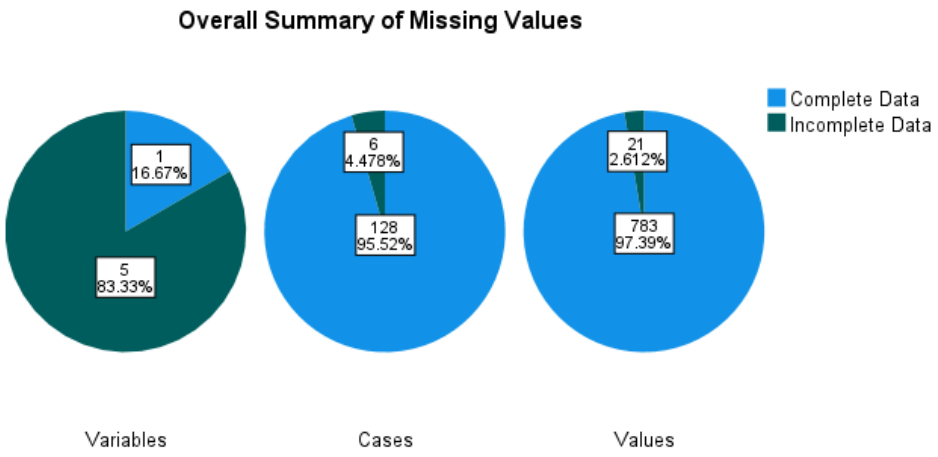
The next block tested, Block 3, was composed of variables of the NASA-TLX Scale. The Summary of Missing Values was executed, and it is displayed in Figure 48:

Figure 48: Summary of Missing Values – Block 3



All the data was complete for the six variables part of the NASA-TLX instrument. Therefore, no further analyses were performed for Block 3. We tested variables for Block 4, aggregating the variables related to the perception of fidelity (realism) and quality of feedback. The results are available in Figure 49:

Figure 49: Summary of Missing Values – Block 4



Considering the data from Block 4, we found 21 missing values (2.61%) while the remaining 97.39% of the values were complete. One hundred twenty-eight

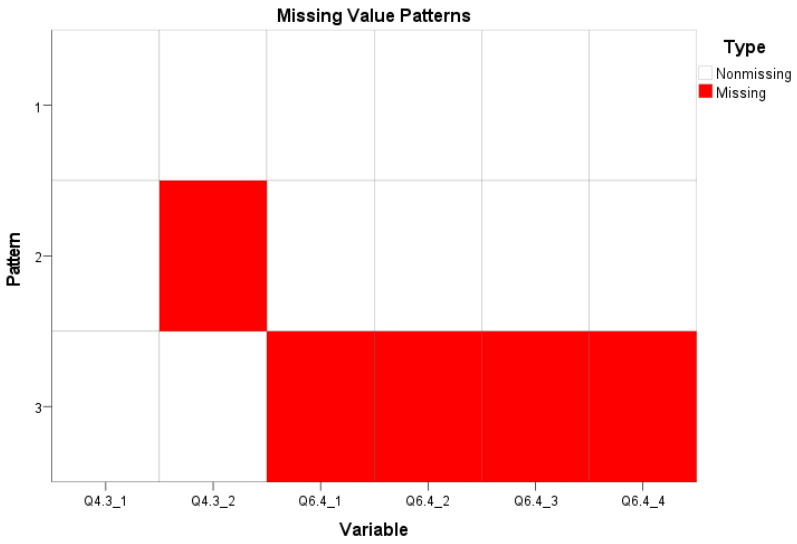
respondents (95.52%) completed all the questions, with only six respondents (3.73%) having at least one missing value for the variables in Block 2. However, in terms of variables, only one variable was completed by all participants, while 5 had at least one missing value. Following the initial analysis, we examined the frequency for all variables composing the fourth block of variables:

Table 43: Missing Data Summary for Block 4 – Fidelity and Feedback Quality

Variable Summary		Valid N	Missing	
Variable	Variable Description		N	Percent
Q4.3_1	Perception of Fidelity (Realism) of the simulation - The scenario resembled a real-life situation	134	0	0.00%
Q4.3_2	Perception of Fidelity (Realism) of the simulation - Real life factors, situation, and variables were built into the simulation scenario	133	1	0.70%
Q6.4_1	Feedback/Guided Reflection - Feedback provided has constructive	129	5	3.70%
Q6.4_2	Feedback/Guided Reflection - Feedback was provided in a timely manner	129	5	3.70%
Q6.4_3	Feedback/Guided Reflection - The simulation allowed me to analyze my own behavior and actions	129	5	3.70%
Q6.4_4	Feedback/Guided Reflection - There was an opportunity after the simulation to obtain guidance/feedback from the teacher in order to build knowledge to another level	129	5	3.70%

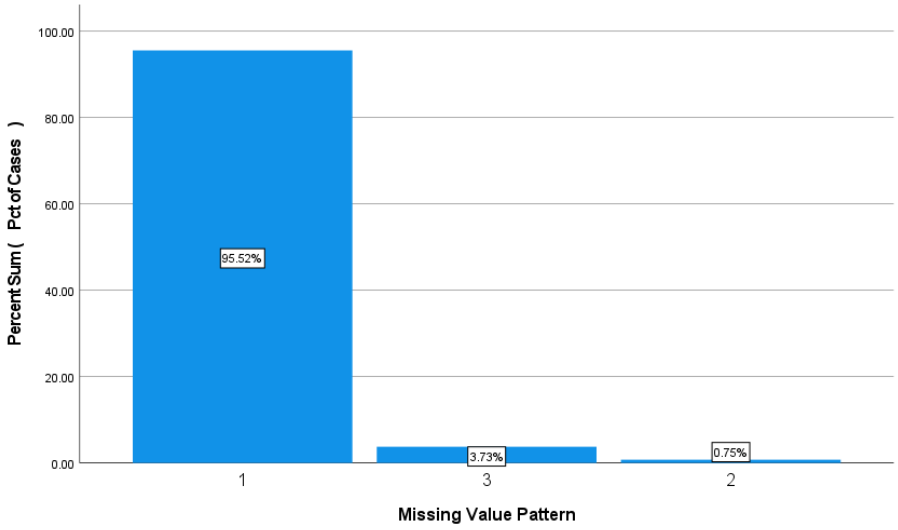
For the fourth Block, only three patterns were detected. The first pattern (1) was represented by a blank line, indicating no missing values. Pattern 2 had one missing value, and the third indicated a more significant portion of the variables detected as missing. Those were related to the feedback section of the block. Figure 50 displays the Missing Value Pattern for the Fidelity and Feedback questions:

Figure 50: Missing Value Pattern – Block 4



We then analyzed the frequency of missing values distributed in the three different patterns. Results are available in Figure 51:

Figure 51: Missing Value Pattern – Pattern Distribution – Block 4



Block 4 had more than 95% of the data complete (represented by pattern 1), while pattern 3 was present in 3.73% (n=5) of the observations, while the second pattern was identified in one respondent. Overall, the analysis for this block demonstrates a low incidence of missing values.

In summary, all our analysis indicates that missing values, although present in our sample, were not a systematic issue and not a likely threat to our study validity. 98.82%

of data was complete, with 113 cases with missing values, while 21 respondents had at least one missed value. Regarding variables, 55.93% were completed by all respondents, with 44.07 % missing at least one response.

Therefore, to treat the missing values encountered, we followed the procedure recommended by Kang (2014) and performed the imputation of the missing data using estimated values. Since the imputation substitutes missing data with a statistically probable value calculated information in the dataset, it allows for preserving all the observations. After replacing missing values, the data set can be analyzed using typical approaches for a complete data set.

We employed the EM (expectation-maximization) method. This method assumes that the pattern of missing data is missing at random, or MAR, allowing the estimated values to be adjusted using information available in the dataset.

After all the procedures, 93 values were replaced from 7813 values present in the questionaries, representing 1.18 % of the data, resulting in 134 usable questionnaires with no missing values, therefore suitable for all statistical tests in this study.

5.4. Outlier analysis

According to Hodge & Austin (2004) outlier detection is a term that refers to various strategies. Although different names are present throughout the specialized literature, many strategies for finding outliers are the same. Authors have been using terms such as outlier detection, novelty detection, anomaly detection, noise detection, deviation detection, and exception mining, among others, to represent the process of detecting deviant observation from data.

An outlying observation, also known as an outlier, is an observation that appears to depart significantly from the rest of the sample (Hodge & Austin, 2004). Therefore, outliers are data points that are not typical of the population and can be considered data contaminants since they do not accurately reflect the population. An outlier is a data point that does not follow the overall trend of the rest of the data in bivariate terms (Osbone & Overbay, 2004)

According to Osbone and Overbay (2004), various factors might lead to inconsistent or abnormal data in a dataset. The reason is errors made during the data input process, including typing errors, conversion, encoding, or calculation failures. Another

factor might be respondent bias, deliberate misreporting, or negligent completion of the measuring questionnaire.

Osbone and Overbay (2004) also point out that outliers can occur when the sampling process is ineffective and sampled subjects have different parameters of interest compared to the target population. These observations may represent valid cases for future analysis, but if they are maintained in the study, they can significantly affect the results without being correctly identified.

Outliers can intensely affect the results and validity of statistical analyses. Osbone and Overbay (2004) point out three main problems due to the non-identification of outliers: the increase in error variance associated with lower statistical power, a departure from data normality, and the influence on the estimators obtained, such as correlation coefficients, means, standard deviation, regression coefficient among others. A point is influential when it disproportionately modifies any regression analysis component, such as responses, slope coefficients, or hypothesis test results.

We employed two complementary procedures for the present study, verifying outliers due to data input errors and applying statistical methods to test for extreme values. In the sequence, both approaches are presented.

A data frequency analysis was performed in SPSS to calculate the range of allowed values for each variable of the instruments employed in this research to detect any potential extreme value resulting from a data encoding error. A summary of the results is presented in Table 44:

Table 44: Instrument's minimum, maximum, and range of values

Instrument	Minimum Value	Maximum Value	Maximum Range	Values out of range (Outliers)
FAS scale	1	6	5	No
Emotion State Scale	-2	2	4	No
Emotions Appraisal Process	1	7	6	No
NASA TLX	1	7	6	No
Sense of Presence and Feedback	1	6	5	No

Source: Compiled by the author

The results indicated the minimum value of 1 for the variables related to the FAS, Emotional Appraisal process and NASA TLX, and sense of presence. In comparison, the

maximum value was seven for the Emotions Appraisal Process and NASA TLX and six for the Sense of Presence and FAS. Finally, the emotion state scale varied from -2 to +2. All data are within the expected range, and no outlier was detected due to the data entry procedure.

In the sequence, we employed the statistical procedure – Grub's Outlier Test – implemented in NCSS 2021, using Grubb's method. The procedure tests the hypothesis that the sample's furthest value means an outlier. Data determines all parameters directly from data, and the approach does not require user-supplied settings. Finally, the significance threshold is compared to the query's Z value to determine the presence of outliers (Hodge & Austin, 2004).

We have selected Grub's outlier test and performed the tests by grouping the variables among items in the same instrument as the previous analysis. We have selected the "Multiple Response Variables, Analyzed Together" option. We retained the default options of a maximum of possible outliers as 10% and a maximum count of 5. We used two-tailed tests (to detect outliers significantly below or higher than the mean value) as recommended by the NCSS manual. The results of the analysis are available in Table 45:

Table 45: Outlier Testing using Grub's outlier test

Group of Variables	Count	Mean	Standard Deviation	Grubbs' Single-Value of Possible Outlier	ESD |Z|	Outlier Test Prob Level	Conclude Outlier by Rosner's Procedure?
FAS Scale	402	3.7637	1.2299	1	2.2471	1	No
				1	2.2643	1	No
				1	2.2818	1	No
				1	2.2997	1	No
				1	2.3181	1	No
Emotion Scale (Baseline, Post-simulation and post-debriefing)	3216	0.4356	0.8766	-2	2.7784	1	No
				-2	2.7822	1	No
				-2	2.786	1	No
				-2	2.7898	1	No
				-2	2.7936	1	No
Emotions Appraisal Process	2546	4.5551	1.4101	1	2.5212	1	No
				1	2.5248	1	No
				1	2.5285	1	No
				1	2.5322	1	No
				1	2.5359	1	No
NASA TLX	804	3.4216	1.775	7	2.0159	1	No
				7	2.0223	1	No
				7	2.0288	1	No
				7	2.0353	1	No
				7	2.0418	1	No
Feedback Dimension	536	5.5709	0.7042	3	3.6508	0.1289	No
				3	3.7007	0.1053	No
				3	3.7527	0.0851	No
				3	3.8069	0.0679	No
				3	3.8636	0.0535	No
Perception of Reality (fidelity)	268	5.0634	0.9519	1	4.2689	0.0038	Yes
				1	4.4313	0.0017	Yes
				1	4.6138	0.0007	Yes
				1	4.8209	0.0002	Yes
				2	3.8323	0.0271	Yes

Source: Compiled by the author

When aggregating the data by instrument, we did not find outliers except in the perception of reality scale, with two items. A total of five observations (5 out of 134 - 3%) had values departing from the expected values only in the perception of reality dimension.

These observations were evaluated individually, and we determined that they were valid cases of participants that experienced a lower perception of fidelity in the simulation sessions. Although the values were statistically deviant from a high perception of fidelity by most participants, therefore being valid values, although with a statistically significant p-value for the test. Consequently, we kept all values in the analysis and determined that outliers' presence was not a threat to the study's validity.

5.5. Data normality tests

Several statistical tests assume that the data are normal, or approximately normal. However, violating these assumptions can lead to biased or unreliable statistical results, with widely varying test sensitivity to violations or departure of the normality. Many methods verify if a variable has a distribution close to the normal curve. Therefore, it is imperative to carry out tests of data normality before applying parametric statistical tests (Hair Jr. et al., 2010).

One group of methods used to verify data normality is based on a visual inspection of frequencies, graphs, and distributions. In contrast, the other group of methods is related to the statistical tests to verify the normality of the data, which can be based on in comparison of asymmetry or kurtosis values of the data, data distribution graphs, where the values found must be close to a normal line, in addition to statistical tests to verify the proximity of the data to a normal distribution (Hair Jr. et al., 2010; Kline, 2011).

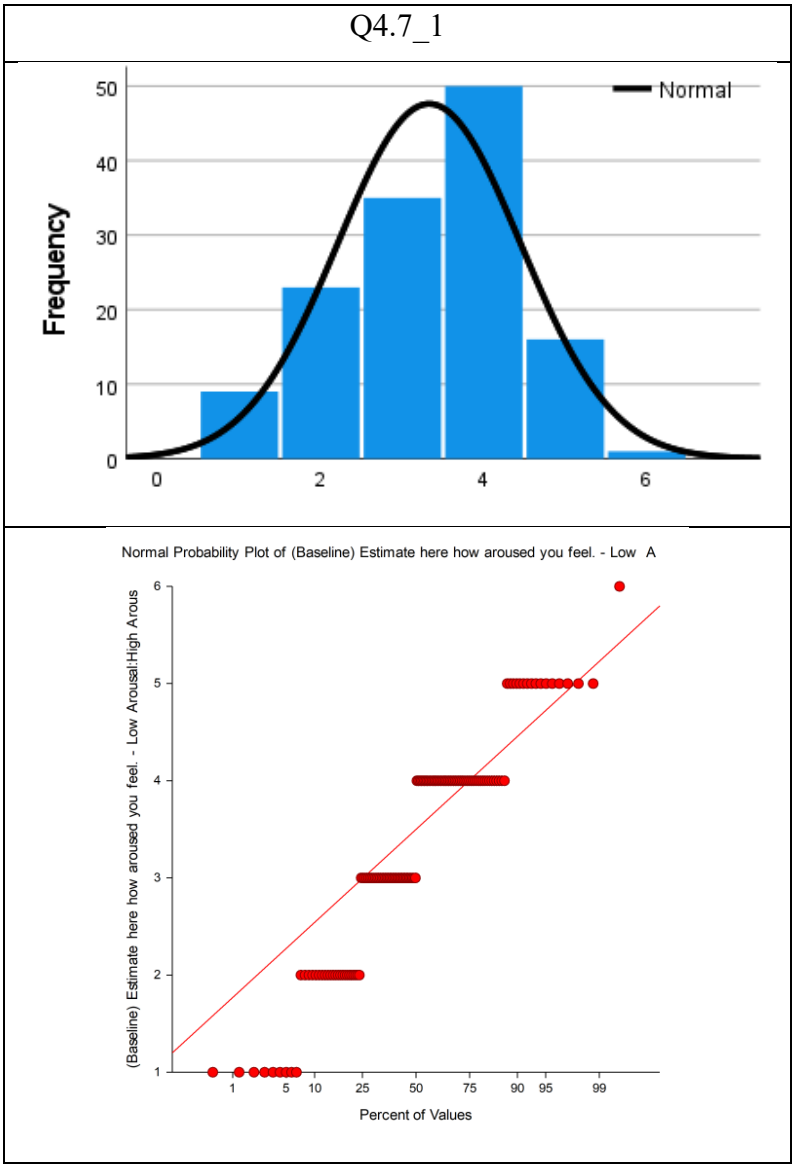
According to Kim (2013), several statistical approaches test data normality. However, there is currently no gold-standard method. The eyeball test (visual inspection) may be suitable for samples of medium to large size (e.g., $n > 50$) but not for small samples.

Considering the known limitations in the visual and statistical methods, we employed both approaches, considering their complementarity to detect significant deviations from normality. Also, the normality of individual items that compose the scales was verified. In a second moment, a statistical test was performed to investigate the normality of the scales/constructs employed in this study. The following items describe the testing process.

Graphic analysis of individual variables

We have performed a variable plot in the SPSS and NCSS to verify the data distribution, with histograms accompanying the normal curve as a reference and normal probability plots (normal P-P plot). A total of 116 plots were executed (2 for each variable) with the following format:

Figure 52: Sample of Frequency and PP-PLOT to check for normality



Note: Frequency Distribution and PP-PLOT

According to the distribution analysis, the results plots are presented in Appendix 2: Frequency distribution of study variables. Our data has a distribution close to normal for the main variables of the study, therefore being generally suitable for parametric tests.

Kurtosis and Skewness analysis of individual variables

The following verification was operationalized by the SPSS FREQUENCIES procedure, where each item's skewness (asymmetry) and kurtosis were calculated. Skewness measures data asymmetry, while kurtosis measures a distribution's 'peakedness'[14].

Verifying the skewness and kurtosis of data distributions is frequently utilized to check for normality (Hair Jr et al., 2010; Kline, 2011). There are several guides regarding the acceptable values for skewness and kurtosis.

For instance, George & Mallery (2018) indicate that skewness and kurtosis values should be in the range of -2 to +2. According to Hair et al. (2010) and Byrne and Van de Vijver (2010), data is deemed normal if the skewness is between 2 and +2 and the kurtosis is between 7 and +7. Lastly, according to Kline (2011), absolute Skewness larger than 3 and Kurtosis values greater than 10 may suggest an issue, with values greater than 20 indicating a more significant problem. Results of kurtosis and skewness for 58 variables can be consulted in Appendix 3: Frequency distribution and Skewness and Kurtosis analysis of study variables. A summary of the results is present in Table 46:

Table 46: Guidelines for normality

Source	Criterion for normality	Results
George and Mallery (2018)	Skewness and kurtosis values should be in the range of -2 to +2	5 variables (out of 58) had higher kurtosis and/or skewness than the guidelines: Feedback and Perception Presence - (Q4.3_1; Q4.3_2; Q6.4_1 ; Q6.4_2; Q6.4_4)
Hair et al. (2010) and Byrne and Van de Vijver (2010)	Skewness is between 2 and +2 and the kurtosis is between 7 and +7.	2 variables (out of 59) had higher skewness (but not kurtosis) than the guidelines. Feedback and Perception Presence - Q6.4_2 (skew=-2.264) and Q6.4_4 (skew=-2.065)
Kline (2011)	Skewness smaller than 3 and/or Kurtosis values smaller than 10	All variables are within the guidelines

Source: Compiled by the author

[14] Peakedness - Lexico Dictionary - https://www.lexico.com/en/definition/peakedness - "The extent to which distribution, as expressed in a graph, is concentrated in a peak or series of peaks."

According to the three guidelines employed, most variables have an approximatively normal distribution, adequate for the conduction of parametric tests. All variables are within Kline's (2011) guidelines.

Normality tests of individual variables

Finally, we tested the normality of single items using the three D'Agostino normality tests implemented in NCSS 2011, performing a z test for skewness, kurtosis, and Omnibus (consisting of a combined test of skewness and kurtosis). The specifics of the tests are described in detail in D'Agostino et al. (1990).

The NCSS 2021 DESCRIPTIVE ANALYSIS procedure indicates that if all three p-values of the tests are lower than the predefined significance level (usually 0.05), the variable is not normal.

The statistical power of normality tests is usually small (power to non-normal distributions) unless the sample sizes are at least over 100 (NCSS, 2021). Like the other variables, the tests were conducted in four blocks: a) emotion state variables, b) feedback and perception of presence, c) emotional appraisal, and NASA TLX workload scale.

Table 47: Test of normality for Emotion State Variables

Variable	Description	Skewness Test	Kurtosis Test	Omnibus Test	Normal according to three criteria?	N of normality criteria met
		Prob	Prob	Prob		
Q2.7_1	(Base) Low Arousal:High Arousal	0.121	0.1744	0.1195	Yes	3
Q2.8_1	(baseline) - Bored:Alert	0.1369	0.3733	0.2225	Yes	3
Q2.8_2	(baseline) - Lethargic:Excited	0.3669	0.2264	0.3202	Yes	3
Q2.8_3	(baseline) - Depressed:Elated	0.0001*	0.0189*	0*	No	0
Q2.8_4	(baseline) - Tense:Calm	0.6652	0.1154	0.2637	Yes	3
Q2.8_5	(baseline) - Sad:Happy	0.396	0.0301*	0.0664	No	3
Q2.8_6	(baseline) - Upset:Contended	0.7947	0.3849	0.6628	Yes	3
Q2.8_7	(baseline) - Stressed:Serene	0.6061	0.4839	0.6853	Yes	3
Q2.8_8	(baseline) - Nervous:Relaxed	0.4228	0.0838	0.1627	Yes	3
Q4.5_1	(PS) Low Arousal:High Arousal	0.0018*	0.2946	0.0043*	No	1
Q4.6_1	(PS) - Bored:Alert	0.0002*	0.0513	0.0002*	No	1
Q4.6_2	(PS) - Lethargic:Excited	0.0214*	0.0434*	0.0092*	No	0
Q4.6_3	(PS) - Depressed:Elated	0.2321	0.0688	0.0936	Yes	3
Q4.6_4	(PS) - Tense:Calm	0.4214	0.2237	0.3452	Yes	3
Q4.6_5	(PS) - Sad:Happy	0.2279	0.0543	0.0759	Yes	3
Q4.6_6	(PS) - Upset:Contended	0.2549	0.041*	0.0648	No	2
Q4.6_7	(PS) - Stressed:Serene	0.0837	0.7599	0.2138	Yes	3
Q4.6_8	(PS) - Nervous:Relaxed	0.4801	0.6789	0.7153	Yes	3
Q6.2_1	(PD) Low Arousal:High Arousal	0.3029	0.4055	0.4162	Yes	3
Q6.3_1	(PD) - Bored:Alert	0.1151	0.5234	0.2358	Yes	3
Q6.3_2	(PD) - Lethargic:Excited	0.7195	0.3273	0.5803	Yes	3
Q6.3_3	(PD) - Depressed:Elated	0.0198*	0.9704	0.0662	No	2
Q6.3_4	(PD) - Tense:Calm	0.5666	0.0428*	0.1092	No	2
Q6.3_5	(PD) - Sad:Happy	0.347	0.0639	0.1155	Yes	3
Q6.3_6	(PD) - Upset:Contended	0.3211	0.0075*	0.0172*	No	2
Q6.3_7	(PD) - Stressed:Serene	0.0891	0.287	0.1337	Yes	3
Q6.3_8	(PD) - Nervous:Relaxed	0.976	0.0966	0.2513	Yes	3

Note: * Significant z-test for $p<0.05$

The next block of variables tested was the items related to feedback, fidelity, and perception of presence. Results are present in Table 48:

Table 48: Test of normality for Feedback and perception of presence

Variable	Description	Skewness Test	Kurtosis Test	Omnibus Test	Normal according to three criteria?	N of normality criteria met
		Prob	Prob	Prob		
Q4.3_1	The scenario resembled a real-life situation	0*	0.0001*	0*	No	0
Q4.3_2	Real life factors, situation, and variables were built into the simulation scenario	0*	0*	0*	No	0
Q4.4_1	Rate your level of presence (1 to 100)	0.0006	0.6342	0.0023	No	1
Q6.4_1	Feedback provided has constructive	0*	0*	0*	No	0
Q6.4_2	Feedback was provided in a timely manner	0*	0*	0*	No	0
Q6.4_3	The simulation allowed me to analyze my own behavior and actions	0*	0.0278*	0*	No	0
Q6.4_4	There was an opportunity after the simulation to obtain guidance/feedback	0*	0*	0*	No	0

Note: * Significant z-test for $p<0.05$

Most items related to feedback and sense of presence rejected the hypothesis of normality, with only one item, sense of presence, meeting one normality criteria. Following this analysis, we proceeded to execute the same test for the variables associated with the emotional appraisal process, as displayed in Table 49:

Table 49: Test of normality for Emotional Appraisal Process

Variable	Description	Skewness Test	Kurtosis Test	Omnibus Test	Normal according to three criteria?	N of normality criteria met
		Prob	Prob	Prob		
Q4.7_1	How enjoyable was your overall experience?	0.001*	0.0391*	0.0005*	No	0
Q4.7_2	How pleasant or unpleasant was your overall experience ?	0.0001*	0.0434*	0.0001*	No	0
Q4.7_3	To what extent was the overall experience consistent to what you wanted to achieve during this session?	0.0146*	0.3238	0.0312*	No	1
Q4.7_4	In comparison to what you desired, how consistent was your overall experience ?	0.0074*	0.2356	0.0136*	No	1
Q4.7_5	To what extent do you see the overall experience in this session important or not important in	0.0151*	0.8498	0.0513	No	2

	achieving your learning goals or needs?					
Q4.7_6	How well did you understand what was happening around you?	0.0677	0.335	0.1184	Yes	3
Q4.7_7	How sure were you about what was happening?	0.0711	0.6671	0.1787	Yes	3
Q4.7_8	To what extent did the overall experience was something you expected or unexpected?	0.0669	0.756	0.1779	Yes	3
Q4.7_9	How usual or unusual was your overall experience ?	0.0322*	0.0828	0.0224*	No	1
Q4.7_10	To what extent was the overall experience you had in the simulation session familiar or unfamiliar ?	0.0212*	0.2556	0.0369*	No	1
Q4.8_1	How well could you predict what was going to happen?	0.0306*	0.4012	0.0679	No	2
Q4.8_2	To what extent did you feel that circumstances beyond anyone's control determined the outcomes/results of this simulation?	0.2501	0.1179	0.152	Yes	3
Q4.8_3	How responsible did you think someone or something other than yourself was for the outcomes/results of the simulation session?	0.9328	0.0052*	0.0201*	No	1
Q4.8_4	When you were in this session, how powerful (dominant) did you feel?	0.8463	0.2456	0.5002	Yes	3
Q4.8_5_I	When you were in this simulation session, how helpless did you feel? (Inverted)	0.0303*	0.3878	0.0659	No	2
Q4.8_6	To what extent did you feel that you could influence what was happening in the simulation session?	0.5216	0.0921	0.1972	Yes	2
Q4.8_7	How important was this simulation session to you?	0.0067*	0.5257	0.0207*	No	1
Q4.8_8	To what extent did you feel cheated or wronged in this simulation session? (inverted)	0*	0.0066*	0*	No	0
Q4.8_9	How fair did you think what happened to you in this simulation session was?	0.0001*	0.204	0.0003*	No	1

Note: * Significant z-test for $p<0.05$

The results on the emotional appraisal are mixed, although most items tested were considered normal. Only three variables were not considered normal in all the three criteria, while 16 variables did meet at least one criterion for normality. Finally, our last block of variables tested for normality was the NASA Workload Scale:

Table 50: Test of normality for NASA Workload Scale

Variable	Description	Skewness Test	Kurtosis Test	Omnibus Test	Normal according to three criteria?	N of normality criteria met
		Prob	Prob	Prob		
Q4.9_1	Mental Demand - How mentally demanding was the simulation?	0.0411*	0.9342	0.1238	No	2
Q4.9_2	Physical Demand - How physically demanding was the task?	0*	0.4417	0.0002*	No	1
Q4.9_3	Temporal Demand - How hurried or rushed was the pace of the simulation?	0.5664	0.0033*	0.0112*	No	1
Q4.9_4	Performance - How successful were you in accomplishing what you were asked to do? (I)	0.9508	0*	0*	No	1
Q4.9_5	Effort - How hard did you have to work to accomplish your level of performance?	0.3441	0.0728	0.1279	Yes	3
Q4.9_6	Frustration - How insecure, discouraged, irritated, stressed, and annoyed were you?	0.0006*	0.8615	0.0027*	No	1

Note: * Significant z-test for $p<0.05$

The results indicate that the variables composing the NASA -TLX scale are close to normal as a block, with all the items meeting at least one criterion. One variable met all three criteria, while another met two, and the remaining four items met one criterion for normality.

Table 51 summarizes the normality tests using the D'Agostino approach (1990) implemented by NCSS 2021. The computations performed were the number of variables considered by one and three criteria, along with information related to the number of items tested:

Table 51: Summary of Statistical tests for normality

Description	N	%
	59	100%
Variables considered normal by three criteria	25	42%
Variables considered normal by at least one criterion	48	81%
Subscales		
Emotion State Variables	n (items) = 27	
Variables tested normal by three criteria	18	67%
Variables tested normal by at least one criterion	25	93%
Feedback and Perception of Presence	7	
Variables tested normal by three criteria	0	0%
Variables tested normal by at least one criterion	1	14%
Emotional Appraisal Process	n (items) = 19	
Variables tested normal by three criteria	6	32%
Variables tested normal by at least one criterion	16	84%
NASA – Workload Scale	n (items) = 6	
Variables tested normal by three criteria	1	17%
Variables tested normal by at least one criterion	6	100%

Note – Test for normality: Skewness, Kurtosis and Omnibus

In summary, by observing all three methods used to verify the normality of data, we have evidence that, as a group, the present study data is approximately normal, and sufficient for the conduction of parametric tests.

The group of items related to Feedback and perception of presence was the set of variables more distant from a normal distribution. However, to improve the robustness of our analysis, our analysis and hypothesis tests will be performed using non-parametric alternatives to the test.

It is important to note that the normality of single Likert items is unlikely to be attained. However, our variables will be combined into their respective constructs, rendering the data closer to normal as the data variability grows, with a less discrete (more continuous) distribution.

5.6. Common Method Variance Bias

The extensive use of single-source instruments poses a risk to the research validity known as common-method variance bias. The common-method variance bias has been a topic of study in psychology since the 1950s. Since self-rated questionnaires are among the most popular and practical methods for gathering data in psychology, the problem is even more prevalent (Karimi & Meyer, 2019).

Podsakoff et al. (2003) indicate that the common variance of the method is also a potential problem in most behavioral research, being recognized by most researchers for more than 60 years. This problem occurs when the variance results from the method used (such as questionaries) and not from a real variation of the measured constructs, creating a bias due to the usage of the instrument on the results of the indicators.

A common method variance bias threatens a study's validity since it introduces systematic errors, which can generate alternative explanations between the different constructs, threatening the validity of the conclusions about the relationship between the variables. The validity of the entire study is compromised by introducing a common factor that influences the measured constructs, which can be confused as an impact between the variables studied (Conway & Lance,2010; Mackenzie & Podsakoff, 2012).

According to Podsakoff et al. (2003), several factors can lead to a bias in the variance common to the technique, including respondent effects, item features, item context, and measurement context effects. Table 52 provides an overview of these factors:

Table 52: Common Method Bias Causes and Examples

Causes	Effects of the responder	Effects of the items' characteristics	Effects of the items' context	Effects of the measurement context
Definition	Artificial covariance between a variable predictor and the criterion produced by using the same responder	Artificial covariance caused by the influence or interpretation that a response may have because of some of the item's proprieties.	Influences induced by the setting and the relationship between one component and the others that make up an instrument	Influences relating to the setting in which the measurements are taken
Examples of factors	Seeking consistency between answers Social Desirability Leniency (positively evaluating familiar people) Acquiescence (tendency to agree or disagree regardless of content)	Social desirability of the item Item ambiguity Same scale format for a group of variables Items written in the form of affirmation or negation	Insertion of the item in a group of variables identifiable to the respondent Inducing behavior or responses Instrument length	Predictor variables and criteria measured at the same time, place, or instrument

Source: Adapted from PODSAKOFF et al. (2003) pg. 882.

Likewise, Karimi & Meyer (2019) highlight two types of bias in research that inquire about participants' perceptions: acquiescence and social desirability. The respondents' propensity to indicate agreement with items on the questionnaire regardless of their content is acquiescence. Another important source of bias is social desirability. It represents the tendency of a respondent to answer the questionnaire in a way that allows them to be portrayed positively or desirably and following the norms and standards imposed by their culture, educational or professional group.

According to Podsakoff et al. (2003) 's and Karimi & Meyer (2019) definitions, the present study is potentially subject to the common-method bias. The present study collects predictor and dependent variables concomitantly. It uses the same type of items to measure the variables in the study (Likert or semantic differential scales), and items

related to emotional responses, feedback, and perceptions of demand and difficulty may be influenced by social desirability.

During the questionnaire construction, we employed suggestions described by Podsakoff et al. (2003) to minimize the potential variance problem common to the method in this research. Respondents were informed of the anonymity of responses to reduce the incidence of socially acceptable responses. Since we employed questionnaires, it presents considerable advantages over other data collection methods, as research indicates that face-to-face interviews are more likely to induce socially desirable responses (PODSAKOFF et al., 2003).

In addition, central topics (such as emotional responses, task difficulty, and pre and post-tests) were separated in the questionnaire to build a psychological barrier between the constructs and decrease the effect of initial responses. Shorter scales with fewer items may cause respondents to check their short-term memory and respond to items, similarly falsely enhancing the association between items with similar content.

We also aimed to make the questionnaire within a reasonable size, with 67 items distributed in different moments: pre-simulation, post-simulation, and post-debriefing, making it less likely for respondents to form associations in their minds between items and previous responses.

Although these steps were considered before data collection, statistical techniques were applied to test for a bias resulting from the common variance method. We employed the single factor Harman test to verify the extent of common method bias present in our study.

Harman's single-factor test is one of the most widely used strategies for detecting common method variation concerns that arise when a single data gathering method is utilized. All study variables are aggregated into a single factor in exploratory factor analysis in this test. Suppose this single factor explains most of the variance between all items (more than 50%). In that case, it can be inferred that this variance is common to the method, inducing systematic error in the measurement due to the data collection method (Podsakoff et al., 2003).

An exploratory factor analysis was performed in SPSS using the FACTOR procedure, with the number of factors limited to 1, using the PRINCIPAL COMPONENT extraction method to explain the maximum variance and not the correlation structure between the variables.

Using the principal components method tends to explain a greater amount of variance, being a more robust test to detect the presence of a common factor of variance (the method also maximizes the variance in the first factor extracted) (Podsakoff et al., 2003). We inputted all the constructs measured by scales in the analysis of 58 variables. Results are available in Table 53:

Table 53: Single Factor Variance- Harman Test

Component	Extraction Sums of Squared Loadings - Total Variance Explained		
	Total	% of Variance	Cumulative %
1	11.85	20.43	20.43

Extraction Method: Principal Component Analysis.

Therefore, in this study, the un-rotated exploratory factor analysis with 58 items using the principal component method indicated that the single factor explained only 20.43% of the variation in the study's instruments. These findings suggest the absence of common method bias in this study, following Podsakoff et al. (2003) indication that common method bias is only present when a single factor explains more than half of the variance of all variables using the same method.

5.7.Confirmatory Factor and Reliability Analysis

We tested the constructs employed in this research using confirmatory factor and reliability analysis of the instruments in the Smart PLS software version 3. We also employed the SPSS 2021 software to compute Cronbach's Alpha for our instruments.

A construct represents a real-world abstract concept, also known as unobservable or latent variables such as intelligence, stress, emotional response, and happiness, usually measured by a set of items or questions to capture the underlying theoretical concept by the level of agreement or disagreement with the proposed statements.

Confirmatory factor analysis (CFA) is a technique based on previous theory, in which the relationships between observed (items) and unobserved (latent constructs) variables are proposed according to a theory. CFA is a type of exploratory factor analysis that enables more powerful assessments of a scale's construct validity, such as measuring constructs at the latent level, the distinction of the error component with the variance shared by a factor (Hair et al., 2010).

We used several tests to determine the reliability and discriminant validity of the instruments: Cronbach's Alpha, composite reliability, Fornell and Larcker criteria, rho_A, and Heterotrait-Monotrait Ratio (HTMT).

Cronbach's Alpha, created by Lee Cronbach in 1951, is one of the most used tests to assess the reliability of an instrument. With a value ranging from 0 to 1, the test measures a test's or scale's internal consistency. In a research instrument, reliability refers to the degree of consistency with which it assesses a construct or the capacity to provide identical results when repeated measurements of a characteristic are taken (Hair Jr. et al., 2010).

Internal consistency tests are required before the scale may be used for research or inference, as they also test the measurement instrument's validity. The reliability of a scale reflects how much random measurement error there is and assesses the correlation between items (Hair Jr. et al., 2010).

The Composite reliability (also known as construct reliability) measures internal consistency in scale items similar to Cronbach's alpha (Hair Jr. et al., 2010).

Discriminant validity for instruments employed in this study was determined by comparing the average extracted variance (AVE) and mean shared variance (MSV), as recommended by Fornell and Larcker (1981). To determine discriminant validity between variables, the variance extracted from a variable (AVE) must be greater than the shared variance (MSV). We can determine that the indicators are more closely related to the element they are supposed to measure than other variables.

We tested our variables in blocks of associated variables and constructs to simplify the interpretation of loadings and reliability. The separation in blocks provides more stability given that the ratio of variables to observations is maintained within the recommended ratio of observations per item (Hair Jr. et al., 2010)

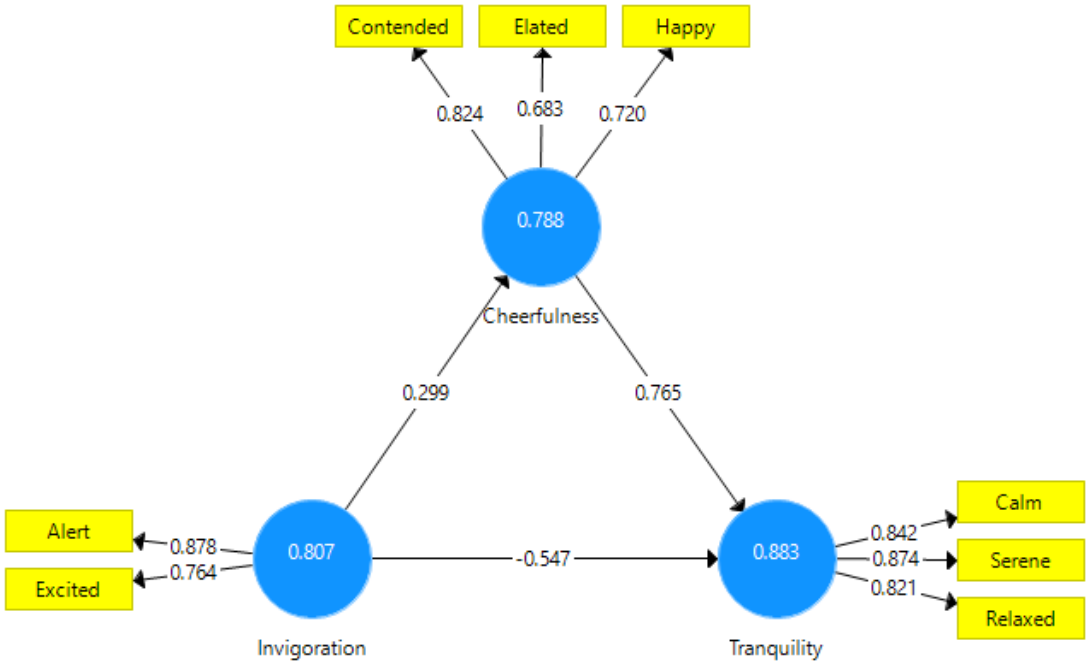
The first scale tested in Smart PLS was the Emotional State Scale. Given that the instrument is measured at three temporal points, we choose the variables associated with the post-simulation moment. It is expected to have more activation and is most commonly studied in healthcare simulation students.

The confirmatory factor analysis was used to identify the emotional scale construct from the eight emotional discrete states: Contented, Elated, Happy, Alert, Excited, Calm, Relaxed, and Serene. The measurement proposed model had three dimensions: Cheerfulness, Invigoration, and Tranquility.

Fraser et al. (2012) determined the two later dimensions and validated them in newer studies. However, the dimension of Cheerfulness was coined after the lack of fitting of the associated items (contended, elated, happy) with the invigoration construct.

The results of the model specification and associated loadings and composite reliability are present in Figure 53:

Figure 53: CFA for Emotion State Scale



Note: Inner Circle number – Composite Reliability / Inner Model – Path Coefficient – Outer Model – item loadings

The model provided satisfactory factor loading (higher than 0.300) for all the variables, indicating that the items are strongly associated with the measured latent constructs. Among the three dimensions, Contented, Elated, and Happy loaded highly to the first component identified as Cheerfulness, with a corresponding factor loading of 0.824, 0.683, and 0.720, respectively.

Two items consisting of Alert and Excited loaded highly to the second component identified as Invigoration with a factor loading of 0.878 and 0.764. The remaining three emotional items were associated with the Tranquility variable, consisting of the Calm, Relaxed and Serene, which loaded highly to the third components with a corresponding factor loading of 0.842, 0.821, and 0.874, respectively.

The model also indicates expected theoretical associations between emotions. Invigoration was negatively associated with Tranquility (path coefficient = -0.547) while positively associated with Cheerfulness (path=0.299). Cheerfulness was also strongly

associated with tranquility (0.765). Table 54 has the factor loadings and factor weights for the variables in this study:

Table 54: Factor Loadings and Factor Weights for the Emotion State Scale

	Factor Loadings			Factor Weights		
Item	Cheerfulness	Invigoration	Tranquility	Cheerfulness	Invigoration	Tranquility
Contended	**0.824**			0.442		
Elated	0.683			0.366		
Happy	**0.720**			0.386		
Alert		**0.878**			0.585	
Excited		**0.764**			0.509	
Calm			**0.842**			0.369
Relaxed			**0.821**			0.360
Serene			**0.874**			0.383

In the sequence, we performed validity and reliability tests for the three dimensions using Cronbach's Alpha, rho_A, composite reliability, and the AVE. Results are available on Table 55:

Table 55: Validity and Reliability for the Emotion State Scale

Dimension	Cronbach's Alpha	rho_A	Composite Reliability	Average Variance Extracted (AVE)
Cheerfulness	0.783	0.795	0.788	0.555
Invigoration	0.803	0.814	0.807	0.677
Tranquility	0.883	0.884	0.883	0.716

The Cronbach's alpha's internal consistency measures were greater than 0.7 across all three emotional dimensions. Likewise, composite reliability shows that the obtained values are greater than 0.6, suggesting that the identified construct is acceptable for capturing an individual's emotional state. The association between items and their respective constructs was also closer to 1 with the construct of cheerfulness with 55.5% average explained variance. In comparison, items of invigoration and the tranquility dimensions accounted for 67.7% and 71.6% of average variance extraction.

Finally, we employed two criteria available in SmartPLS to establish the discriminant validity of the measured constructs. Table 56 presents both analyses:

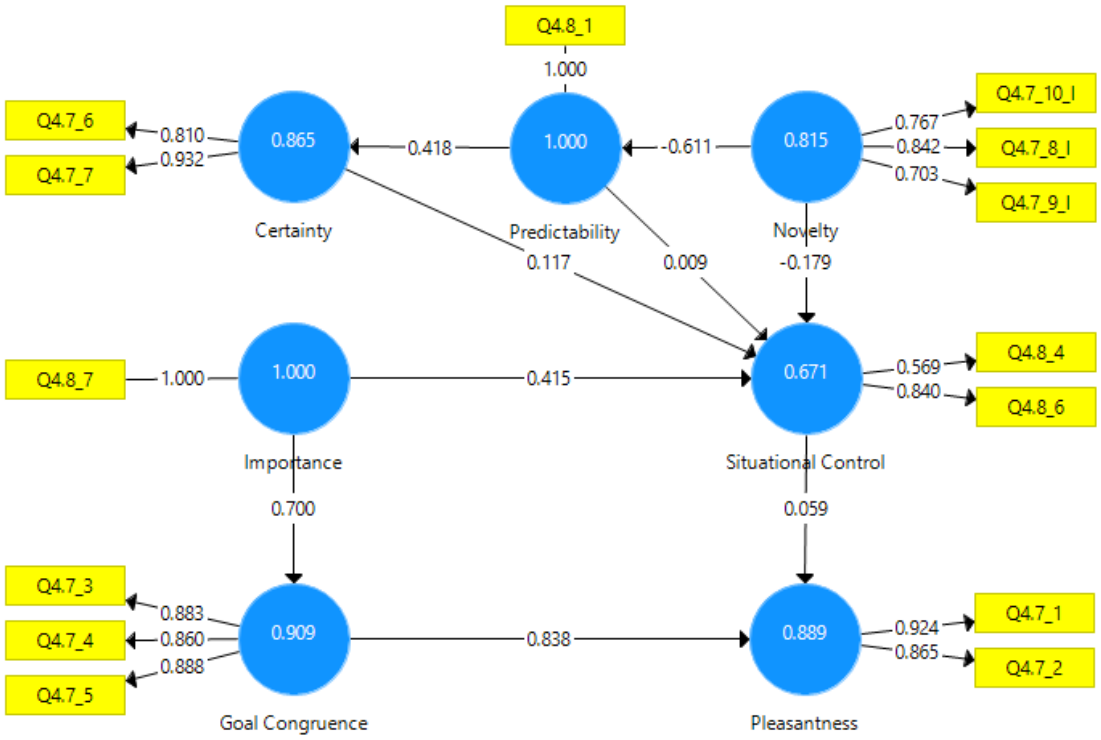
Table 56: Discriminant Validity for the Emotion State Scale

Dimension	Fornell-Larcker Criterion			Heterotrait-Monotrait Ratio (HTMT)		
	Cheerfulness	Invigoration	Tranquility	Cheerfulness	Invigoration	Tranquility
Cheerfulness	**0.745**					
Invigoration	0.299	**0.823**		0.322		
Tranquility	0.601	-0.319	**0.846**	0.597	0.314	

The two criteria for establishing discriminant validity were attained for the Emotion State Scale for all the dimensions, namely the AVE (Cheerfulness=0.745; Invigoration=0.823; Tranquility=0.846) than higher than the shared variance with the other constructs (ranging from 0.601 to -0.319). Likewise, the Heterotrait-Monotrait Ratio (HTMT) was below 1 for the three dimensions.

After executing the confirmatory factor analysis for the Emotion State Scale, we proceeded to test the emotion appraisal scale based on 14 emotional survey items. The result of the analysis classified the 14 items into seven appraisal dimensions. The model specified, associated with the variables, composite reliability, and factor loadings are in Figure 54:

Figure 54: CFA for Emotion Appraisal Scale



Note: Inner Circle number – Composite Reliability / Inner Model – Path Coefficient – Outer Model – item loadings

Items Q4.7_1 and Q4.7_2 was found to load highly to the first component Pleasantness with corresponding loading factor of 0.924 and 0.865, Q4.7_3 to Q4.7_ 5 were highly loaded to the second components identified as Goal Congruence with loading factor of 0.883, 0.860, and 0.888.

Items Q4.7_6 and Q4.7_7 loaded to the same factor Certainty with a factor loading of 0.810 and 0.932. The novelty was formed based on the variables Q4.7_8_1 to Q4.7_10_1. The situational control construct was formed by variables Q4.8_4 and Q4.8_6 with a factor loading of 0.569 and 0.840.

Other items composing the original sub-dimension of Situational Control were removed because their factorial loadings were below 0.300 (items removed were: Q4.8_2, Q4.8_3, and Q4.8_5) and caused a low Cronbach's Alfa to the variables, therefore not being suitable for further analysis. Predictability and Importance were measured by a single item, with the variables Q4.8_1 and Q4.8_7, with a factor loading of 1. The factor loadings and weights for all the variables are present in Table 57:

Table 57: Factor Loadings for the Emotional Appraisal Process

Factor Loadings							
	Pleasant.	Goal Cong.	Certainty	Novelty	Predic.	Sit. Control	Import.
Q4.7_1	0.924						
Q4.7_2	0.865						
Q4.7_3		0.883					
Q4.7_4		0.860					
Q4.7_5		0.888					
Q4.7_6			0.810				
Q4.7_7			0.932				
Q4.7_8_I				0.842			
Q4.7_9_I				0.703			
Q4.7_10_I				0.767			
Q4.8_1					1.000		
Q4.8_4						0.569	
Q4.8_6						0.840	
Q4.8_7							1.000
Factor Weights							
	Pleasant.	Goal Cong.	Certainty	Novelty	Predict.	Sit. Control	Import.
Q4.7_1	0.545						
Q4.7_2	0.51						
Q4.7_3		0.365					
Q4.7_4		0.355					
Q4.7_5		0.367					
Q4.7_6			0.496				
Q4.7_7			0.571				
Q4.7_8_I				0.426			
Q4.7_9_I				0.356			
Q4.7_10_I				0.388			
Q4.8_1					1		
Q4.8_4						0.467	
Q4.8_6						0.689	
Q4.8_7							1

Following the specification of the model and determination of the factor weights and loadings, we have executed tests for the Validity and Reliability of the scale, presented in Table 58:

Table 58: Validity and Reliability for the Emotion Appraisal Scale

	Cronbach's Alpha	rho_A	Composite Reliability	Average Variance Extracted (AVE)
Certainty	0.861	0.873	0.865	0.763
Goal Congruence	0.909	0.909	0.909	0.770
Novelty	0.816	0.821	0.815	0.597
Pleasantness	0.889	0.891	0.889	0.801
Situational Control	0.647	0.713	0.671	0.515

Note: Importance and Predictability are single items measures and all values are by definition 1

The internal consistency measures by the Cronbach alpha were higher than 0.7 , except for the Situational control, which had a value of 0.647. The composite reliability was greater than 0.6 for all variables implying that the identified construct is acceptable for capturing the information on the emotional appraisal process. The AVE of each identified emotional appraisal dimension was closer to 1, with each construct accounting for more than 50% explained variance.

Finally, in the following two tables, we present the tests for the discriminant validity of the emotion appraisal process using the Fornell-Larcker Criterion and Heterotrait-Monotrait Ratio (HTMT) Criterion.

Table 59: Discriminant Validity for the Emotion Appraisal Process - Fornell-Larcker Criterion

	Certainty	Goal Congr.	Import.	Novelty	Pleasant.	Predict.	Sit. Control
Certainty	**0.873**						
Goal Congruence	0.445	**0.877**					
Importance	0.216	0.7	**1**				
Novelty	-0.6	-0.413	-0.215	**0.773**			
Pleasantness	0.545	0.868	0.558	-0.514	**0.895**		
Predictability	0.418	0.119	-0.004	-0.611	0.228	**1**	
Situational Control	0.318	0.523	0.479	-0.343	0.497	0.165	**0.717**

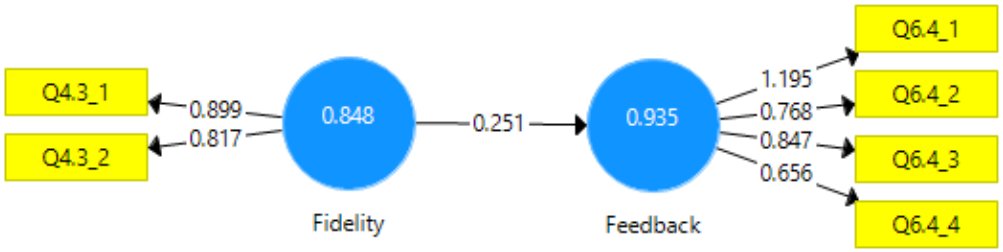
Table 60: Discriminant Validity for the Emotion Appraisal Process

	Certainty	Goal Congr.	Import.	Novelty	Pleasant.	Predict.
Goal Congruence	0.454					
Importance	0.221	0.699				
Novelty	0.597	0.413	0.215			
Pleasantness	0.559	0.869	0.557	0.512		
Predictability	0.417	0.12	0.004	0.608	0.229	
Situational Control	0.325	0.545	0.479	0.367	0.512	0.168

The two criteria for establishing discriminant validity were established for all the constructs tested. The AVE for the Certainty (0.873), Goal Congruence (0.877), Importance(1), Novelty(0.773), Pleasantness(0.895), Predictability(1) and Situational Control (0.717) was than higher than the shared variance with the other constructs (ranging from 0.601 to -0.319). To confirm the discriminant validity of the instruments, the Heterotrait-Monotrait Ratio (HTMT) was below 1.00 for all the dimensions.

The next block of variables tested was the perception of fidelity (realism) and the subscale of debriefing feedback with six variables. Two items were associated with Fidelity, and four were associated with Feedback. The model specified, associated with the variables, composite reliability, and factor loadings are in Figure 55:

Figure 55: CFA for Fidelity (Realism) and Feedback dimensions



Note: Inner Circle figure – Composite Reliability / Inner Model – Path Coefficient – Outer Model – item loadings

The result of the factor analysis showed that items Q4.3_1 and Q4.3_2 loaded to the first components with factor loading 0.899 and 0,817, respectively, identified as the fidelity construct. Fidelity was positively associated with the quality of feedback during debriefing (0.251). The remaining four survey items were loaded to the second extracted component identified as the feedback construct. Table 61 has all the factor loadings and weights:

Table 61: Factor Loadings and Factor Weights for the Fidelity and Feedback dimensions

	Factor Loadings		Factor Weights	
Variables	Fidelity	Feedback	Fidelity	Feedback
Q4.3_1	0.899		0.563	
Q4.3_2	0.817		0.511	
Q6.4_1		1.195		0.376
Q6.4_2		0.768		0.242
Q6.4_3		0.847		0.267
Q6.4_4		0.656		0.207

Following the specification of the model and determination of the factor weights and loadings, we present the validity and reliability statistics of the items related to the perception of fidelity and quality of feedback during debriefing in Table 62:

Table 62: Validity and Reliability for the Fidelity and Feedback Scales

	Cronbach's Alpha	rho_A	Composite Reliability	Average Variance Extracted (AVE)
Feedback	0.934	0.992	0.935	0.791
Fidelity	0.847	0.852	0.848	0.737

The internal consistency using Cronbach's alpha was greater than the recommended value of 0.7 (Hair Jr et al., 2010). The composite reliability statistic also was over the 0.6 thresholds. The result implies that the two obtained constructs are reliable and apt to be used in future analysis with evidence of satisfactory reliability. The average variance extracted was closer to 1 for each construct, with an average extraction of over 70% of the variance. Lastly, we have determined the discriminant validity of both constructs using two criteria as displayed in Table 63:

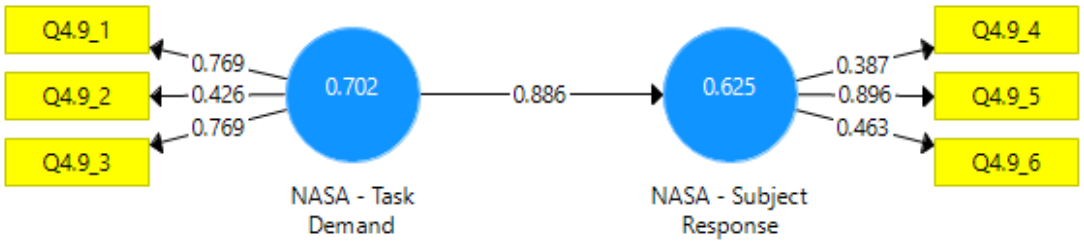
Table 63: Discriminant Validity for the Fidelity and Feedback Scales

	Fornell-Larcker Criterion		Heterotrait-Monotrait Ratio (HTMT)	
	Feedback	Fidelity	Feedback	Fidelity
Feedback	0.890			
Fidelity	0.251	0.859	0.246	

The two criteria for establishing discriminant validity were established for both constructs. The AVE for the Feedback (0.890) and Fidelity (0.859) were both higher than the shared variance (0251). The Heterotrait-Monotrait Ratio (HTMT) was also satisfactory since its value (0.246) was lower than 1.

The last group of constructs tested for reliability and validity was the mental workload defined by the NASA TLX instrument with six items, including Q4.9_1 to Q4.9_6. We tested two different models, one for the two-dimensional model (task demand and subject response) while the second model assumed NASA-TLX as a single dimension. Figure 56 presents the model specified for the first model (two dimensional), with associated factor loadings, composite reliability, and the path between the measured constructs:

Figure 56: CFA for NASA-TLX



Note: Inner Circle figure – Composite Reliability / Inner Model – Path Coefficient – Outer Model – item loadings

Between the two factors, variables Q4.9_1 to Q4.9_3 loaded highly to the first component identified as NASA-Task Demand with a corresponding factor loading of 0.769, 0.426, and 0.769, respectively. The three remaining survey items consisting of Q4.9_4 to Q4.9_6 loaded highly to the second component identified as NASA-Subject Response with a factor loading of 0.387, 0.896, and 0.463, respectively.

Another observable finding is the high association between the two variables (0.886), indicating a strong positive correlation between the task demand and the associated subject response. Table 64 presents the factor weights and loadings for the two dimensions:

Table 64: Factor Loadings and Factor Weights for the NASA-TLX

Variables	Factor Loadings		Factor Weights	
	NASA - Task Demand	NASA - Subject Response	NASA - Task Demand	NASA - Subject Response
Q4.9_1	0.769		0.487	
Q4.9_2	0.426		0.270	
Q4.9_3	0.769		0.487	
Q4.9_4		0.387		0.287
Q4.9_5		0.896		0.665
Q4.9_6		0.463		0.343

After determining the weights and loadings, we executed the tests for the validity and reliability of the scale. Results are available in Table 65:

Table 65: Validity and Reliability for the NASA-TLX Scale

	Cronbach's Alpha	rho_A	Composite Reliability	Average Variance Extracted (AVE)
NASA - Subject Response	0.618	0.75	0.625	0.389
NASA - Task Demand	0.699	0.747	0.702	0.455

The internal consistency measures using the Cronbach alpha are greater than 0.6 across all the two identified mental workload dimensions. Likewise, the composite reliability shows that the obtained values are greater than 0.6, suggesting that the identified construct is satisfactory for capturing an individual's NASA scale.

However, the average variance extracted by the items for each construct was 38.9% for the subject response, and 45.5% explained variance by the task demand. These values are below the ideal average variance extracted of 50%. Finally, we have tested the discriminant validity of the two factors measured. The results are displayed in Table 66:

Table 66: Discriminant Validity for the NASA-TLX Scales

	Fornell-Larcker Criterion		Heterotrait-Monotrait Ratio (HTMT)	
	NASA - Subject Response	NASA - Task Demand	NASA - Subject Response	NASA - Task Demand
NASA - Subject Response	0.624			
NASA - Task Demand	0.886	0.674	0.857	

The results indicate a high correlation between subject response dimensions and task demand. One of the criteria (Fornell-Larcker indicates that the average extracted variance is lower than the shared variance, suggesting a unidimensional model. Otherwise, the HTMT ratio is lower than the 1.0 recommendation in SmartPLS, establishing discriminant validity.

We tested the second theorized model for NASA-TLX as a single construct to measure the mental workload. Both two-dimensional (task and demand on the subject) and single-dimensional (mental workload) models are widely employed and supported by the literature. The results for the single model indicated that the Cronbach's Alpha was 0.760 for all six items.

Finally, we executed the reliability analysis procedure in SPSS to determine the instrument's stability and any problematic items by using the option available to calculate multiple reliabilities alternatives "if items are deleted" from the scale and its impact on the instrument. The results are available in Table 67:

Table 67: Cronbach's Alpha Reliability – NASA-TLX Single Dimension

Item-Total Statistics	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Mental Demand - How mentally demanding was the simulation?	16	31.609	0.568	0.71
Physical Demand - How physically demanding was the task?	18.03	31.097	0.387	0.761
Temporal Demand - How hurried or rushed was the pace of the simulation?	16.89	28.04	0.67	0.676
Performance - How successful were you in accomplishing what you were asked to do?	17.1	31.733	0.377	0.761
Effort - How hard did you have to work to accomplish your level of performance?	16.55	29.798	0.666	0.684
Frustration - How insecure, discouraged, irritated, stressed, and annoyed were you?	18.07	33.679	0.409	0.747

The results showed that the instrument as a single-dimensional construct is reliable, even with modifications in the instrument with the Cronbach's Alfa potentially ranging from (0.676 to 0.747). In summary, our scales were determined to be reliable and with evidence of validity, including discriminant validity. Therefore, we generated the factors using factor weights determined during the confirmatory factor analysis.

However, for the variables related to the NASA instrument, given the established usage of the instrument as a simple sum of the items, in the NASA-TLX-RAW and the

reporting of such model, we choose to maintain this method for computing these variables. The resulting structure method for item aggregation for each scale is presented in Table 68:

Table 68: Final Constructs Loadings and computations for Study's variables

Scale	Resulting Structure	Aggregation Method
Emotion State (ESS)	Three-dimensional model	Confirmatory Factor Analysis Weights
Appraisal Process Scale	Original Instrument Structure	Confirmatory Factor Analysis Weights
RAW NASA TLX	Original Instrument Structure	Simple sum of items composing the scale
NASA TLX DEMAND AND TASK (Excluding physical demand dimension)	Original Instrument Structure	Simple sum of items composing the scale (excluding the physical component)
Feedback Scale	Original Instrument Structure (excluded Situational Agency items)	Confirmatory Factor Analysis Weights
Perception of Fidelity	Original Instrument Structure	Confirmatory Factor Analysis Weights

5.8.Descriptive Statistics

In this section, we have described the responses in three blocks of analysis comprising the variables of feedback and fidelity, mental workload as measured by NASA TLX, and the items related to the appraisal process of emotions for a complete sample. In the sequence, we present the frequency of responses of the Likert-type items related to the Feedback and Fidelity dimensions:

Table 69: Descriptive information for the Feedback and Fidelity items

(FD) The scenario resembled a real-life situation		Strongly disagree	Disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
	N	3	3	0	28	54	46
	%	2.20%	2.20%	0%	20.90%	40.30%	34.30%
(FD) Real life factors, situation, and variables were built into the simulation scenario		Strongly disagree	Disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
	N	1	0	2	21	61	49
	%	0.70%	0%	1.50%	15.70%	45.50%	36.60%
(FB) Feedback provided has constructive		Strongly disagree	Disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
	N	0	0	4	5	33	92
	%	0%	0%	3.00%	3.70%	24.60%	68.70%
(FB) Feedback was provided in a timely manner		Strongly disagree	Disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
	N	0	0	4	1	34	95
	%	0%	0%	3.00%	0.70%	25.40%	70.90%
(FB) The simulation allowed me to analyze my own behavior and actions		Strongly disagree	Disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
	N	0	0	5	12	39	78
	%	0%	0%	3.70%	9.00%	29.10%	58.20%
(FB) There was an opportunity after the simulation to obtain guidance/feedback from the teacher in order to build knowledge to another level		Strongly disagree	Disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
	N	0	0	3	1	38	92
	%	0%	0%	2.20%	0.70%	28.40%	68.70%

Note: FD – Perception of Fidelity , FB - Feedback

The perception of the quality of the feedback received was very positive. We have aggregated the responses between “Agree” and “Strongly agree” to illustrate the agreement with the statements.

The “Feedback provided has constructive” for 93% of the respondents, while 96% felt that the “Feedback was provided in a timely manner”. More than % 87 of respondents agreed that “The simulation allowed me to analyze my own behavior and actions” and finally, there was a significant amount of agreement over “There was an opportunity after the simulation to obtain guidance/feedback from the teacher in order to build knowledge to another level” with 97% agreement.

Regarding the participant’s perception of fidelity or realism of the simulation, 74.60% of respondents indicated that the “scenario resembled a real-life situation”. In comparison, 82.10 % indicated that they had recognized that “Real life factors, situation, and variables were built into the simulation scenario”.

The next block of questions concerned the NASA-TLX scale to measure workload. This scale had 7 points ranging from different anchor points, it would range from “very low” to “very high”, while the points for performance were “good” and “poor”. The results are available in Table 70 :

Table 70: Descriptive information for the NASA-TLX Scale

Mental Demand - How mentally demanding was the simulation?		Very Low	2	3	4	5	6	Very High
	N	4	9	13	36	39	24	9
	%	3.00%	6.70%	9.70%	26.90%	29.10%	17.90%	6.70%
Physical Demand - How physically demanding was the task?		Very Low	2	3	4	5	6	Very High
	N	66	17	11	15	14	5	6
	%	49.30%	12.70%	8.20%	11.20%	10.40%	3.70%	4.50%
Temporal Demand - How hurried or rushed was the pace of the simulation?		Very Low	2	3	4	5	6	Very High
	N	16	21	25	31	21	14	6
	%	11.90%	15.70%	18.70%	23.10%	15.70%	10.40%	4.50%
Performance - How successful were you in accomplishing what you were asked to do?		Good	2	3	4	5	6	Poor
	N	30	19	15	24	29	14	3
	%	22.40%	14.20%	11.20%	17.90%	21.60%	10.40%	2.20%
Effort - How hard did you have to work to accomplish your level of performance?		Very Low	2	3	4	5	6	Very High
	N	7	18	22	33	36	14	4
	%	5.20%	13.40%	16.40%	24.60%	26.90%	10.40%	3.00%
Frustration - How insecure, discouraged, irritated, stressed, and annoyed were you?		Very Low	2	3	4	5	6	Very High
	N	47	29	26	19	9	3	1
	%	35.10%	21.60%	19.40%	14.20%	6.70%	2.20%	0.70%

For the dimensions of mental workload, we have aggregated the responses to 5 to 7 as “high” and responses from 1 to 3 as “low”, and 1 to 3 as “Good” and 5 to 7 as “Poor” for the performance item, to facilitate the description.

For the first dimension, mental demand, 54% considered the mental demand high, while 20% considered the demand low. For the physical demand dimension, most respondents (70%) considered a low physical demand, while 18% considered the demand high. The temporal demand was also considered low by 46% of participants, while 30% believed it was hurried or rushed. Still, 48% considered their performance good, while 34% self-evaluated their performance as poor.

Participants agreed they had to work hard to accomplish their objectives (40%). A smaller number of respondents reported a low effort required (35%) to complete the simulation scenarios. Finally, many respondents reported low frustration with the simulation experience (76%), and 10% reported high frustration during the task.

Finally, we have conducted a frequency analysis of the distribution of responses regarding the emotional appraisal process for the dimensions of pleasantness, goal congruence, certainty, novelty, predictability, situational control, importance, and fairness. Results are available in Table 71:

Table 71: Descriptive information for the Emotional Appraisal Process

Variables	not at all enjoyable	2	3	4	5	6	extremely enjoyable
(P) How enjoyable was your overall experience ?	N 1	3	4	18	49	37	22
	% 0.70%	2.20%	3.00%	13.40%	36.60%	27.60%	16.40%
	not at all pleasant	2	3	4	5	6	extremely pleasant
(P) How pleasant or unpleasant was your overall experience ?	N 2	5	6	17	47	39	18
	% 1.50%	3.70%	4.50%	12.70%	35.10%	29.10%	13.40%
	very much inconsistent	2	3	4	5	6	very much consistent
(GC) To what extent was the overall experience consistent to what you wanted to achieve during this session?	N 1	2	7	24	44	35	21
	% 0.70%	1.50%	5.20%	17.90%	32.80%	26.10%	15.70%
	very much inconsistent	2	3	4	5	6	very much consistent
(GC) In comparison to what you desired, how consistent was your overall experience ?	N 2	2	10	22	47	30	21
	% 1.50%	1.50%	7.50%	16.40%	35.10%	22.40%	15.70%
	not important	2	3	4	5	6	extremely important
(GC) To what extent do you see the overall experience in this session important or not important in achieving your learning goals or needs	N 0	2	8	21	38	45	20
	% 0.00%	1.50%	6.00%	15.70%	28.40%	33.60%	14.90%
	no unders.	2	3	4	5	6	great deal of understanding
(CT) How well did you understand what was happening around you?	N 0	2	10	22	41	37	22
	% 0.00%	1.50%	7.50%	16.40%	30.60%	27.60%	16.40%
	not sure at all	2	3	4	5	6	extremely sure
(CT) How sure were you about what was happening?	N 1	1	15	22	51	30	14
	% 0.70%	0.70%	11.20 %	16.40%	38.10%	22.40%	10.40%
	unexpected	2	3	4	5	6	expected
(NV)To what extent did the overall experience was something you expected or unexpected?	N 11	27	39	34	16	4	3
	% 8.20%	20.10%	29.10 %	25.40%	11.90%	3.00%	2.20%
	unusual	2	3	4	5	6	usual
	N 11	26	45	38	10	1	3

(NV) How usual or unusual was your overall experience ?	%	8.20%	19.40%	33.60 %	28.40%	7.50%	0.70%	2.20%
(NV) To what extent was the overall experience you had in the simulation session familiar or unfamiliar ?	unfamiliar		2	3	4	5	6	familiar
	N	8	23	47	35	13	5	3
	%	6.00%	17.20%	35.10 %	26.10%	9.70%	3.70%	2.20%
(PD) How well could you predict what was going to happen?			Not	2	3	4	5	Extremely
	N		10	16	27	50	27	4
	%		7.50%	11.90 %	20.10%	37.30%	20.10%	3.00%
(SC)- To what extent did you feel that circumstances beyond anyone's control determined the outcomes/results of this simulation?			Not	2	3	4	5	Extremely
	N		17	39	40	28	9	1
	%		12.70%	29.10 %	29.90%	20.90%	6.70%	0.70%
(SC) - How responsible did you think someone or something other than yourself was for the outcomes/results of the simulation session?			Not	2	3	4	5	Extremely
	N		2	20	37	35	32	8
	%		1.50%	14.90 %	27.60%	26.10%	23.90%	6.00%
(SC) When you were in this session, how powerful (dominant) did you feel?			Not	2	3	4	5	Extremely
	N		8	27	45	36	16	2
	%		6.00%	20.10 %	33.60%	26.90%	11.90%	1.50%
(SC) When you were in this simulation session, how helpless did you feel?			Extrem.	2	3	4	5	Not at ALL (I)
	N		2	12	21	39	43	17
	%		1.50%	9.00%	15.70%	29.10%	32.10%	12.70%
(SC) To what extent did you feel that you could influence what was happening in the simulation session?			Not	2	3	4	5	Extremely
	N		0	10	27	47	37	13
	%		0%	7.50%	20.10%	35.10%	27.60%	9.70%
(IM) How important was this simulation session to you?	Not at all		2	3	4	5	6	Extremely
	N	1	8	16	45	50	1	13
	%	0.70%	5.90%	11.90 %	33.60%	37.30%	0.70%	9.70%
(FN) To what extent did you feel cheated or wronged in this simulation session ?			Extrem.	2	3	4	5	Not at ALL
	N		1	3	11	14	13	92
	%		0.70%	2.20%	8.20%	10.40%	9.70%	68.70%
(FN) How fair did you think what happened to you in this simulation session was?			Not	2	3	4	5	Extremely
	N		3	3	16	28	45	39
	%		2.20%	2.20%	11.90%	20.90%	33.60%	29.10%

Note: P – Pleasantness, GC – Goal Congruence, CT- Certainty, NV – Novelty, PD- predictability, SC – Situational control, IM – Importance, FN - Fairness

The items were combined into the same dimensions to provide an overview of responses according to each dimension, with items from 5 to 7 in the Likert Scale considered agreeable with the statement. In contrast, responses 1 to 3 were considered to disagree with the main statement.

Most participants related that their experience was pleasant (79%), with 8% reporting a non-enjoyable experience. A similar pattern emerged when considering the congruence with goals, with 74% reporting that the experience was compatible with their goals (9% reported not being consistent). Among the respondents, 74% reported being sure of what they were doing, while 10 % reported the contrary.

In terms of novelty, results were less defined. 59% reported that the experience was familiar, unusual, or unexpected, while 14% reported the situation as familiar, usual, and expected. The remaining 28% chose the mid-point of the scale. 60% reported being able to predict what would happen during the experience.

For the situational control items, on average, participants agreed on feeling in control and dominant during the experience (62%), while 14% of participants felt less in control, hopeless, or could not influence the outcome of the results. Regarding the attributed importance, 47% considered the simulation important to them, with a smaller group not finding it important (12%).

Finally, the last dimension was fairness perception. Most participants believe that the outcomes of the simulation session were fair (86%), while only 3% reported feeling that the outcomes were not fair given their participation and efforts.

In general, we can assert that the simulation sessions had good fidelity and realism and were satisfied with the feedback received during the debriefing. The results show a balance in mental workload, with a sizeable number defining it as low or high, perhaps depending on other factors, such as their prior experiences and class standing. The only dimension that did not show the same pattern was the physical demand, considered low by the majority.

The dimension of emotional appraisal was overly classified as positive as well. Overall, participants found the experience enjoyable, aligned with their goals, novel, and fair, with less situational control and novelty.

5.9. Hypothesis Tests

In this section, we will demonstrate the process of testing the eight hypotheses of this study. The following items have been compiled into a table for each hypothesis to provide a detailed and independent description of the tests: the research question, the researcher objectives, the formal *null* and alternative hypothesis, experimental design, variables involved, statistical analysis performed to test the hypothesis and the rejection criteria.

5.9.1. Non-inferiority margins for hypothesis testing

In the sequence, we present the margins for the hypothesis tested for non-inferiority (Hypothesis 3,4,6,7, and 8). In the case of a non-inferiority test, the alternative hypothesis is that the difference in treatments is less than the margin (Δ).

As described in the literature review chapter dedicated to explaining the concept of non-inferiority studies, the definition of a non-inferiority margin is a fundamental aspect of the non-inferiority approach. We specify that the alternative treatment is not inferior to the established alternative by that margin by defining a margin. Defining the margin impacts the results of the tests and must be reasonable and based on clinical or practical differences.

Put in another way, using a margin allows us to move from testing if a difference between two groups is equal to zero (potentially meaningless) to verify if a difference is more prominent than quantity with clinical or practical relevance.

After carefully considering prior literature and the implications of different margins, we choose to replicate Berg & Steinsbekk's defined margin of 13% in our study. Their study, named "Is individual practice in an immersive and interactive virtual reality application non-inferior to practicing with traditional equipment in learning systematic clinical observation? A randomized controlled trial" was published in The Bmc Medical Education, part of the Springer Nature (ISSN: 1472-6920), the most similar article found to our study.

Additionally, Berg & Steinsbekk (2020) cited that their analysis indicated that prior authors studying clinical learning outcomes have demonstrated that a non-inferiority limit between 10-15% is fair.

Descriptive statistics were generated for the variables to be compared with results limited to the in-person sample (n=72), and the margin of 13% was applied, as defined in

a previous study Berg & Steinsbekk (2020). The test for non-inferiority was conducted in NCSS 2021 using the Two Sample-T test for non-inferiority.

We conducted the non-inferiority test using the two-sample t-test and the specified margin for the variable against the in-person group. Generally, our non-inferiority assumption considered that higher values are better (higher emotional response, higher demand on participants as a positive outcome – except for arousal response in a debriefing where the deactivation was considered a better outcome).

We employed three different statistical approaches available in the NCSS package to test for non-inferiority in the presence of:

a) data normality and equal variances between groups (Equal-Variance T-Test for Non-Inferiority)

b) data normality and different variances between groups (Aspin-Welch Unequal-Variance T-Test for Non-Inferiority)

c) non-parametric assumptions (Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority)

All three statistical tests were presented for all the variables tested, and the final determination of which test to use depended on the variable data distribution. Results of the descriptive analysis, the resulting calculated margins, and the directionality of the non-inferiority are presented in Table 72:

Table 72: Variables used in the Hypothesis testing and associated non-inferiority margins

Variables	N	Mean	Std. Deviation	Margin Δ (13%)	Type of non-inferiority
Arousal - Post-Simulation (simulation-baseline)	72	0.8194	1.45666	0.107	Higher is better
Arousal - Post-Debriefing (post-debriefing-post-baseline)	72	-0.5694	1.36171	-0.074	Lower is better
Cheerfulness Response - Simulation	72	0.0904	0.63024	0.012	Higher is better
Cheerfulness Response - Debriefing	72	0.0698	0.61936	0.009	Higher is better
Invigoration Response - Simulation	72	0.7519	0.81689	0.098	Higher is better
Invigoration Response - Debriefing	72	-0.4406	0.82835	-0.057	Higher is better
Tranquility Response – Simulation	72	-0.2327	0.95729	-0.030	Higher is better
Tranquility Response – Debriefing	72	0.6715	1.10409	0.087	Higher is better
Pleasantness	72	5.7468	1.23727	0.747	Higher is better
Goal Congruence	72	5.854	1.27379	0.761	Higher is better
Certainty	72	5.5596	1.16922	0.723	Higher is better
Novelty	72	3.8615	1.1822	0.502	Higher is better
Predictability	72	3.4722	1.29975	0.451	Higher is better
Situational Control	72	4.5208	1.10649	0.588	Higher is better
Appraisal- Importance	72	4.3394	1.04919	0.564	Higher is better
Perception of Fidelity	72	5.539	0.69344	0.720	Higher is better
Level of Presence 0 – 100	72	70.4583	22.61928	9.160	Higher is better
Feedback Quality	72	6.0928	0.69468	0.792	Higher is better
NASA-TLX Raw Score	72	21.2917	7.14131	2.768	Higher is better
NASA-TLX Raw Score (excl. physical demand)	72	18.0278	5.84054	2.344	Higher is better
NASA-TLX Task Demand	72	11.5417	4.3925	1.500	Higher is better
NASA-TLX Task Demand (excludes physical demand)	72	8.2778	2.88377	1.076	Higher is better
NASA-TLX Demand on Subjects	72	9.75	3.46715	1.268	Higher is better
Valid N (listwise)	72				

For all the tests that failed to reject the non-inferiority, we employed the two-sided differences test to detect whether the groups differ in the variables. When differences between groups were found, we conducted one-sided tests to test whether the mean of the in-person simulation in the given variable was significantly greater than telesimulation.

These analyses are not part of the hypothesis testing and were performed as post hoc tests to support the discussion of results and implications of the inability to reject the hypothesis of non-inferiority of telesimulation.

5.9.2. Hypothesis 1

The first hypothesis of this study tested whether the in-person simulation in emotional response would evoke emotional responses, using a pre-posttest approach starting with the baseline values collected before the simulation (t0) and assessing for changes in t1 (post-simulation) and t2 (post debriefing).

The ANOVA Repeated measures is a case of paired-t samples for more than two comparisons. In this and the following sections, we will use the ANOVA Repeated measure to test for differences between the three moments in the research (baseline, post-simulation, and pre-simulation).

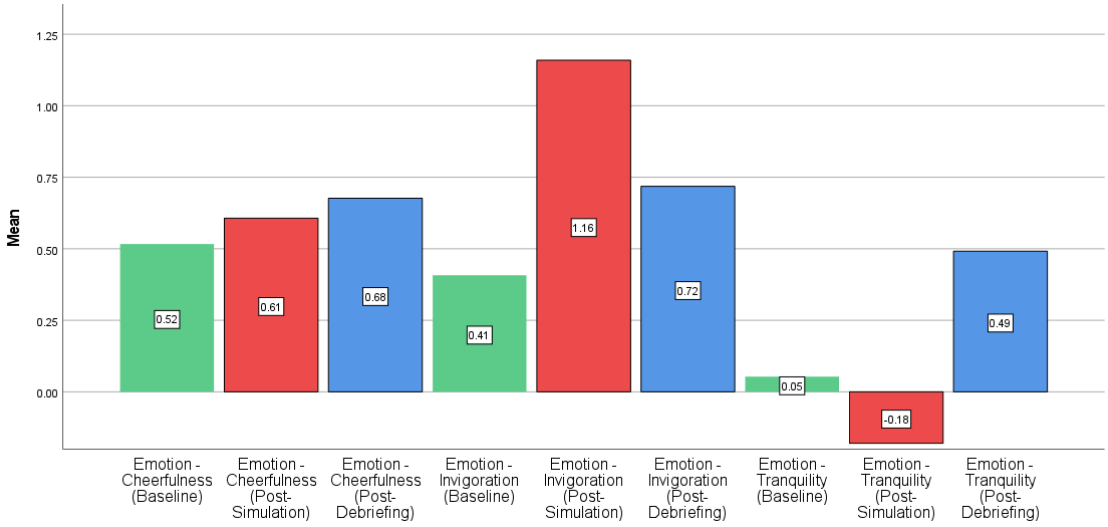
The emotional response was defined as a statistically significant change in the scores of the constructs related to emotional state (greater than zero). Details of the hypothesis, including the tests performed and additional information, is presented in Table 73:

Table 73: Summary of included variables and testing procedures for Hypothesis 1

Research Question: Does in-person simulation evoke emotional responses between baseline, post-simulation, and post-debriefing stages?
Objective: Demonstrate that in-person simulation evokes an emotional response between the different simulation moments.
Hypothesis 1
H1 **(null)** – In-person simulation **does not** evoke an emotional response during simulation and debriefing
Alternative hypothesis - In-person simulation **evokes** an emotional response during simulation and debriefing
Sub-hypotheses
H1a **(null)**: In-person simulation **does not evoke** an emotional state response during simulation and debriefing (time factor) as measured by the Emotional State Scale (ESS)
Alternative Hypothesis - In-person simulation **evokes** an emotional state response during simulation (time factor) as measured by the Emotional State Scale (ESS)
H1b **(null)**: In-person simulation **does not evoke** an arousal response emotional state response during simulation (time factor) as measured by Felt Arousal Scale (FAS)
Alternative Hypothesis - In-person simulation **evokes** an arousal response emotional state response during simulation and debriefing (time factor) as measured by Felt Arousal Scale (FAS)
Experimental Design: Pretest-Posttest Design
Statistical Analysis to test the hypothesis: Paired-t test / Repeated Measures ANOVA with Post-Hoc Tests
Independent variable (s): In-person effect (Treatment) on Simulation and Debriefing
Dependent variable(s): Emotional State Scale (ESS) and Felt Arousal Scale (ESS)
Rejection rule: Reject if p-value is less than 0.05 level of significance
One-tailed / two-tailed or non-inferiority: two-tailed

The first procedure to understand the dynamics of emotional response in the three moments and for the three dimensions was to plot the emotional dimensions graphically. Figure 57 has the three dimensions for each emotional dimension of the study:

Figure 57: Emotional Responses during Baseline, Simulation and Debriefing – In-person



The figure shows that emotional responses of cheerfulness are higher in post-debriefing than in post-simulation moments. We also observed that the post-simulation in the invigoration was higher than the invigoration from the post-debriefing. Likewise, the emotional responses based on tranquility were higher in the post-debriefing than in the post-simulation moment.

Next, we tested the induced emotional response to the in-person simulation by comparing the three testing moments (baseline period (t0), simulation (t1), and debriefing (t2) using ANOVA repeated measures for the three dimensions for the within-subjects factor:

Table 74: ANOVA – Repeated Measures (Cheerfulness, Invigoration and Tranquility) – In-person

Cheerfulness	**ANOVA - Within Subjects Effects - Cheerfulness**					
	Cases	Sum of Squares	df	Mean Square	F	p
	Time	0.93	2	0.465	2.084	0.128
	Residuals	31.669	142	0.223		
	Non-parametric tests					
	Friedman Test	Chi-Squared		df	p	Kendall's W
	Time	2.4		2	0.301	0.017
Invigoration	**ANOVA - Within Subjects Effects - INVIGORATION**					
	Cases	Sum of Squares	df	Mean Square	F	p
	Time	20.554	2	10.277	26.676	< .001
	Residuals	54.706	142	0.385		
	Non-parametric tests					
	Friedman Test	Chi-Squared		df	p	Kendall's W
	Time	40.075		2	< .001	0.278
Tranquility	**ANOVA - Within Subjects Effects - Tranquility**					
	Cases	Sum of Squares	df	Mean Square	F	P
	Time	16.742	2	8.371	16.475	< .001
	Residuals	72.153	142	0.508		
	Non-parametric tests					
	Friedman Test	Chi-Squared		df	p	Kendall's W
	Time	27.942		2	< .001	0.194

Note. Type III Sum of Squares

The one-way repeated measures ANOVA results show that emotional responses in baseline (M=0.517, SD=0.741), simulation (M=0.607, SD=0.627) on the cheerfulness construct was statistically insignificant (F(2,142)=2.084, p=0.128) at 5% level.

However, the result of the test for emotional responses based on the invigoration construct shows that the responses in the three moments were significantly different (F(2,142)=26.676, p <0.001), with the average score on the debriefing (M=0.677, SD=0.514), simulation (M=0.607, SD=0.627) and baseline (M=0.517, SD=0.741) being statistically significant among them.

Finally, the ANOVA result of the emotional responses based on the participants responses in the tranquility construct that there is significant difference (F(2,142)= 16.475, p <0.001) between the baseline (M=0.052;SD=0.882)simulation (M=-0180;SD=0.857) and debriefing (M=0.492;SD=0.698).

Following the ANOVA tests, we performed post hoc comparisons between the pairs of emotions in the three moments, namely: a) simulation – baseline (simulation response) b) debriefing – simulation (debriefing response), and c) debriefing – baseline (total emotional response). Results are available in Table 75:

Table 75: PostHoc comparisons – Cheerfulness, Invigoration, and Tranquility – Simulation and Debriefing in-person

	Comparison	Estimate	SE	df	t	P
Cheerfulness	Simulation - Baseline	0.09	0.08	142	1.15	0.252
	Debrief – Simulation	0.07	0.08	142	0.89	0.377
	Debriefing - Baseline	0.16	0.08	142	2.04	0.044
Invigoration	Simulation - Baseline	0.752	0.1	142	7.27	< .001
	Debrief – Simulation	-0.441	0.1	142	-4.26	< .001
	Debriefing - Baseline	0.311	0.1	142	3.01	0.003
Tranquility	Simulation - Baseline	-0.233	0.12	142	-1.96	0.052
	Debrief – Simulation	0.672	0.12	142	5.65	< .001
	Debriefing - Baseline	0.439	0.12	142	3.69	< .001

When comparing the pairs using the post hoc tests, initially for the cheerfulness dimension, the score on the debriefing (M=0.677, SD=0.514) was not significantly different from the simulation cheerfulness (M=0.607, SD=0.627). However, it was higher than the baseline (M=0.517, SD=0.741) ($t(142)= 7.27$, $p<0.001$).

The result indicates that the observed cheerfulness score as a measure of emotional responses in the in-person simulation from the start of the simulation session (baseline) and its completion (debriefing) experienced a significant increase, meaning the participants felt more cheerful after they participated in the simulation sessions. Even though statistically, the three measurements do not differ when considered as a whole before the post hoc test.

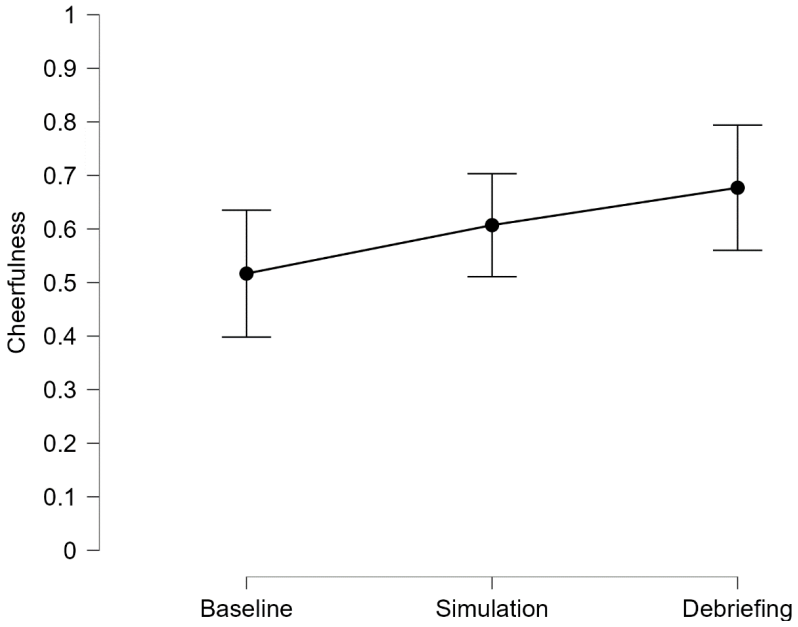
For the dimension of invigoration, post hoc tests showed that the invigoration during the simulation is (M=1.159; SD=0.794) significantly ($t(142)= -4.26$, $p<0.001$) greater than debriefing invigoration (M=0.719; SD=0.759) and likewise simulation invigoration is significantly ($t(142)= 7.27$, $p<0.001$) greater than baseline (M=0.407; SD=0.91).

The post hoc test also shows that the invigoration level during debriefing was significantly (t(142)=, p=0.003) different from the baseline. Participants felt more invigorated during the simulation, with their level of invigoration lowering during debriefing, although higher than in the baseline.

For the dimension of tranquility, the post hoc test showed that tranquility after debriefing (M=0.492; SD=0.698) was significant (t(142)= 5.65, p<0.001) higher than simulation (M=-0180;SD=0.857). We also found that debriefing score of tranquility was significantly (t(142)= 3.69, p<0.001) greater than baseline (M=0.052;SD=0.882). Finally, the level of tranquility during the simulation was significantly (t(142)= -1.96, p=0.003) lower than the baseline.

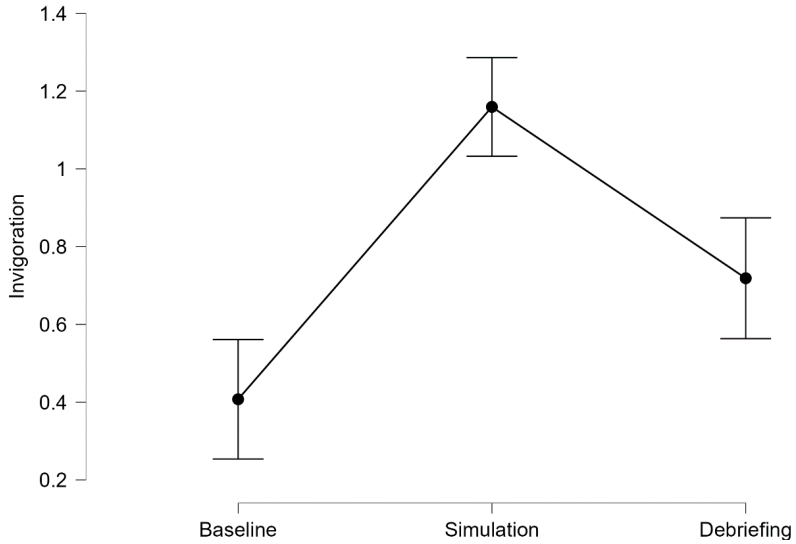
Participants felt more tranquil after participating in the in-person simulation, with a diminished level of tranquility after the simulation with a calming effect after the debriefing feedback. In the sequence, we present a series of Figures with the graphical representation of the three dimensions, plotted against the three moments of data collection:

Figure 58: Cheerfulness during Baseline, Simulation, and Debriefing – In-person



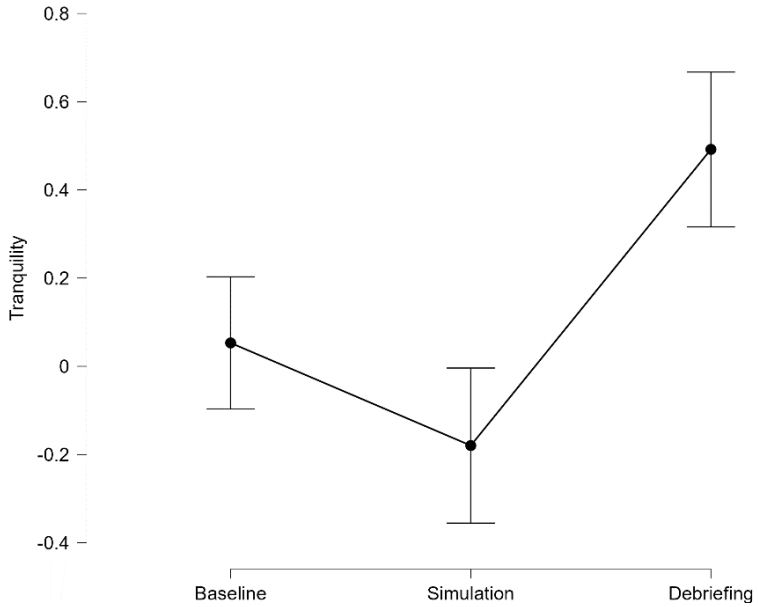
BARS represent CI = 95 %

Figure 59: Invigoration during Baseline, Simulation and Debriefing – In-person



BARS represent CI = 95 %

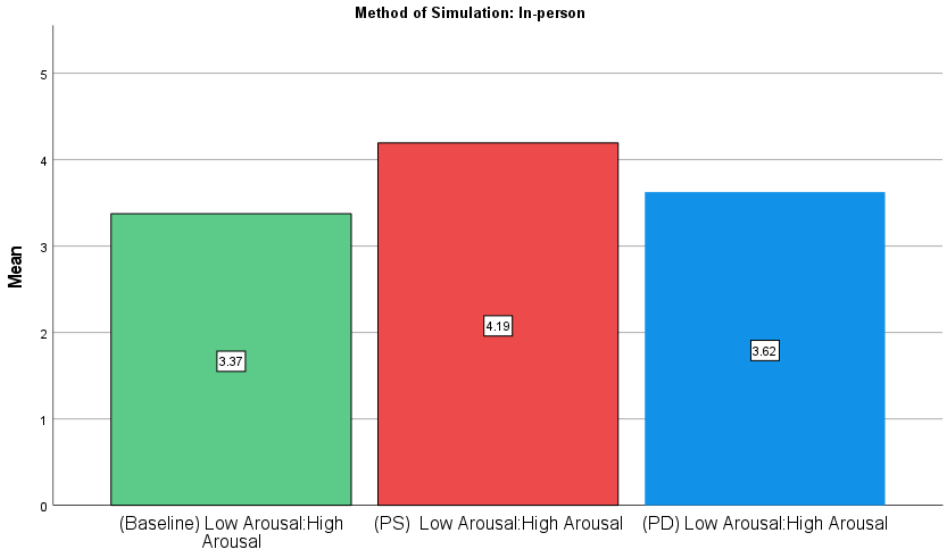
Figure 60: Tranquility during Baseline, Simulation and Debriefing – In-person



Note: Bars represent CI – 95 %

We performed the same analysis for the arousal response during simulation and debriefing in the sequence. Figure 60 presents the arousal scores during baseline, simulation, and debriefing for the in-person approach:

Figure 61: Arousal Scores during Baseline, Simulation and Debriefing – In-person



Next, we tested the arousal response on the in-person simulation by comparing the three testing moments (baseline period (t0), simulation (t1), and debriefing (t2) using ANOVA repeated measures for the three dimensions for the within-subjects factor:

Table 76: ANOVA – Repeated Measures - Arousal during Baseline, Simulation, and Debriefing – in-person

	ANOVA - Within Subjects Effects - Arousal					
	Cases	Sum of Squares	df	Mean Square	F	p
Arousal	Time	25.398	2	12.699	12.645	< .001
	Residuals	142.602	142	1.004		
	Non-parametric tests					
	Friedman Test	Chi-Squared		df	P	Kendall's W
	Time	29.636		2	< .001	0.206

Note. Type III Sum of Squares

The analysis shows that arousal was significantly different ($F(2,142)= 12.645$, $p<0.001$) between the three moments based on the repeated measure ANOVA carried out, with the scores on the baseline ($M=3.375$ $SD =1.156$). We have performed post hoc tests to compare the moments in the emotional dimensions. The results of the comparison are available in Table 77:

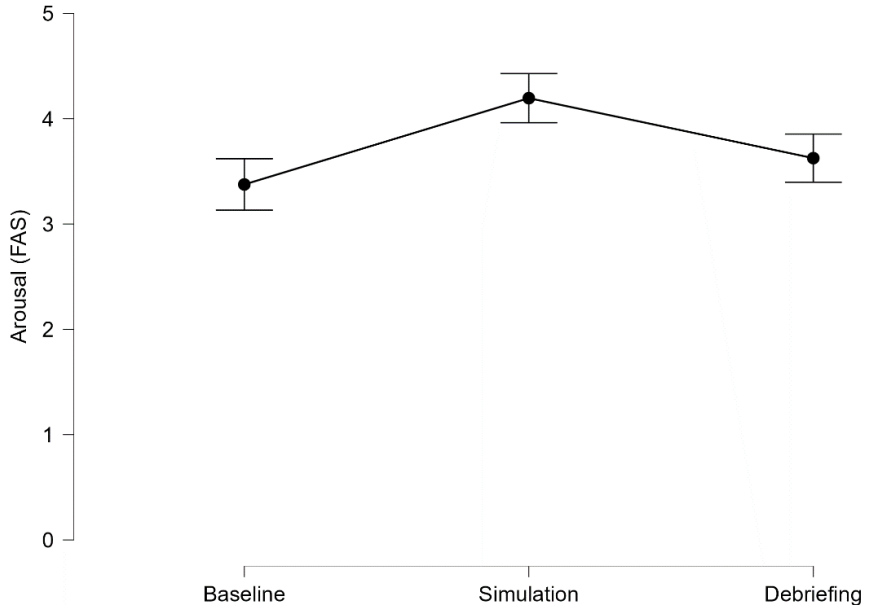
Table 77: Post-Hoc comparisons – Arousal – in-person

Comparison	Estimate	SE	df	T	p
Simulation - Baseline	0.819	0.167	142	4.906	< .001
Debrief - Simulation	-0.569	0.167	142	-3.409	< .001
Debriefing - Baseline	0.25	0.167	142	1.497	0.137

The post hoc test also shows that the arousal level during the simulation was significantly (t(142)=, p=0.003) different from the baseline. Likewise, the arousal level was significantly different during debriefing. The arousal response between the start and end of the simulation session is not significant, explained by the dynamics of arousal, which increased during the simulation and lowered during debriefing, bringing the felt arousal after debriefing to levels statistically equal to the baseline.

Following, we have created a graphical representation of the three dimensions, plotted against the three moments of data collection, available in Figure 62:

Figure 62: Arousal during Baseline, Simulation, and Debriefing – In-person



BARS represent CI = 95 %

Finally, we present a summary of the results for the test of the first hypothesis, the related statistical test (and p-value), followed by the decision:

Table 78: Summary of Results for Hypothesis 1

Hypothesis	Statistic test result (p-value)	Decision
H1 – In-person simulation does not evoke an emotional response during simulation or debriefing	Tested by the sub-hypothesis	Reject the hypothesis
H1a: In-person simulation **does not evoke** an emotional state response between baseline and post-simulation as measured by the Emotional State Scale (ESS)	Hypothesis rejected for tranquility ($p<0.001$) and invigoration ($p<0.001$)	Reject the hypothesis
H1b: In-person simulation **does not evoke** an arousal response between baseline and post-simulation as measured by Felt Arousal Scale (FAS)	$p< .001$	Reject the hypothesis

5.9.3. Hypothesis 2

The second hypothesis tested whether the telesimulation in emotional response would evoke emotional responses, using a pre-posttest approach starting with the baseline values collected before the simulation and assessing for changes in t1 (post-simulation) and t2 (post debriefing).

This hypothesis is virtually the same as the first hypothesis, with the difference being that it tested whether telesimulation would evoke emotional responses. Details of the hypothesis, including the tests performed and additional information, are presented in Table 79:

Table 79: Summary of included variables and testing procedures for Hypothesis 2

Research Question: Does telesimulation evoke emotional responses between baselines, post-simulation, and post-debriefing?
Objective: Demonstrate that telesimulation evokes an emotional response between the different simulation moments.
Hypothesis 2
H2 (**null**): Telesimulation **does not evoke** an emotional response during simulation and debriefing
Alternative Hypothesis - Telesimulation **evokes** an emotional response during simulation and debriefing
Sub-hypotheses
H2a**(null)**: Telesimulation **does not evoke** an emotional state response (ESS) during simulation and debriefing (time factor)
Alternative hypothesis - Telesimulation **evokes** an emotional state response (ESS) during simulation and debriefing (time factor)
H2b**(null)**: Telesimulation does not evokes an arousal response (FAS) between baseline and post-simulation
Alternative hypothesis - Telesimulation **evokes** an arousal response (FAS) between baseline and post-simulation
Experimental Design: Pretest-Posttest Design
Independent variable (s): Telesimulation effect (Treatment) on Simulation and Debriefing
Dependent variable(s): Emotional State Scale (ESS) and Felt Arousal Scale (ESS)
Statistical Analysis to test the hypothesis: Paired-t test / Repeated Measures ANOVA
One-tailed / two-tailed or non-inferiority: two-tailed
Rejection rule: Reject if p-value is less than 0.05 level of significance

The first procedure in this part is to understand the dynamics of emotional response in the three moments and the three dimensions in the telesimulation approach by plotting the emotional dimensions graphically:

Figure 63: Emotional Responses during Baseline, Simulation and Debriefing – Telesimulation

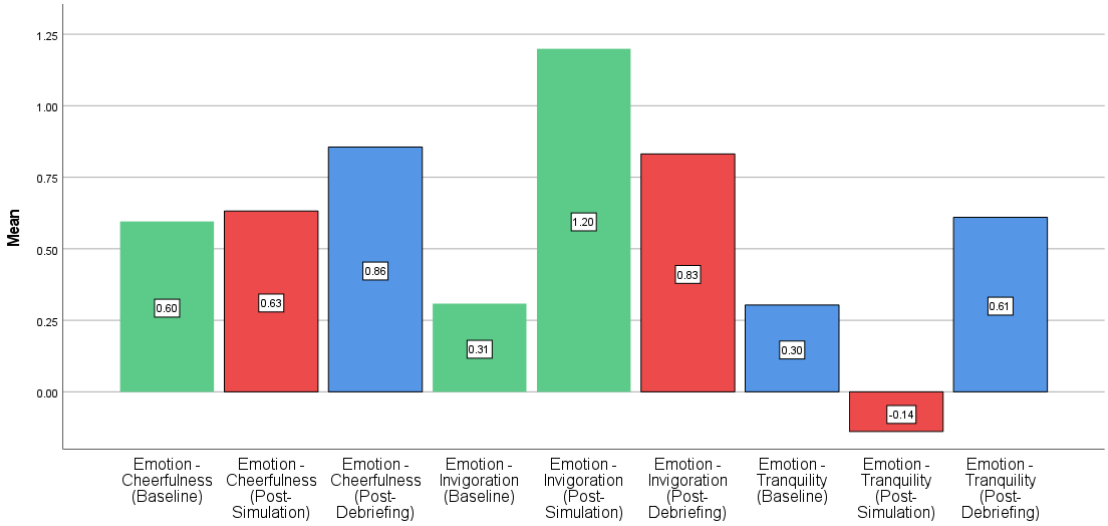


Figure 62 features the three emotional dimensions in telesimulation: the figure shows that emotional responses using cheerfulness are higher in post-debriefing than in the post-simulation moment. Post-simulation levels of the invigoration were higher than invigoration from the post-debriefing. Likewise, the emotional responses based on tranquility were higher in the post-debriefing than in the post-simulation moment.

Next, we tested the induced emotional response to the telesimulation by comparing the three testing moments (baseline period (t0), simulation (t1), and debriefing (t2) using ANOVA repeated measures for the three dimensions for the within-subjects factor:

Table 80: ANOVA – Repeated Measures (Cheerfulness, Invigoration and Tranquility) - Telesimulation

Cheerfulness	**ANOVA - Within Subjects Effects – Cheerfulness**					
	Cases	Sum of Squares	Df	Mean Square	F	p
	Time	2.454	2	1.227	5.878	0.004
	Residuals	25.469	122	0.209		
	Non-parametric tests					
	Friedman Test	Chi-Squared		df	p	Kendall's W
	Time	15.401		2	< .001	0.124
Invigoration	**ANOVA - Within Subjects Effects - INVIGORATION**					
	Cases	Sum of Squares	df	Mean Square	F	p
	Time	24.832	2	12.416	37.515	< .001
	Residuals	40.378	122	0.331		
	Non-parametric tests					
	Friedman Test	Chi-Squared		df	p	Kendall's W
	Time	53.66		2	< .001	0.433
Tranquility	**ANOVA - Within Subjects Effects - Tranquility**					
	Cases	Sum of Squares	df	Mean Square	F	p
	Time	17.553	2	8.776	21.669	< .001
	Residuals	49.412	122	0.405		
	Non-parametric tests					
	Friedman Test	Chi-Squared		df	p	Kendall's W
	Time	35.49		2	< .001	0.286

Note. Type III Sum of Squares

The result of the ANOVA shows that the emotional responses based on cheerfulness were significantly different (F(2,122)=5.878, p=0.004) between the evoked response from the telesimulation approach in the three moments, which are the baseline (M=0.60) simulation(M=0.63) and the debriefing (M=0.86).

Also, the result of the ANOVA test for emotional responses based on the invigoration construct shows that the responses on the three moments were significantly different (F(2,122)=37.515, p>0.001), with the average score on the simulation moment (M=1.20) higher than debriefing (M=0.83) and baseline (M=0.31).

Finally, the ANOVA result of the emotional responses based on the responses in the tranquility construct shows the existence of statistical differences (F(2,122)= 21.669, p<0.001) from the telesimulation approach between the baseline (M=0.30), simulation (M=-0.014) and debriefing (M=0.61).

Following the ANOVA tests, we performed post hoc comparisons between the pairs of emotions in the three moments, namely: a) simulation – baseline (simulation response), b) debriefing – simulation (debriefing response), and c) debriefing – baseline (total emotional response). Results are available in Table 81:

Table 81: Post Hoc comparisons – Cheerfulness, Invigoration and Tranquility – Simulation and Debriefing - Telesimulation

Variable	Comparison	Estimate	SE	Df	t	p
Cheerfulness	Simulation - Baseline	0.036	0.082	122	0.44	0.661
	Debrief - Simulation	0.224	0.082	122	2.725	0.007
	Debriefing - Baseline	0.26	0.082	122	3.165	0.002
Invigoration	Simulation - Baseline	0.89	0.103	122	8.618	< .001
	Debrief - Simulation	-0.367	0.103	122	-3.554	< .001
	Debriefing - Baseline	0.523	0.103	122	5.064	< .001
Tranquility	Simulation - Baseline	-0.442	0.114	142	-3.867	< .001
	Debrief – Simulation	0.748	0.114	142	6.547	< .001
	Debriefing - Baseline	0.306	0.114	142	2.681	0.008

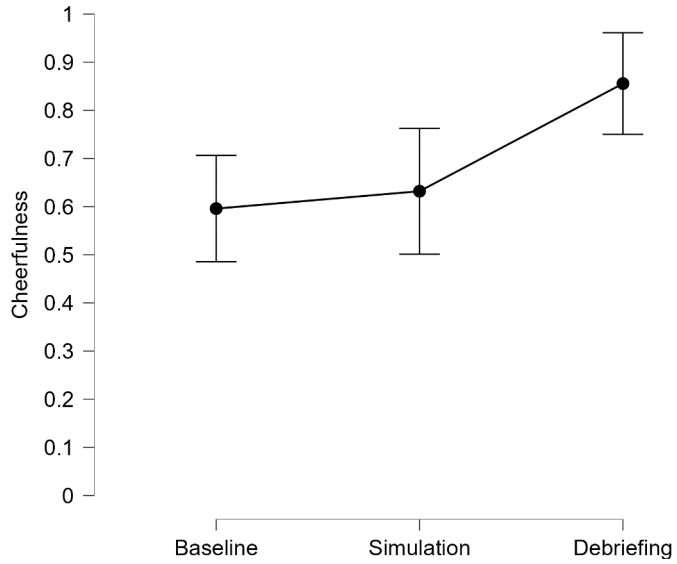
When comparing the pairs of post-simulation and post debriefing using the post hoc tests, the result shows that the cheerfulness dimension was not significantly ($t(122)=2.725$, $p=0.007$), different, with a mean difference of 0.224 ($se=0.082$).

The result indicates that the observed cheerfulness score as a measure of emotional responses in the telesimulation from the start of the simulation session (baseline) and its completion (debriefing) experienced a significant increase. Participants feel more cheerful after participating in the simulation sessions using the telesimulation approach.

For the invigoration using the telesimulation, post hoc tests showed that the invigoration responses from the post-simulation moment is significantly ($t(122)= -3.554$, $p<0.001$) different from the response on the invigoration from the post-debriefing moment with an average mean difference of -0.367 ($se=0.103$). That is, participants felt more invigorated during the post-simulation, with their level of invigoration lowering during debriefing, although higher than in the baseline.

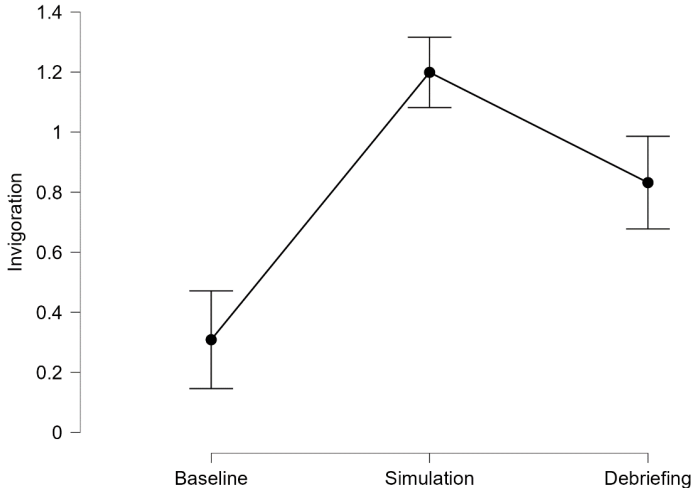
For the dimension of tranquility, the post hoc test showed that tranquility after debriefing was significant ($t(142)= 6.547$, $p<0.001$), different from the evoked responses at the post-simulation with the telesimulation approach. Implying that participants felt more tranquil after participating in the telesimulation, lowering tranquility post-simulation. In the sequence, we present a series of Figures with the graphical representation of the three dimensions, plotted against the three moments of data collection:

Figure 64: Emotional Responses during Baseline, Simulation and Debriefing – Telesimulation



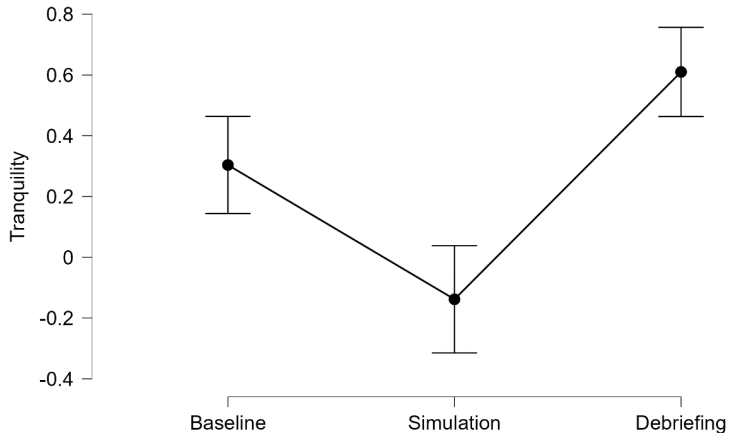
BARS represent CI = 95 %

Figure 65: Invigoration during Baseline, Simulation and Debriefing – telesimulation



BARS represent CI = 95 %

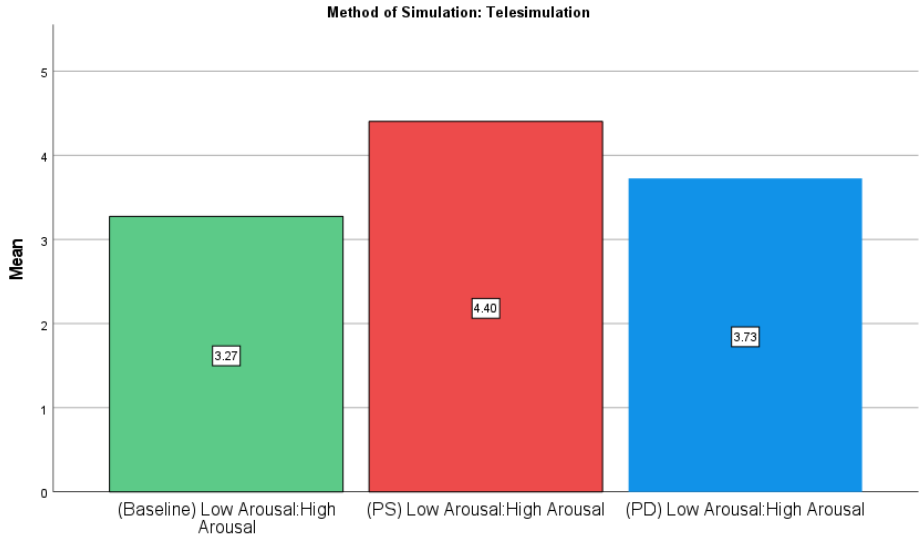
Figure 66: Tranquility during Baseline, Simulation and Debriefing – in-person



BARS represent CI = 95 %

We performed the same analysis for the arousal response during simulation and debriefing for the telesimulation approach in the sequence. Figure 67 presents the arousal scores during baseline, simulation, and debriefing for the telesimulation approach. The result shows that the arousal responses in the post-simulation moment were higher than arousal responses from the post-debriefing and the baseline moments.

Figure 67: Arousal during Baseline, Simulation, and Debriefing – telesimulation



The analysis shows that arousal was significantly different ($F(2,122)= 24.767$, $p<0.001$) between the three moments based on the repeated measure ANOVA carried out, with the scores of the arousal (on average) in the post-simulation ($M=4.403$, $SD=1.063$) greater than the scores from the post debriefing ($M=3.726$, $SD=1.089$) and the baseline

(M=3.274, SD =1.089). We have performed post hoc tests to compare the moments in the emotional dimensions. The results of the comparison are available in Table 82:

Table 82: ANOVA – Repeated Measures - Arousal during Baseline, Simulation, and Debriefing – Telesimulation

Arousal	**ANOVA - Within Subjects Effects – Arousal**					
	Cases	Sum of Squares	df	Mean Square	F	p
	Time	40.043	2	20.022	24.767	< .001
	Residuals	98.624	122	0.808		
	Non-parametric tests					
	Friedman Test	Chi-Squared		df	P	Kendall's W
	Time	35.202		2	< .001	0.284

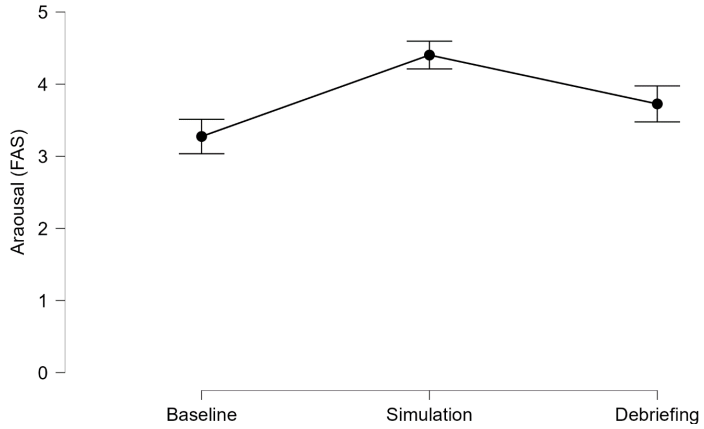
The post hoc test also shows that the arousal level during simulation and debriefing was significantly ($t(122)=-4.195$, $p<0.001$) different. Arousal response between the start and end of the simulation session is significantly explained by the arousal dynamics, which increased during the simulation and lowered during debriefing. The results are available in Table 83:

Table 83: Post Hoc comparisons – Arousal – Telesimulation

Comparison	Estimate	SE	df	T	p
Simulation - Baseline	1.129	0.161	122	6.992	< .001
Debrief - Simulation	-0.677	0.161	122	-4.195	< .001
Debriefing - Baseline	0.452	0.161	122	2.797	0.006

Following, we have created a graphical representation of the three dimensions, plotted against the three moments of data collection, available in Figure 68:

Figure 68: Arousal during Baseline, Simulation and Debriefing – telesimulation



BARS represent CI = 95 %

Table 84 presents the hypothesis and sub hypothesis tested during this section, along with the statistics and the decision based on the statistical tests:

Table 84: Summary of Results for Hypothesis 2

Hypothesis	Statistic test result (p-value)	Decision
H2 (**null**): Telesimulation **does not evoke** an emotional response during simulation or debriefing	Tested by the sub-hypothesis	Reject the hypothesis
H2a**(null)**: Telesimulation **does not evoke** an emotional state response (ESS) between baseline and post-simulation	Hypothesis rejected for Tranquility ($p<0.001$) and invigoration ($p<0.001$)	Reject the hypothesis
H2b**(null)**: Telesimulation does not evokes an arousal response (FAS) between baseline and post-simulation	$p< .001$	Reject the hypothesis
H2c **(null)**: Telesimulation **does not** evoke an emotional state (ESS) response between baseline (post-simulation) and post-debriefing	Hypothesis rejected for cheerfulness($p=0.007$) tranquility ($p<0.001$) and invigoration ($p<0.001$)	Reject the hypothesis
H2d **(null)**: Telesimulation **does not evoke** an arousal response (FAS) between baseline (post-simulation) and post-debriefing	$p< .001$	Reject the hypothesis

5.9.4. Hypothesis 3

The third hypothesis compared the emotional responses between telesimulation and in-person simulation, using a pre-posttest approach starting with the baseline values collected before the simulation and assessing for changes in t1 (post-simulation) – emotional response during the simulation. We also verified the changes between t1 (post-simulation) serving as the baseline and t2 (post debriefing) to estimate the debriefing response. We used a combination design using elements of the pre-post-test approach (within-subjects) and between-subjects design (non-equivalent group design).

This hypothesis employed the non-inferiority approach as defined in the literature review chapter. The objective is to test whether telesimulation is a non-inferior approach to in-person simulation to create emotional responses considering the pre-defined margin of 13% following previous studies' definition of margins to similar situations (Berg & Steinsbekk, 2020).

Details of the hypothesis, including the tests performed and additional information, is presented in Table 85:

Table 85: Summary of included variables and testing procedures for Hypothesis 3

Research Question: Are emotional responses in telesimulation non-inferior to in-person simulation?
Objective: Demonstrate that telesimulation is a non-inferior method to in-person simulation regarding emotional responses.
Hypothesis 3
H3 (**null**): The emotional response evoked by telesimulation is inferior to in-person simulation during simulation and debriefing
Alternative Hypothesis - The emotional response evoked by telesimulation is not inferior to in-person simulation during simulation and debriefing
Sub-hypotheses
H3a (**null**): Arousal response (FAS) between baseline and post-simulation **is inferior** in telesimulation compared to in-person simulation
Alternative hypothesis - Arousal response (FAS) between baseline and post-simulation **is not inferior** in telesimulation compared to in-person simulation
H3b: (**null**) Arousal response (FAS) between post-simulation and post-debrief **is inferior** in telesimulation compared to in-person debriefing
Alternative hypothesis - Arousal response (FAS) between post-simulation and post-debrief **is not inferior** in telesimulation compared to in-person debriefing
H3c: (**null**) Emotion state response (ESS) between baseline and post-simulation **is inferior** in telesimulation compared to in-person simulation
Alternative hypothesis - Emotion state response (ESS) between baseline and post-simulation **is not inferior** in telesimulation compared to in in-person
H3d: (**null**) Emotion state response (ESS) between post-simulation and post-debrief **is inferior** in telesimulation compared to in-person debriefing
H3d: Emotion state response (ESS) between post-simulation and post-debrief **is not inferior** in telesimulation compared to in-person debriefing
Alternative hypothesis - Emotion state response (ESS) between post-simulation and post-debrief **is not inferior** in telesimulation compared to in-person debriefing
Experimental Design: Combination Design (Pre-posttest and nonequivalent group design)
Independent variable (s): Method of simulation (in-person vs telesimulation)
Dependent variable(s): Emotional State Scale (ESS) and Felt Arousal Scale (ESS)
Statistical Analysis to test the hypothesis: Two Sample T-Test (Equal-Variance T-Test for Non-Inferiority, Aspin-Welch Unequal-Variance T-Test for Non-Inferiority and Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority)
One-tailed / two-tailed or non-inferiority comparison: non-inferiority (one-tailed)
Rejection rule: Reject if p-value is less than 0.05 level of significance

This section compared telesimulation and the in-person simulation approach based on the emotional and arousal response evoked. The first procedure to compare the two approaches was to graphically display the dynamics between the two approaches over the three moments of data collection. The results are available in Figure 69:

Figure 69: Telesimulation and in-person arousal during baseline, simulation, and debriefing

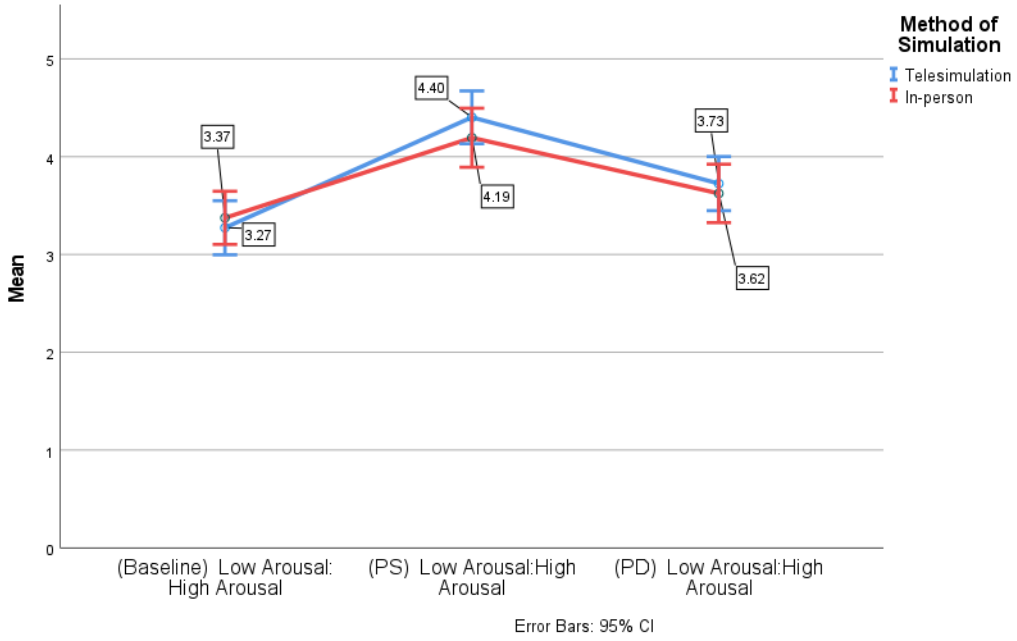


Figure 68 indicates that both approaches have very similar temporal dynamics with numerically close results, indicating plausible non-inferiority.

Emotional responses were operationalized by calculating the difference between baseline and simulation (simulation response) and between responses collected after simulation and after debriefing (debriefing response).

We have used three t-tests to provide more robust testing in the presence of non-normality (Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test) and non-equal variances between the groups compared (Aspin-Welch Unequal-Variance) in addition to the standard Two samples Equal-Variance T-Test for Non-Inferiority. All three options were implemented in the NCSS software version 2021.

We examined the arousal responses in the telesimulation against the in-person approach, having as the null hypothesis that the telesimulation is an inferior approach. The results of the tests are available in Table 86:

Table 86: Test for non-inferiority for Telesimulation – Arousal Response during Simulation

Test: Arousal Simulation		**NInf Hypothesis: (Telesimulation) > (In-person) - 0.107**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	1.1290	1.138	0.145	1.994	0.840	1.417
In-person	72	0.819	1.457	0.172	1.994	0.477	1.161
Confidence Intervals for the Mean Difference							
Variance Assumption		Df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	0.310	0.229	1.978	-0.142	0.762
Unequal		130.82	0.310	0.224	1.978	-0.134	0.753
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	Df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.107	0.309588	0.228	1.823	132	0.0353	Yes	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	Df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.107	0.309588	0.224	1.856	130.82	0.033	Yes	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		P	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.761			3.139		0.001	Yes	

Shapiro-Wilk 0.8937 0.00002 Yes
Modified-Levene=0.8004 p-0.37261

The normality test indicates a rejection of the normality assumption (W=0.893, p<0.001), suggesting the usage of the non-parametric version of the test. However, using the Levene we can maintain the assumption of equal variance between the two groups (F=0.8004, p=0.373).

All three tests for non-inferiority resulted significative. The t-test showed a significant result (t(132)=1.832, p=0.0353), implying the rejection of the inferiority null hypothesis. The same was observed with the Mann-Whitney U non-parametric test (z=3.139, p=0.001). The results indicate that the telesimulation is non-inferior to the in-person simulation in healthcare for the variable of arousal response during simulation.

Table 87 presents the non-inferiority tests employed for the arousal response during the debriefing:

Table 87: Test for non-inferiority for Telesimulation – Arousal Response during Debriefing

Test: Arousal Debriefing		**NInf Hypothesis: (Telesimulation) > (In-person) +0.074**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	-0.677	1.212	0.154	1.999	-0.985	-0.370
In-person	72	-0.569	1.362	0.160	1.993	-0.889	-0.249
Confidence Intervals for the Mean Difference							
Variance Assumption		Df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-0.108	0.224	1.978	-0.552	0.336
Unequal		131.85	-0.108	0.222	1.978	-0.548	0.332
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT < μC + 0.074	-0.108	0.224	-0.811	132	0.209	No	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT < μC + 0.074	-0.108	0.222	-0.818	131.85	0.207	No	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		P	Conclude Non-Inferiority at α = 0.05?	
LocT < LocC + 0.074			-2.584		0.0049	Yes	
Modified-Levene	0.6898	0.40772	No				
Shapiro-Wilk	0.9057	0.00017	Yes				

The distribution of the arousal during debriefing was not compatible with a normal distribution(W=0.9057; p<0.001); as such, a non-parametric counterpart of the t-test was employed to compare the telesimulation of the in-person simulation at the post-debriefing period.

The result supports the rejection (z= -2.584, p=0.0049) of the null hypothesis, indicating that the arousal obtained for the telesimulation is not inferior compared to the in-person (using the margin of 13%) debriefing. Therefore, telesimulation is non-inferior to the in-person simulation at the debriefing. After testing the non-inferiority of telesimulation on arousal in simulation and debriefing, the next set of comparisons was between the three dimensions of emotional response. In the subsequent Figures, we have

illustrated the temporal pattern of cheerfulness, invigoration, and tranquility, comparing the in-person and telesimulation.

Figure 70: Cheerfulness temporal pattern – Telesimulation vs In-Person

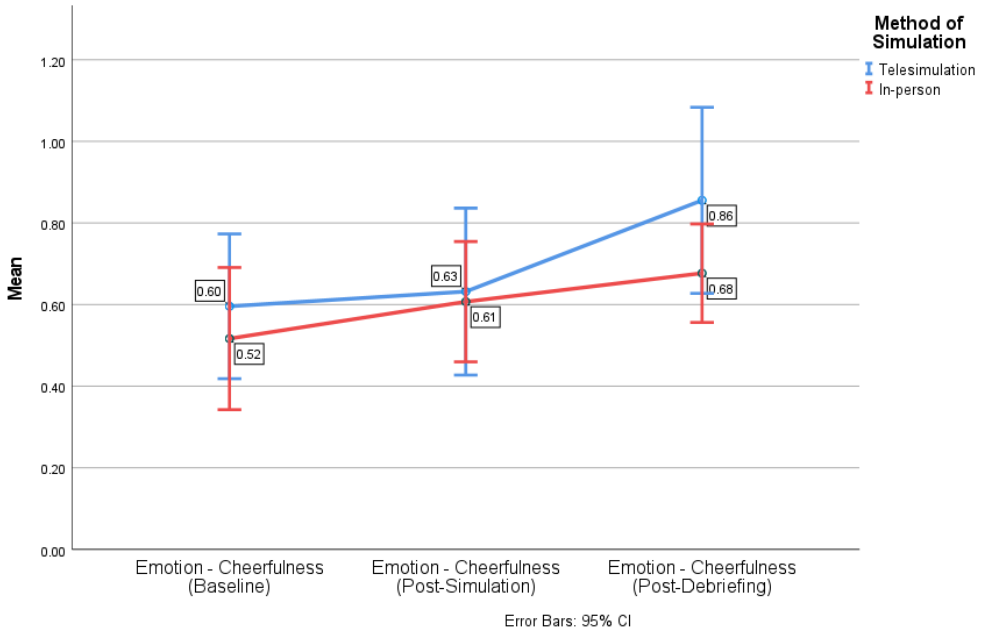


Figure 71: Invigoration temporal pattern – Telesimulation vs In-Person

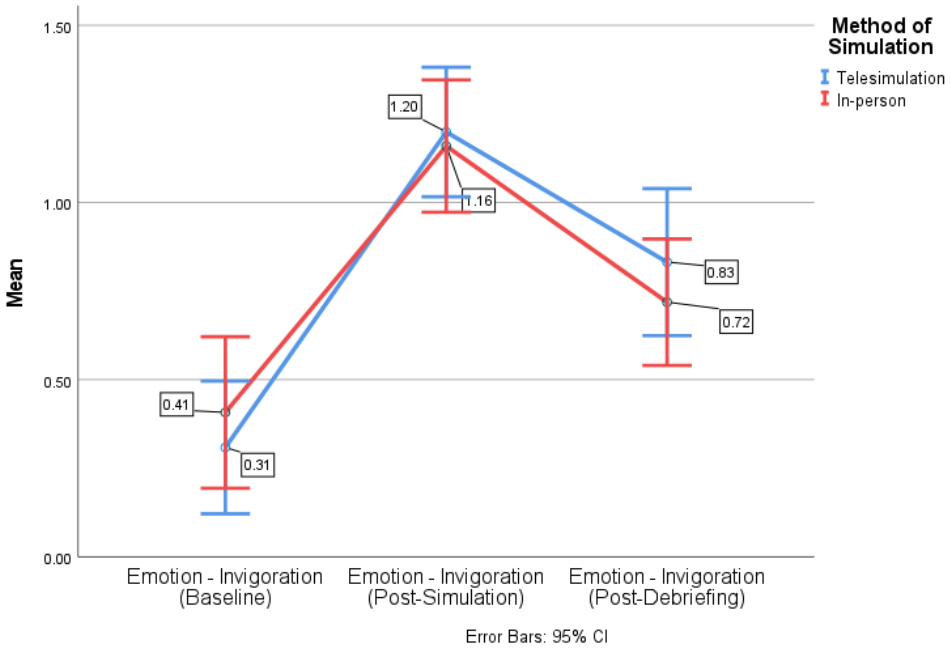
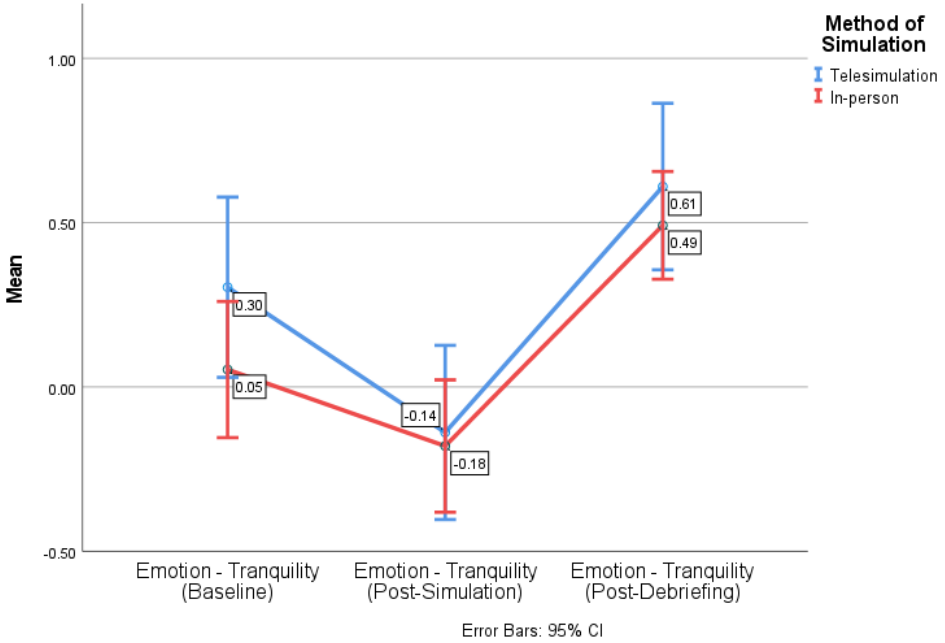


Figure 72: Tranquility temporal pattern – Telesimulation vs In-Person



For the second group of tests, we examined the emotional responses to the telesimulation against the in-person approach, having the null hypothesis that the telesimulation is an inferior approach. The results of the tests for the Cheerfulness during simulation are available in Table 88:

Table 88: Test for non-inferiority for Telesimulation – Cheerfulness response during Simulation

Test: Cheerfulness Simulation		**NInf Hypothesis: (Telesimulation) > (In-person) - 0.012**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	0.036	0.700	0.0889	2.000	-0.142	0.214
In-person	72	0.090	0.630	0.074	1.994	-0.058	0.238
Confidence Intervals for the Mean Difference							
Variance Assumption		Df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-0.054	0.115	1.978	-0.282	0.173
Unequal		124	-0.054	0.116	1.979	-0.284	0.175
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	P	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.012	-0.054	0.115	-0.368	132	0.643	No	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.012	-0.054	0.116	-0.366	124	0.642	No	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		P	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.012			2.090		0.018	Yes	
Modified-Levene	0.0012	0.97261	No				
Shapiro-Wilk	0.8216	0.00000	Yes				

The distribution of the variable cheerfulness during the simulation was not normal ($W=0.822$; $p<0.001$). Therefore a non-parametric counterpart of the t-test was selected as the primary test to determine the non-inferiority of telesimulation

The non-parametric test indicated the non-inferiority of telesimulation ($z=2.090$; $p=0.018$), with conflicting results for the non-inferiority tests compared to the other tests. The parametric t-test indicated that we could not reject the hypothesis of inferiority ($t(132)= -0.368$, $p=0.643$) between the two groups during the simulation for cheerfulness.

We can conclude that the telesimulation was non-inferior to person simulation because the non-parametric test employed is the most appropriate given the data distribution.

We performed the same test following the cheerfulness response during debriefing. The results are available in Table 89:

Table 89: Test for non-inferiority for Telesimulation – Cheerfulness response during Debriefing

Test: Cheerfulness Debriefing		NInf Hypothesis: (Telesimulation) > (In-person) - 0.009					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	0.224	0.676	0.086	2.000	0.052	0.395
In-person	72	0.070	0.619	0.073	1.994	-0.076	0.215
Confidence Intervals for the Mean Difference							
Variance Assumption		Df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	0.154	0.112	1.978	-0.068	0.375
Unequal		124.96	0.154	0.113	1.979	-0.069	0.377
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	P	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.009	0.154	0.112	1.454	132	0.074	No	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.009	0.154	0.113	1.445	124.96	0.075	No	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		P	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.009			3.219		0.0006	Yes	
Shapiro-Wilk	0.8623	0.00001	Yes				
Modified-Levene	1.0701	0.30281	No				

The normality of variable cheerfulness during debriefing was rejected (W=0.862; p<0.001), indicating that the non-parametric Mann-Whitney U was recommended to determine the non-inferiority of telesimulation.

The non-parametric test indicated the non-inferiority of telesimulation (z=3.219; p=0.0006). However, the parametric t-test results indicated that we could not reject the hypothesis of inferiority ((t(132)= 1.454, p= 0.074)) for cheerfulness.

Like in the previous tests, telesimulation was non-inferior to person simulation in their arousal response during the debriefing if using the non-parametric test. However, we failed to reject the same hypothesis when using a parametric test. We retained the

results of the non-parametric tests, given the presence of the non-normality of the variable. We tested the non-inferiority for the invigoration response during simulation in the sequence. The results are presented in Table 90:

Table 90: Test for non-inferiority for Telesimulation – Invigoration response during Simulation

Test: Invigoration Simulation		**NInf Hypothesis: (Telesimulation) > (In-person) - 0.098**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	0.890	0.764	0.097	1.9996	0.696	1.085
In-person	72	0.752	0.817	0.096	1.9939	0.560	0.943
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	0.138	0.137	1.978	-0.133	0.410
Unequal		131.06	0.138	0.137	1.978	-0.132	0.409
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.098	0.138	0.137	1.721	132	0.0438	Yes	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	Df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.098	0.138	0.137	1.730	131.06	0.043	Yes	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.098			3.0736		0.0011	Yes	
Modified-Levene	0.6308	0.42848	No				
Shapiro-Wilk	0.8926	0.00006	Yes				

The variable of invigoration during the simulation was non-normal (W=0.8925; p<0.001). However, all three tests found a significant result (t(132)=1.721, p=0.0438) for the parametric test). Moreover, the non-parametric tests (z=3.073;p=0.011) imply the rejection of the null hypothesis that telesimulation is inferior regarding the dimension of invigoration.

We conclude that the telesimulation is non-inferior to in-person simulation in the invigoration response. The same analysis was performed during the debriefing phase to determine the non-inferiority:

Table 91: Test for non-inferiority for Telesimulation – Invigoration response during Debriefing

Test: Invigoration Debriefing		**NInf Hypothesis: (Telesimulation) > (In-person) - 0.057**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	-0.367	0.709	0.090	1.999	-0.547	-0.187
In-person	72	-0.440	0.828	0.098	1.993	-0.635	-0.246
Confidence Intervals for the Mean Difference							
Variance Assumption		Df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	0.073	0.134	1.9781	-0.192	0.339
Unequal		132	0.073	0.132	1.9781	-0.189	0.336
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.057	0.0734	0.134	0.970	132	0.1668	No	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.057	0.0734	0.133	0.982	132	0.164	No	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.057			2.3071		0.0105	Yes	
Modified-Levene	2.649	0.106	No				
Shapiro-Wilk	0.8926	0.00006	Yes				

The Invigoration variable during debriefing was not normally distributed (W=0.892, p=0.006), although the Levene test of equal variance suggested equal variances (F=2.649, p=0.106), supporting the use of the Mann-Whitney U, given the non-normality of the variable.

The Mann-Whitney U non-parametric test supported the non-inferiority of telesimulation (z=2.3071; p=0.0105) for the invigoration during debriefing. However, we could not reject the hypothesis of inferiority ((t(132)= 0.970, p= 0.1668)) using the

parametric t-test. In summary, we rejected the non-inferiority of invigoration during debriefing, given the adequacy of the non-parametric test.

Next, we tested the non-inferiority of telesimulation for the tranquility dimension of the simulation. The results are present in Table 92:

Table 92: Test for non-inferiority for Telesimulation – Tranquility response during Simulation

Test: Tranquility Simulation		**NInf Hypothesis: (Telesimulation) > (In-person) - 0.03**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	-0.441	0.974	0.124	1.9996	-0.689	-0.194
In-person	72	-0.232	0.957	0.113	1.9939	-0.457	-0.007
Confidence Intervals for the Mean Difference							
Variance Assumption		Df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-0.209	0.167	1.9781	-0.540	0.122
Unequal		128.34	-0.209	0.167	1.9786	-0.540	0.122
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.03	-0.209	0.167	-1.072	132	0.8571	No	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	P	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.03	-0.209	0.167	-1.071	128.34	0.8568	No	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.03			0.1809		0.4282	No	
Shapiro-Wilk	0.9562	0.01364					
Modified-Levene	0.2481	0.61923	No				

Results indicated non-normal data distribution (W=0.956;p=0.134) and equal variances(F=0.24;p=0.62). The parametric test for non-inferiority of telesimulation in tranquility dimension (t(132)= -1.072, p=0.8571) was statistically insignificant, likewise the same conclusion was reached for the non-parametric test (z=0.181p=0.428) between the two groups during the simulation.

We could not establish that telesimulation was non-inferior regarding the tranquility dimension, therefore retaining the null hypothesis that telesimulation is inferior to in-person when evoking tranquility during simulation.

Table 93: Test for non-inferiority for Telesimulation – Tranquility response during Debriefing

Test: Tranquility Debriefing		**NInf Hypothesis: (Telesimulation) > (In-person) - 0.087**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	0.748	0.910	0.115	1.9996	0.517	0.980
In-person	72	0.671	1.104	0.130	1.9939	0.412	0.930
Confidence Intervals for the Mean Difference							
Variance Assumption		Df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	0.0769	0.176	1.9781	-0.272	0.426
Unequal		131.76	0.0769	0.174	1.9781	-0.267	0.421
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.087	0.077	0.176	0.928	132	0.177	No	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	P	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.087	0.077	0.174	0.942	131.76	0.174	No	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.087			1.552		0.0603	No	

Note: Modified-Levene 0.9712 0.32618 No

Likewise, tranquility tests were statistically insignificant (t(132)= 0.928, p=0.177). between the two groups during the debriefing, in all three non-inferiority tests. Therefore, we could not establish that telesimulation was non-inferior regarding the tranquility dimension, therefore retaining the null hypothesis that telesimulation is inferior to in-person when evoking tranquility during debriefing.

Table 94: Summary of Results for Hypothesis 3

	Statistic test result (p-value)	Decision
H3 (**null**): The emotional response evoked by telesimulation is inferior to in-person simulation during simulation and debriefing	Tested by subhypothesis	Reject the hypothesis (partially)
H3a (**null**): Arousal response (FAS) between baseline and post-simulation **is inferior** in telesimulation compared to in-person simulation	Arousal during simulation FAS (z=3.139, p=0.001)	Reject the hypothesis
H3b: (**null**) Emotion state response (ESS) between baseline and post-simulation **is inferior** in telesimulation compared to in-person simulation	Responses during simulation: Cheerfulness (z=2.090 ; p=0.018) Invigoration (z=3.073;p=0.011) Tranquility (t(132)= -1.072, p=0.8571)	Reject two hypotheses, retain one
H3c: (**null**) Arousal response (FAS) between post-simulation and post-debrief **is inferior** in telesimulation compared to in-person debriefing	Arousal during debriefing (z= -2.584 , p=0.0049)	Reject hypothesis
H3d: (**null**) Emotion state response (ESS) between post-simulation and post-debrief **is inferior** in telesimulation compared to in-person debriefing	Responses during debriefing: Cheerfulness ((z=3.219 ; p=0.0006) Invigoration ((z=2.3071; p=0.0105)) Tranquility (t(132)= 0.928, p=0.177).	Reject two hypotheses, retain one

Post Hoc Tests

Post hoc tests were conducted to compare the emotional responses between telesimulation and in-person simulation. Tests were executed to verify if the failure to reject the non-inferiority would be detected as a difference between the groups. The results are available in Table 95:

Table 95: Post Hoc tests for differences in emotional responses - Independent Samples T-Test Simulation and in-person

Variable	Test	Statistic	Df	p
Arousal Response (Simulation)	Student	1.355	132	0.178
	Mann-Whitney	2395.5		0.451
Arousal Response (Debriefing)	Student	-0.481	132	0.631
	Mann-Whitney	2160.5		0.744
Cheerfulness Response (Simulation)	Student	-0.473	132	0.637
	Mann-Whitney	2350		0.591
Invigoration Response (Simulation)	Student	1.008	132	0.315
	Mann-Whitney	2437.5		0.355
Invigoration Response (Debriefing)	Student	0.546	132	0.586
	Mann-Whitney	2345		0.607
Tranquility Response (Simulation)	Student	-1.251	132	0.213
	Mann-Whitney	1980		0.26
Tranquility Response (Debriefing)	Student	0.435	132	0.664
	Mann-Whitney	2320.5		0.694
Cheerfulness Response (Debriefing)	Student	1.374	132	0.172
	Mann-Whitney	2613		0.081

There were no significant differences for all the variables compared, except for the Mann-Whitney test for the debriefing response during debriefing. These further supports using a more robust and applicable testing approach for non-inferiority instead of assuming equality when the t-tests for comparison fail to reject the hypothesis of differences.

5.9.5. Hypothesis 4

The fourth hypothesis compared the mental workload between telesimulation and in-person simulation, using a nonequivalent groups design. The study in this aspect is a between-subjects study in which participants are not allocated to conditions at random, given that it is a natural experiment, as explained in the methodology section.

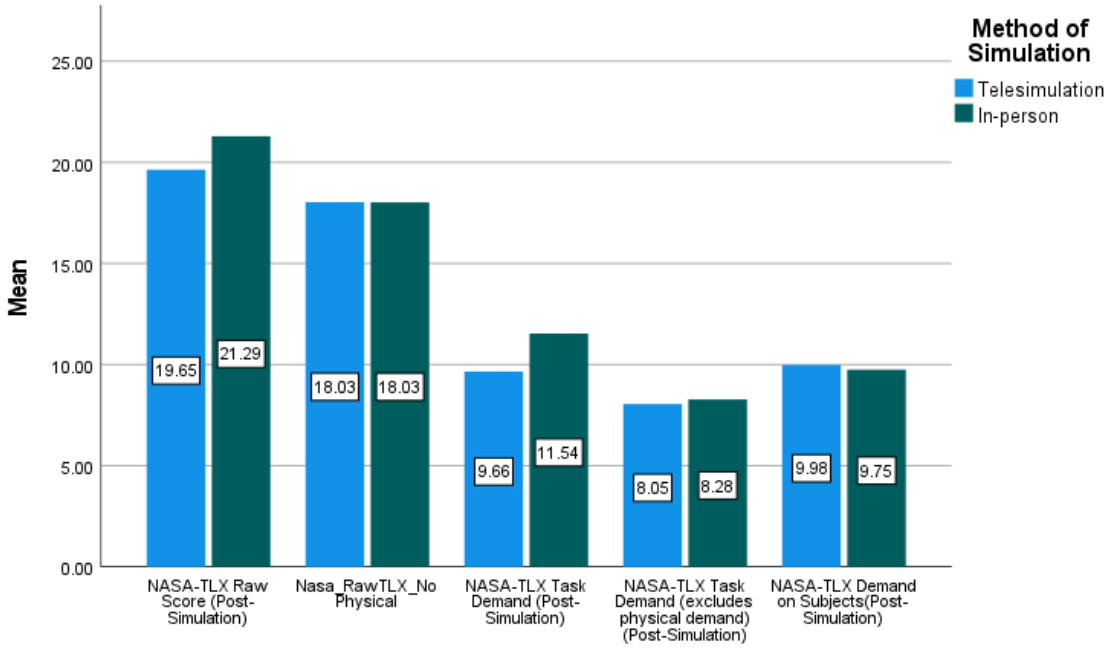
The fourth hypothesis also employed the non-inferiority approach. This hypothesis aims to test whether telesimulation is a non-inferior approach to healthcare simulation to impact mental workload in participants considering the pre-defined margin (13 %) following Berg & Steinsbekk (2020). Details of the hypothesis, including the tests performed and additional information, are presented in Table 96:

Table 96: Summary of included variables and testing procedures for Hypothesis 4

Research Question: Is the mental workload, measured by NASA-TLX, non-inferior to in-person simulation?
Objective: Demonstrate that telesimulation is a non-inferior method compared to in-person simulation regarding generated mental workload
Hypothesis 4
H4 (**null**): The mental workload (as measured by NASA-TLX) in telesimulation **is inferior** compared to in-person simulation
Alternative Hypothesis - The mental workload (as measured by NASA-TLX) in telesimulation **is not inferior** compared to in-person simulation
Sub-hypotheses
H4a (**null**): Task demand on the subject (Mental, Physical and Temporal) **is inferior** in telesimulation compared in in-person simulation (it is inferior)
Alternative hypothesis - Task demand on the subject (Mental, Physical and Temporal) **is not inferior** in telesimulation compared in in-person simulation (it is inferior)
H4b (**null**): Subject response to the task (Perceptions of effort, performance, and frustration) **is inferior** in telesimulation compared in-person simulation
Alternative hypothesis - Subject response to the task (Perceptions of effort, performance, and frustration) **is not inferior** in telesimulation compared in-person simulation
Experimental Design: Non-equivalent group design
Independent variable (s): Method of simulation (in-person vs telesimulation)
Dependent variable(s): NASA (TLX), and associated subdimensions (Task Demand) and (Demand on Subject)
Statistical Analysis to test the hypothesis: Two Sample T-Test (Equal-Variance T-Test for Non-Inferiority, Aspin-Welch Unequal-Variance T-Test for Non-Inferiority and Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority)
One-tailed / two-tailed or non-inferiority comparison: non-inferiority (one-tailed)
Rejection rule: Reject if p-value is less than 0.05 level of significance

The first method to visualize the scores for the dimensions of NASA-TLX for Telesimulation and in-person was to execute a mean comparison plot. The result indicates that average responses from telesimulation were only lower than that of the in-person approach on the NASA-TLX Raw score construct and on the NASA-TLX task demand construct. In contrast, the numerical average score values are relatively close to the remaining mental workload dimension.

Figure 73: NASA-TLX Telesimulation vs. In-person



Then we tested the mental workload responses in the telesimulation against the in-person approach, having as the null hypothesis that telesimulation is an inferior approach. The results of the tests for the NASA-TLX Raw during post-simulation are available in Table 97:

Table 97: Test for non-inferiority for Telesimulation – Mental Workload during Simulation (NASA TLX RAW)

NASA-TLX Raw		**NInf Hypothesis: Telesimulation > In-person - 0.107**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	19.645	5.657	0.718	2.000	18.209	21.082
In-person	72	21.292	7.141	0.842	1.994	19.614	22.970
Confidence Intervals for the Mean Difference							
Variance Assumption		Df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-1.647	1.126	1.978	-3.873	0.580
Unequal		131.130	-1.647	1.107	1.978	-3.836	0.543
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC-2.768	-1.647	1.126	0.996	132	0.160	No	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC -2.768	-1.647	1.107	1.014	131.130	0.156	No	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 2.768			0.199		0.421	No	

For the Student t-test, location parameter is given by mean difference. For the Mann-Whitney test, location parameter is given by the Hodges-Lehmann estimate

Modified-Levene 4.2807 0.041 Yes

Shapiro-Wilk 0.9732 0.193 No

The NASA-TLX Raw scores were normally distributed (W=0.9732, p=0.193), with the Levene test of equal variance between the two groups suggesting unequal variances (F=4.281, p=0,041). Thus, the t-test of unequal variance was employed, being was insignificant ((t(131.13)=1.014, p=0.156)) which implies the non-rejection of the inferiority null hypothesis that telesimulation is inferior in evoking mental workload based on the NASA-TLX Raw.

The same insignificant result was found for the non-parametric method (z=0.199, p=0.421). The test concludes that the telesimulation is inferior to in-person simulation in evoking mental workload as measured by NASA-TLX Raw.

The same test was performed on the NASA-TLX Raw dimension, excluding the physical demand. The result of the test is provided in Table 98:

Table 98: Test for non-inferiority for Telesimulation – Mental Workload during Simulation (NASA TLX RAW - Excluding Physical Demand)

NASA-TLX (excl. physical)		**NInf Hypothesis: Telesimulation > In-person - 2.344**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	18.032	5.301	0.673	2.000	16.686	19.378
In-person	72	18.028	5.841	0.688	1.994	16.655	19.400
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	0.004	0.970	1.978	-1.914	1.923
Unequal		131.62	0.004	0.963	1.978	-1.900	1.909
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 2.344	0.004	0.970	2.422	132.000	0.008	Yes	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 2.344	0.004	0.963	2.439	131.620	0.008	Yes	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 2.344			2.061		0.020	Yes	
Shapiro-Wilk	0.9773	0.30485	No				
Modified-Levene	0.5503	0.45953	No				

The test supports the normality assumption (W=0.977, p=0.305) of NASA-TLX Raw (excluding physical demand), providing evidence for using the parametric testing method. The Levene test for equal indicates equal variance assumption between the two groups is not violated (F=0.550, p=0.460).

Hence, the inferiority test based on an equal variance t-test showed a significant difference (t(132)=2.422, p=0.008). This result means that we reject the null hypothesis

of non-inferiority of the telesimulation on the NASA-TLX Raw (excluding physical demand).

We can conclude that telesimulation is non-inferior to the in-person simulation approach on the NASA-TLX Raw (excluding physical demand). The same test was performed on the Task Demand, including the physical demand. The result of the test is provided in Table 99:

Table 99: Test for non-inferiority for Telesimulation – Task Demand during Simulation

Task Demand		**NInf Hypothesis: Telesimulation > In-person - 1.5**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	9.661	2.953	0.375	2.000	8.911	10.411
In-person	72	11.542	4.392	0.518	1.994	10.509	12.574
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-1.880	0.658	1.978	-3.181	-0.580
Unequal		125	-1.880	0.639	1.979	-3.145	-0.615
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	Df	P	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 1.5	-1.880	0.658	-0.578	132	0.718	No	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	P	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 1.5	-1.880	0.639	-0.595	125	0.724	No	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		P	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 1.5			-0.7661		0.778	No	

Shapiro-Wilk 0.964 0.03734 Yes
Modified-Levene 13.4364 0.00036 Yes

The normality test rejects the normality assumption (W=0.964, p=0.037) of Task Demand, providing evidence for using the non-parametric testing method. The Levene test for equal indicates equal variance assumption between the two groups is violated (F=13.436, p<0.001).

The inferiority test based on the non-parametric method showed an insignificant difference (z=-0.766, p=0.778). The same result was obtained using the Welch approach

t-test ((t(125)= -0.595, p=0.724), failing to reject the null hypothesis of the inferiority of the telesimulation on the Task Demand.

Therefore, we retain the hypothesis that telesimulation is inferior to the in-person simulation approach on the NASA Task Demand construct. In the sequence, we perform the non-inferiority test for the Task Demand, excluding the physical demand component:

Table 100: Test for non-inferiority for Telesimulation – Task Demand during Simulation (excluding physical demand)

Task Demand (excl. physical)		**NInf Hypothesis: Telesimulation > In-person - 1.076**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	8.048	2.425	0.308	2.000	7.432	8.664
In-person	72	8.278	2.884	0.340	1.994	7.600	8.955
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-0.229	0.465	1.978	-1.148	0.690
Unequal		131.93	-0.229	0.459	1.978	-1.137	0.678
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 1.076	-0.229	0.465	1.822	132	0.035	Yes	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 1.076	-0.229	0.459	1.846	131.93	0.034	Yes	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 1.076			2.328		0.010	Yes	
Shapiro-Wilk	0.973	0.183	No				
Modified-Levene	2.639	0.107	No				

The test supports the normality assumption (W=0.973, p=0.0183) of Task Demand without the physical demand providing evidence for using the parametric testing method. Likewise, the Levene test for equal indicates equal variance assumption between the two groups is not violated (F=3.639, p=0.107).

The t-test showed a significant result (t(132)= 1.822, p=0.035), indicating that we can reject the null hypothesis of the inferiority of the telesimulation on the Task Demand, excluding physical demand. We, therefore, conclude that telesimulation is non-inferior to the in-person simulation of the Task Demand (excluding physical demand). Continuing the test for mental workload dimensions, we tested the Task Demand on the subject for non-inferiority. The result of the test is provided in Table 101:

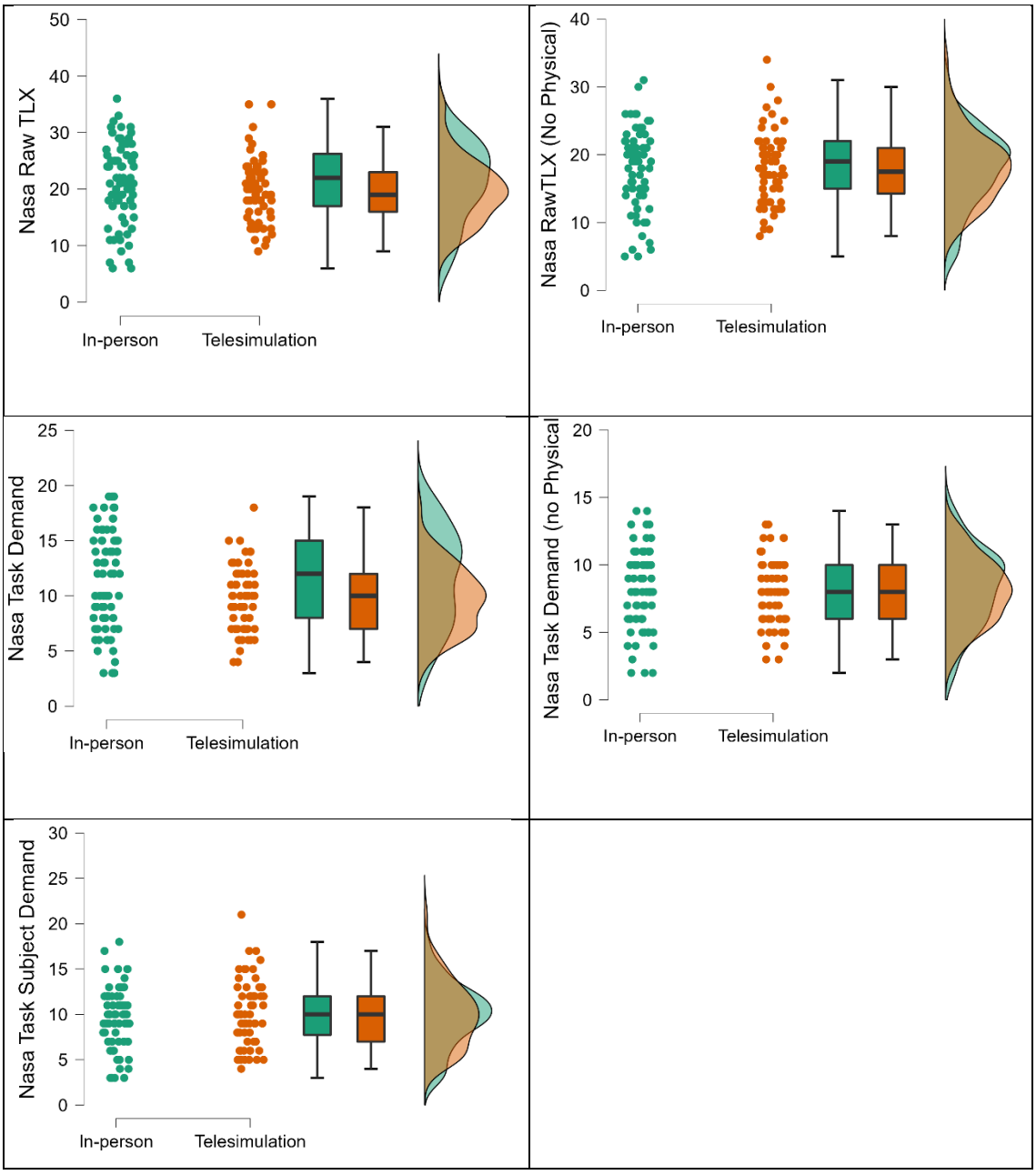
Table 101: Test for non-inferiority for Telesimulation – Demand on subjects during Simulation

Demand on Subjects		**NInf Hypothesis: Telesimulation > In-person - 1.268**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	9.984	3.660	0.465	2.000	9.054	10.913
In-person	72	9.750	3.467	0.409	1.994	8.935	10.565
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	0.234	0.616	1.978	-0.985	1.453
Unequal		126.7	0.234	0.619	1.979	-0.991	1.458
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 1.268	0.234	0.616	2.437	132	0.008	Yes	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 1.268	0.234	0.619	2.427	126.7	0.008	Yes	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 1.268			2.388		0.008	Yes	
Shapiro-Wilk	0.9621	0.05297	No				
Modified-Levene	0.3565	0.55145	No				

The normality test rejects the normality assumption (W=0.962, p=0.053) for Demand on the subject, indicating the usage of the non-parametric testing method. The Levene test for equal variance indicates equal variance assumption between the two groups is not violated (F=0.357, p=0.551).

The inferiority test based on the non-parametric method showed a significant difference (z=2.388, p=0.0008). The same result was obtained using the t-test (t(132)= 2.437, p=0.008), indicating that we reject the null hypothesis of the inferiority of the telesimulation on the Demand on the subject. Therefore, we retain the alternative hypothesis that telesimulation is non-inferior to the in-person simulation approach on the Demand on Subject. In the sequence, we present a raincloud plot for the NASA mental workload:

Figure 74: Raincloud plots for the variables of NASA mental workload vs. Simulation method



In the sequence, we present a summary of the results of hypothesis testing conducted in this item:

Table 102: Summary of Results for Hypothesis 4

	Statistic test result (p-value)	Decision
H4 (**null**): The mental workload (as measured by NASA-TLX) in telesimulation **is inferior** compared to in-person simulation	Tested in sub hypothesis	Reject the hypothesis
H4a (**null**): Task demand on the subject (Mental and Temporal) **is inferior** in telesimulation compared in in-person simulation (it is inferior)	($t(132)= 1.822$, $p=0.035$)	Reject the hypothesis
H4b: (**null**): Subject response to the task (Perceptions of effort, performance, and frustration) **is inferior** in telesimulation compared in-person simulation	($t(132)= 2.437$, $p=0.008$)	Reject the hypothesis

5.9.6. Hypothesis 5

The fifth hypothesis compared the impact of the scenario complexity on participants' emotional and cognitive response, using a nonequivalent groups design, comparing groups that participated in low complexity and medium complexity scenarios.

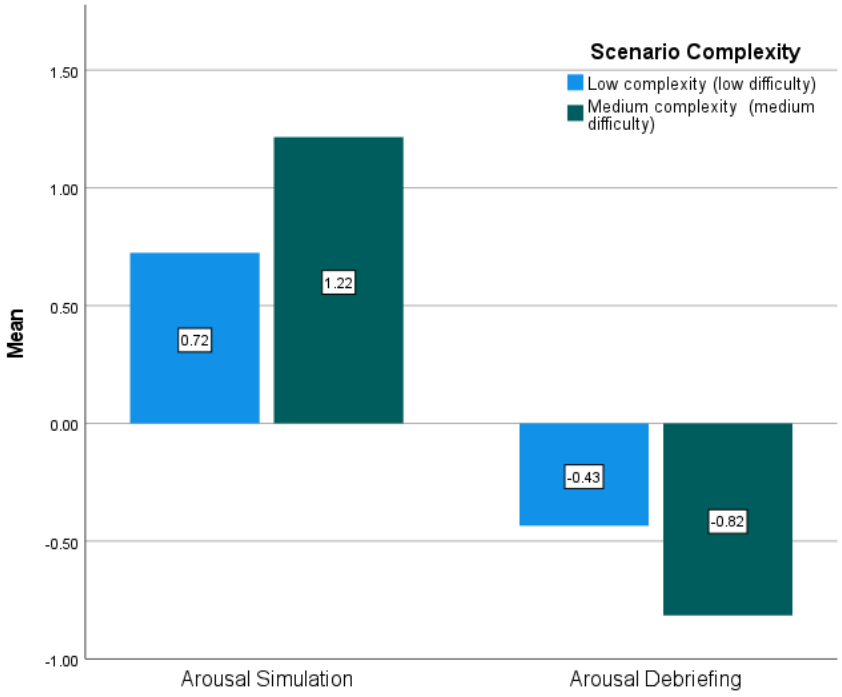
This hypothesis aims to test whether the scenario complexity has a role in the level of emotional response and cognitive response. Details of the hypothesis, including the tests performed and additional information, are presented in Table 103:

Table 103: Summary of included variables and testing procedures for Hypothesis 5

Research Question: Does scenario complexity affects the magnitude of emotional and cognitive responses?
Objective: Demonstrate that higher scenario complexity evokes higher emotional and cognitive responses.
Hypothesis 5
H5 (**null**): Scenario complexity does **not affect** the level of emotional and cognitive responses during telesimulation and in-person simulation
Alternative Hypothesis - Scenario complexity **affects** the level of emotional and cognitive responses during telesimulation and in-person simulation
Sub-hypotheses
H5a (**null**): Higher complexity scenarios is not associated with stronger emotional state (ESS)
Alternative hypothesis - Higher complexity scenarios **is associated** with stronger emotional state (ESS)
H5b (**null**): Higher scenario complexity is not associated with stronger arousal response (FAS)
Alternative hypothesis - Higher scenario complexity **is associated** with stronger arousal response (FAS)
H5c (**null**): Higher scenario complexity **is not associated** with a higher task demand on subjects (NASA-TLX)
Alternative hypothesis - Higher scenario complexity **is associated** with a higher task demand on subjects (NASA-TLX)
H5d (**null**): Higher scenario complexity **is not associated** with a higher subject response to the task (NASA-TLX)
Alternative hypothesis - Higher scenario complexity **is associated** with a higher subject response to the task (NASA-TLX)
Experimental Design: Non-equivalent group design
Independent variable (s): Scenario Complexity (low vs medium complexity)
Dependent variable(s): Emotional State Scale (ESS), Felt Arousal Scale (ESS), NASA-TLX
Statistical Analysis to test the hypothesis: Independent Sample-T test
One-tailed / two-tailed or non-inferiority: one-tailed
Rejection rule: Reject if p-value is less than 0.05 level of significance

For the fifth hypothesis, we executed the mean plot of the arousal responses based on the scenario complexity during the simulation and the debriefing. The plot is available in Figure 75:

Figure 75: Arousal Response During Simulation and Debriefing – Low and Medium Complexity



The plot indicates arousal responses evoked during the simulation period were higher for medium complexity in numerical value than the low complexity. At the debriefing moment, the reverse was observed, with induced arousal responses from the medium complexity being relatively lower than that of the low complexity when compared numerically.

Proceeding further, the Levene test of equal variance from the simulation and debriefing support that the equal variability assumption between the two groups of scenario complexity for the arousal construct responses was insignificant (F=1.576, p=0.212) for the simulation insignificant (F=1.727, p=.191) for the debriefing period. The results are displayed in the tables in sequence:

Table 104: Test of Equality of Variances for Arousal Responses

Variable	F	df	P
Arousal Response Simulation	1.576	1	0.212
Arousal Response Debriefing	1.727	1	0.191

Table 105: Arousal Responses vs Scenario Complexity – Independent T-Test

Variable	**Test**	**Statistic**	**Df**	**p**	**Location Parameter**	**SE Diff.**
Arousal Response Simulation	Student	2.176	132	0.016	0.491	0.226
	Welch	2.19	128.86	0.015	0.491	0.224
	Mann-Whitney	2585		0.057	3.699×10^{-5}	
Arousal Response Debriefing	Student	-1.718	132	0.044	-0.381	0.221
	Welch	-1.727	130.664	0.043	-0.381	0.22
	Mann-Whitney	1819.5		0.026	-2.722*10-5	

Note. For the Arousal response during Simulation - the alternative hypothesis specifies that group Medium complexity (medium difficulty) is greater than group Low complexity (low difficulty) For the Arousal response during Debriefing, the alternative hypothesis specifies that group Medium complexity (medium difficulty) is less than group Low complexity (low difficulty) .For the Student t-test and Welch t-test, the location parameter is given by mean difference. For the Mann-Whitney test, the location parameter is given by the Hodges-Lehmann estimate.

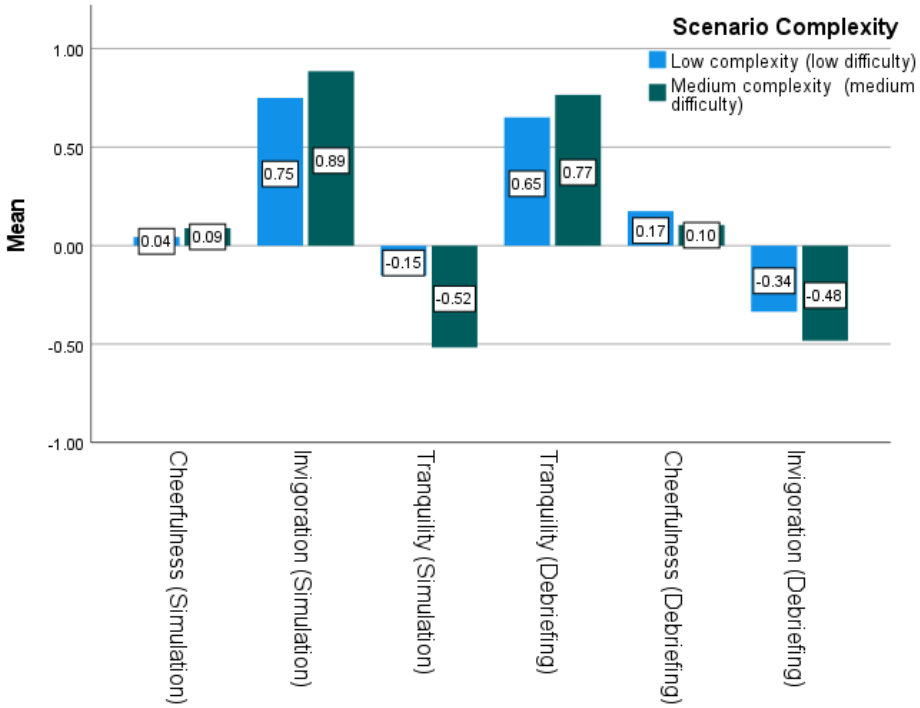
Results indicate that the arousal responses differed significantly between the complexity levels (t(132)= 2.176, p=0.016). Significant result (t(132)= -1.718, p=0.044) was obtained between the low and medium scenario complexity during the debriefing.

Results indicate that the null hypothesis was rejected at the simulation phase, confirming that the arousal responses in medium complexity are greater than those in low complexity scenarios.

During the debriefing period, the null was rejected. The result supports the claim that the arousal responses in the medium complexity are lower than those during low complexity scenarios.

Proceeding with this hypothesis, we examined the role of the relationship of scenario complexity with response evoked based on the emotional dimension during both the simulation and the debriefing period, as displayed in Figure 76:

Figure 76: Emotional Appraisal Dimensions During Simulation and Debriefing – Low and Medium Complexity



The plot demonstrates that the cheerfulness of the two complexity levels was similar. However, scenarios with medium complexity displayed more significant scores for invigoration. Scores for the tranquility construct were lower for medium complexity when compared with low complexity.

After debriefing, the mean comparison shows cheerfulness score value was close between the medium and low complexity. Invigoration was lower in the medium complexity than the low complexity, while tranquility was higher in the medium complexity.

In the sequence, we tested the emotional responses between the two complexity scenarios (medium and low) considering the debriefing and the simulation period across all the emotional dimensions, first by testing the equality of variances and then using an independent t-test:

Table 106: Test of Equality of Variances for Emotional Appraisal Dimensions

Test of Equality of Variances (Levene's)			
Variable	**F**	**df**	**p**
Cheerfulness response (Simulation)	0.059	1	0.808
Cheerfulness response (Debriefing)	1.534	1	0.218
Invigoration response (Simulation)	0.789	1	0.376
Invigoration response (Debriefing)	0.002	1	0.967
Tranquility response (Simulation)	0.095	1	0.758
Tranquility response (Debriefing)	0.02	1	0.887

The Levene-test of equal variances between the two complexity was insignificant for all the emotional response dimensions at a 5% level with an obtained p-value greater than 0.05, indicating the plausibility of the equality of variance assumption. Then, we performed the independent t-test for all the emotional variables for the simulation and debriefing period against the scenario complexity:

Table 107: Emotional Responses – Simulation and Debriefing vs. Complexity – Independent T-Test

Variable	**Test**	**Statistic**	**df**	**P**	**Loc. Param.**	**SE Diff.**
Cheerfulness response (Simulation)	Student	0.4	132	0.345	0.046	0.115
	Welch	0.402	131.031	0.344	0.046	0.114
	Mann-Whitney	2342		0.326	3.323×10^{-6}	
Cheerfulness response (Debriefing)	Student	-0.624	132	0.733	-0.07	0.112
	Welch	-0.629	127.514	0.735	-0.07	0.111
	Mann-Whitney	2301.5		0.395	4.180×10^{-5}	
Invigoration response (Simulation)	Student	0.982	132	0.164	0.135	0.137
	Welch	0.979	128.026	0.165	0.135	0.138
	Mann-Whitney	2326		0.354	4.799×10^{-5}	
Invigoration response (Debriefing)	Student	-1.104	132	0.864	-0.147	0.134
	Welch	-1.103	131.014	0.864	-0.147	0.134
	Mann-Whitney	1974		0.89	-50.39	
Tranquility response (Simulation)	Student	-2.208	132	0.986	-0.364	0.165
	Welch	-2.211	131.958	0.986	-0.364	0.165
	Mann-Whitney	1877		0.949	-0.36	
Tranquility response (Debriefing)	Student	0.652	132	0.258	0.115	0.176
	Welch	0.652	131.452	0.258	0.115	0.176
	Mann-Whitney	2359.5		0.301	2.500×10^{-5}	

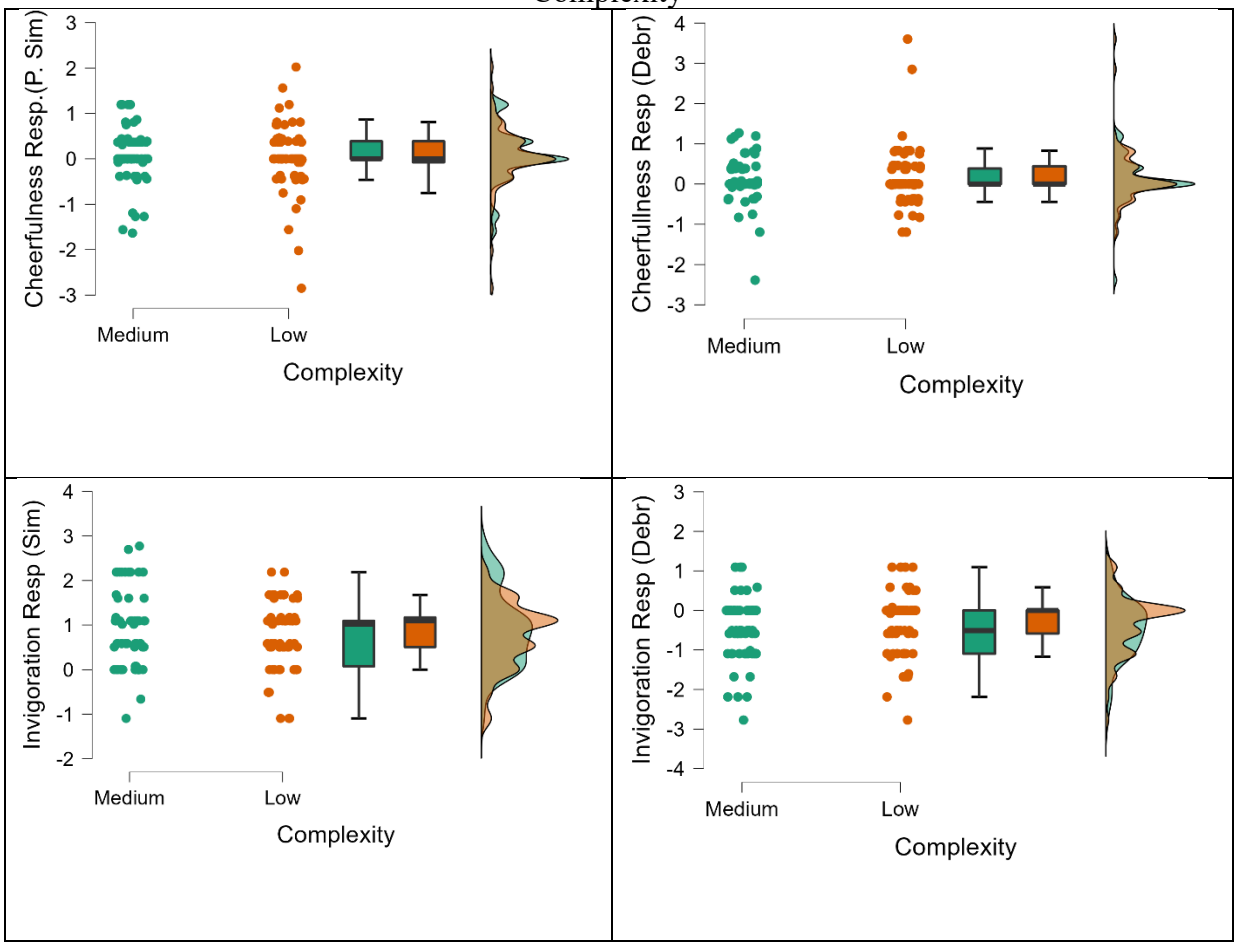
Note. For all tests, the alternative hypothesis specifies that group Medium complexity (medium difficulty) is greater than group Low complexity (low difficulty) For the Student t-test and Welch t-test, location parameter is given by mean difference. For the Mann-Whitney test, location parameter is given by the Hodges-Lehmann estimate.

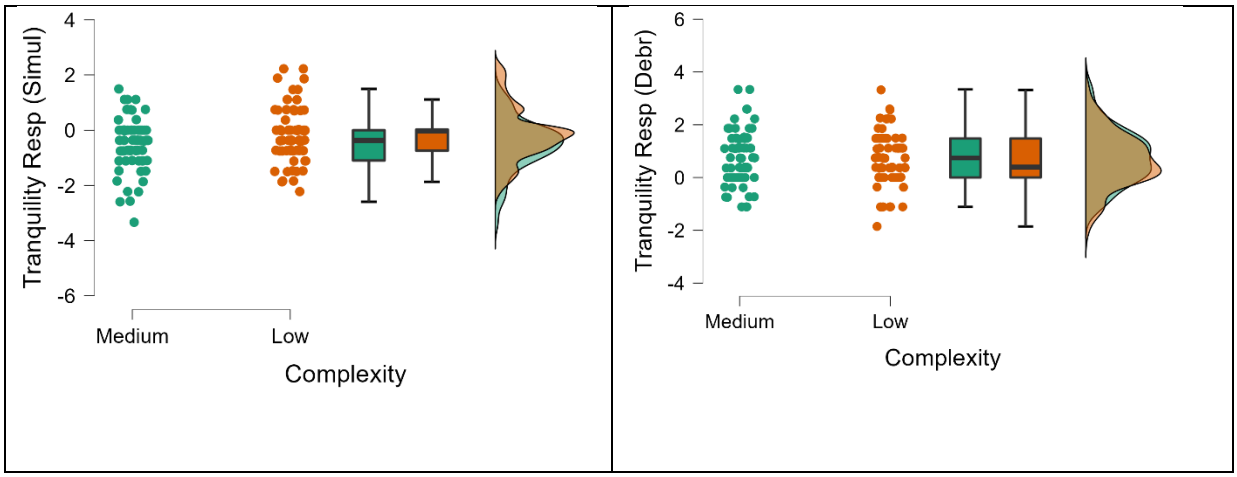
The test result shows that responses induced at the simulation and during the debriefing between the complexity were insignificant at 5% for all the three considered emotional dimensions.

The tests supported the associated null hypotheses. The result confirms that irrespective of the period (simulation or debriefing), the emotional responses from medium complexity are not greater than those from low complexity for those dimensions.

The density distribution, boxplots, and scatter points for telesimulation and in-person in the dimensions emotional responses dimensions are available in the sequence:

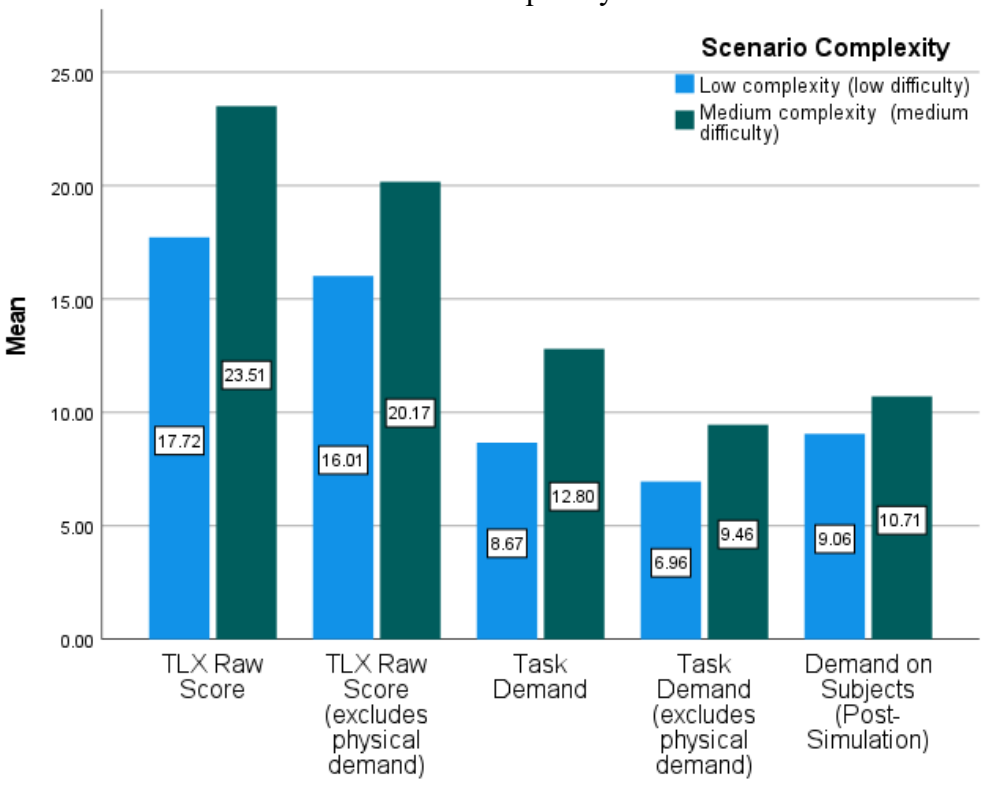
Figure 77: Raincloud plots for the variables of Emotional Responses vs Complexity





Finally, we compared scenario complexity levels in the mental workload dimensions. Figure 78 presents the mean plot for all mental workload dimensions against scenario complexity:

Figure 78: NASA Mental Workload Scale and subdimensions – Low and Medium Complexity



The mean plot indicated that the mental workload construct's mean score was numerically higher in the medium complexity than the medium complexity. Then the variables were subjected to a t-test with the alternative hypothesis stating that the mental

workload in the medium complexity is greater than the mental workload in the low complexity tested at a 5% level.

Table 108: Test of Equality of Variances for the Mental Workload Scale and Subdimensions (NASA-TLX)

Test of Equality of Variances (Levene's)			
Variable	**F**	**df**	**p**
Nasa RAW	0.109	1	0.742
Nasa Raw (Excludes Physical)	0.335	1	0.564
Nasa Task Demand	1.73	1	0.191
Nasa Task Demand (Excludes Physical)	1.413	1	0.237
Nasa Task Subject Demand	0.014	1	0.905

The obtained result from Levene's test assumes equal variances between the two complexities. The following Table demonstrates the comparison tests between the Nasa Complexity using the complexity as the group factor. Results are displayed in Table 109:

Table 109: Mental Workload Scale and Subdimensions (NASA-TLX) vs Complexity – Independent Samples T-Test

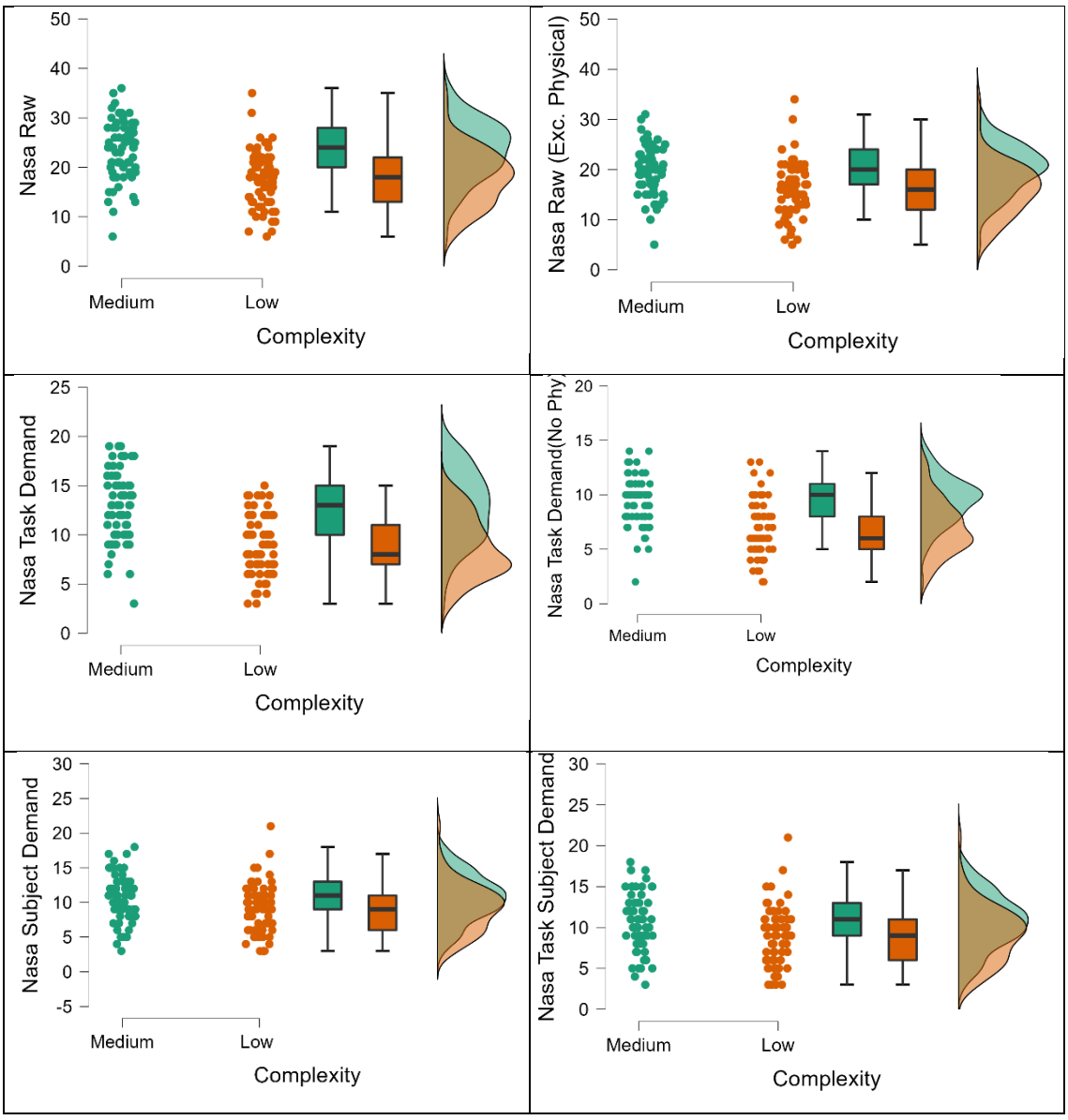
Variable	**Test**	**Statistic**	**df**	**p**	**Location Parameter**	**SE Difference**
NASA Raw	Student	5.702	132	< .001	5.783	1.014
	Welch	5.697	130.86	< .001	5.783	1.015
	Mann-Whitney	3430		< .001	6	
NASA Raw (No Physical)	Student	4.63	132	< .001	4.155	0.897
	Welch	4.642	131.88	< .001	4.155	0.895
	Mann-Whitney	3282.5		< .001	4	
Nasa Task Demand	Student	7.221	132	< .001	4.133	0.572
	Welch	7.183	125.19	< .001	4.133	0.575
	Mann-Whitney	3614		< .001	4	
Nasa Task Demand (No Physical)	Student	6.117	132	< .001	2.505	0.41
	Welch	6.138	131.57	< .001	2.505	0.408
	Mann-Whitney	3500.5		< .001	3	
Nasa Task Subject Demand	Student	2.758	132	0.003	1.65	0.598
	Welch	2.761	131.94	0.003	1.65	0.598
	Mann-Whitney	2865		0.003	2	

Note. For all tests, the alternative hypothesis specifies that group Medium complexity (medium difficulty) is greater than group Low complexity (low difficulty) For the Student t-test and Welch t-test, location parameter is given by mean difference. For the Mann-Whitney test, location parameter is given by the Hodges-Lehmann estimate.

The result of the t-test showed that the test supports the non-rejection of the alternative hypotheses for all the constructs of the mental workload construct at a 1% level of significance, indicating that mental workload responses from the medium complexity were more significant than the mental workload induced from of low complexity scenarios.

We compared the density distribution, boxplots, and scatter points for telesimulation and in-person in the dimensions of mental workload using a raincloud plot, as illustrated in the figure below:

Figure 79: Raincloud plots for the variables of NASA mental workload vs Complexity



Finally, we present a summary of the results of the hypothesis testing for the fifth hypothesis:

Table 110: Summary of Results for Hypothesis 5

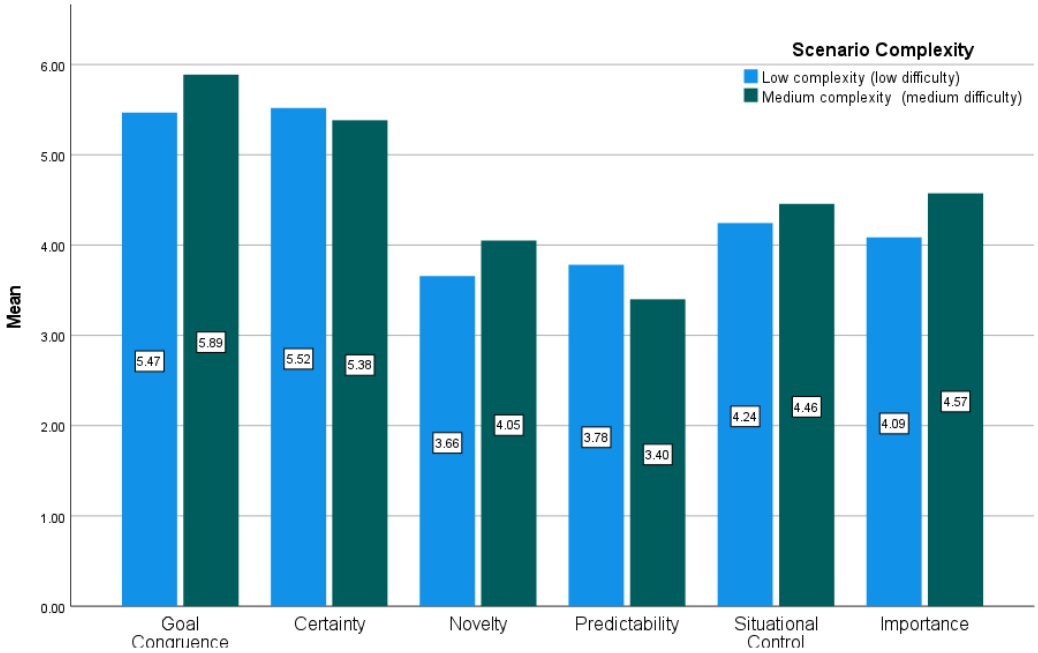
Hypothesis	Statistic test result (p-value)	Decision
H5 (**null**): Scenario complexity does **not affect** the level of emotional and cognitive responses	Tested in sub hypothesis	Reject the hypothesis
H5a: Higher complexity scenarios are not associated with stronger emotional state (ESS)	$p>0.005$	Fail to reject the hypothesis
H5b: Higher scenario complexity are not associated with stronger arousal response (FAS)	$p<0.001$	Reject the hypothesis
H5c: Higher scenario complexity are not associated with a higher task demand on subjects (NASA-TLX)	$p<0.001$	Reject the hypothesis
H5d: Higher scenario complexity are not associated with a higher subject response to the task (NASA-TLX)	$p=0.003$(and lower)	Reject the hypothesis

POST HOC ANALYSIS

In this section, we have performed tests not pre-established by hypothesis development but were executed given the practical and potential theoretical importance. We examined the emotion appraisal dimension differences regarding low and medium complexity scenarios.

The first attempt to visualize the comparison between the two groups was to execute a mean plot, as displayed in the following Figure:

Figure 80: Emotion Appraisal Dimensions– Low and Medium Complexity



The means plot provides an insight that reveals that induced responses were higher in the medium complexity than low complexity for four dimensions, including goal congruence, novelty, situational control, and importance. In comparison, responses from the medium were only lower than low complexity in two dimensions: predictability and certainty.

Finally, the emotional appraisal dimensions were compared to determine the differences between the different complexities. The alternative hypothesis was that the induced emotional appraisals were greater in the medium complexity than low complexity.

Table 111: Emotional Appraisal Process – Comparison between Medium and Low Complexity – Independent Samples T-Test

Variable	Student t	df	p	Loc. Param.	SE Diff.
Pleasantness	1.287	132	0.100	0.276	0.215
Goal Congruence	2.000	132	0.024	0.421	0.21
Certainty	-0.651	132	0.742	-0.136	0.21
Novelty	1.785	132	0.038	0.394	0.221
Predictability	-1.746	132	0.042	-0.383	0.211
Situational Control	1.125	132	0.131	0.213	0.189
Importance	2.742	132	0.003	0.486	0.177
Fidelity	1.667	132	0.049	0.272	0.163
Presence	3.051	132	0.001	11.651	3.819
Feedback	0.474	132	0.318	0.058	0.122

Note. For all tests, the alternative hypothesis specifies that group Medium complexity (medium difficulty) is greater than group Low complexity (low difficulty) Student's t-test. For the predicability the alternative hypothesis specifies that group Medium complexity (medium difficulty) is less than group Low complexity (low difficulty) .

The result from the t-test showed that at a 5% level, the result from the goal congruences, novelty, importance, fidelity, and presence were found to be significant, which confirm that the emotional appraisal process variables of these five dimensions were higher in the medium difficulty than in the low difficulty.

The result from pleasantness, certainty, predictability, situational control, and feedback was not significant at 5%, supporting the non-rejection of the null hypothesis, indicating that in these five dimensions, the emotional appraisal from the medium difficulty was not greater than those from the low difficulty.

5.9.7. Hypothesis 6

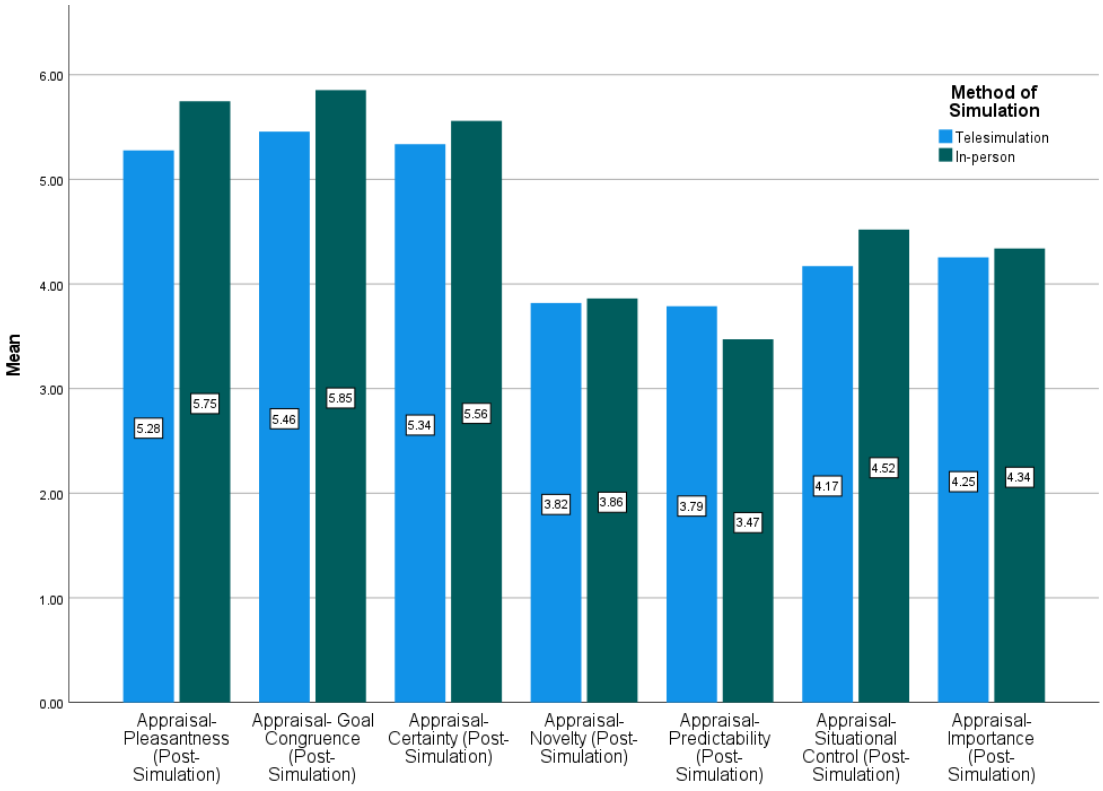
The sixth hypothesis compared the emotional appraisal process between telesimulation and in-person simulation, using a nonequivalent groups design, and conducting a between-subjects comparison. To test this hypothesis, we again employed the non-inferiority approach. This hypothesis tests whether telesimulation is a non-inferior approach to in-person simulation regarding the emotional appraisal process. Details of the hypothesis, including the tests performed and additional information, are presented in Table 112:

Table 112: Summary of included variables and testing procedures for Hypothesis 6

Research Question: Is the emotional appraisal process in telesimulation non-inferior to in-person?
Objective: Demonstrate that the appraisal process in telesimulation is not inferior to in-person simulation.
Hypothesis 6
H6 (**null**): The scores of emotional appraisal process (Pleasantness, Goal Congruence, Certainty, Novelty, Predictability, Situational Control, Importance and Fairness) scores in telesimulation **are inferior** to in-person simulation
Alternative Hypothesis - The scores of emotional appraisal process (Pleasantness, Goal Congruence, Certainty, Novelty, Predictability, Situational Control, Importance and Fairness) scores in telesimulation **are not inferior** to in-person simulation
Experimental Design: Non-equivalent group design
Independent variable (s): Method of simulation (in person vs telesimulation)
Dependent variable(s): Emotional Appraisal Process Scale and its dimensions
Statistical Analysis to test the hypothesis: Two Sample T-Test (Equal-Variance T-Test for Non-Inferiority, Aspin-Welch Unequal-Variance T-Test for Non-Inferiority and Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority)
One-tailed / two-tailed or non-inferiority: non-inferiority (one-tailed)
Rejection rule: Reject if p-value is less than 0.05 level of significance

The first approach to compare the two methods was to plot a mean comparison between Telesimulation and in-person simulation. The Results are available in the figure in the sequence:

Figure 81: Emotion Appraisal Dimensions– Telesimulation and In-person



The mean comparison suggests that the appraisal, goal congruence, certainty, and situational control are visually higher in the in-person group compared to the telesimulation. However, the remaining item will test for non-inferiority for each dimension.

The pleasantness dimension was tested for normality assumption and equal variances. The test outcome found that evoked response to the pleasantness follows a non-normal distribution ($W=0.8885$, $p<0.001$) with equal variances between the telesimulation and in-person simulation not violated ($F=,0.3711$, $p=0.5435$), which underlines the usage of the non-parametric method as an alternative to t-testing.

Table 113: Test for non-inferiority for Telesimulation – Pleasantness

Pleasantness		**NInf Hypothesis: Telesimulation > In-person - 0.747**					
Comparison	Count	Mean	Std Dev of Data	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	5.285	1.218	0.155	2.000	4.976	5.595
In-person	72	5.747	1.237	0.146	1.994	5.456	6.038
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-0.4613	0.2128	1.978	-0.882	-0.040
Unequal		129.64	-0.4613	0.2125	1.978	-0.882	-0.041
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	Df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.747	-0.461	0.213	1.342	132	0.091	No	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	Df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.747	-0.461	0.213	1.344	129.64	0.091	No	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.747			0.622		0.267	No	
Shapiro-Wilk	0.8885	0.00001	Yes				
Modified-Levene	0.3711	0.54345	No				

The result obtained from the Mann-Whitney test indicates a non-significant test (z=0.622, p=0.267) of non-inferiority of pleasantness in telesimulation (M=5.228, SD=1.218) compared to the in-person approach (M=5.747, SD=1.237). The non-significant test was observed from the t-test, supporting the non-acceptance of the non-inferiority alternative. In summary, the telesimulation approach is inferior to the in-person approach based on the induced pleasantness dimension.

Next, the goal congruence dimension was tested for normality assumption and equal variances. The test outcome found that evoked response on the goal congruence at the simulation is non-normal (W=0.9365, p<0.0012) with the equal variances assumption between the telesimulation and in-person simulation not being violated (F=1.1178, p=0.2923); therefore, we used the non-parametric method for testing.

Table 114: Test for non-inferiority for Telesimulation – Goal Congruence

Goal Congruence		**NInf Hypothesis: Telesimulation > In-person - 0.761**					
Comparison	Count	Mean	Std Dev of Data	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	5.459	1.154	0.147	2.000	5.166	5.752
In-person	72	5.854	1.274	0.150	1.994	5.555	6.153
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-0.3946	0.211362	1.978	-0.813	0.023
Unequal		131.64	-0.3946	0.209804	1.978	-0.810	0.020
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.761	-0.395	0.211	1.733	132	0.043	Yes	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.761	-0.395	0.210	1.746	131.64	0.042	Yes	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.761			1.8309		0.034	Yes	
Shapiro-Wilk	0.9365	0.00127	Yes				
Modified-Levene	1.1178	0.29232	No				

The result obtained from the Mann-Whitney test indicates the rejection of the null hypothesis (z=1.8309, p=0.034) of the inferiority of goal congruence response between the telesimulation approach (M=5.459, SD=1.154) and in-person approach (M=5.854, SD=1.274). The test's significance was also observed from the t-test, supporting the acceptance of the non-inferiority alternative, indicating that telesimulation is non-inferior to the in-person approach based on the goal congruence dimension.

The next dimension of emotional appraisal was Certainty. First, we tested the normality and equal variances assumptions. The test outcome found that evoked response on the Certainty at the simulation is non-normal (W=0.9606, p=0.024) with equal variances between the telesimulation and in-person simulation not violated (F=0.2363, p=0.628), prompting us to use the non-parametric method. The results of the non-inferiority tests are available in sequence:

Table 115: Test for non-inferiority for Telesimulation – Certainty

Certainty		**Telesimulation > In-person - 0.723**					
Comparison	Count	Mean	Std Dev of Data	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	5.329	1.254	0.159	2.000	5.010	5.647
In-person	72	5.560	1.169	0.138	1.994	5.285	5.834
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-0.2307	0.20949	1.978	-0.645	0.184
Unequal		125.9	-0.2307	0.2106	1.979	-0.647	0.186
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.723	-0.231	0.209	2.350	132	0.010	Yes	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.723	-0.231	0.211	2.338	125.9	0.010	Yes	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.723			2.421		0.008	Yes	
Shapiro-Wilk	0.9606	0.024	Yes				
Modified-Levene	0.2363	0.628	No				

The result obtained from the non-parametric Mann-Whitney test indicates significant results (z=2.421, p=0.008), rejecting the null hypothesis of the inferiority of telesimulation (M=5.329, SD=1.254) to in-person (M=5.560, SD=1.169) for the certainty dimensions. We, therefore, retain the alternative hypothesis that telesimulation is non-inferior to the in-person approach based on the Certainty dimension.

Following the list of variables, we proceed to test the Novelty dimension. We tested the variable for the normality and equal variances assumptions. The test indicates that the Novelty dimension during the simulation can be assumed to be normal (W=0.9766, p=0.199) with equal variances assumption between the telesimulation and in-person simulation (F=1.6591, p=0.200), supporting the usage of the parametric t-test with equal variance approach for the non-inferiority test.

Table 116: Test for non-inferiority for Telesimulation – Novelty

Novelty		**Telesimulation > In-person – 0.502**					
Comparison	Count	Mean	Std Dev of Data	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	3.834	1.410	0.179	2.000	3.476	4.192
In-person	72	3.862	1.182	0.139	1.994	3.584	4.139
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-0.0279	0.22393	1.978	-0.471	0.415
Unequal		119.6	-0.0279	0.22689	1.980	-0.477	0.421
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.502	-0.028	0.224	2.117	132	0.018	Yes	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.502	-0.028	0.227	2.090	119.56	0.019	Yes	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.502			2.3773		0.009	Yes	
Shapiro-Wilk	0.9766	0.199	No				
Modified-Levene	1.6591	0.2	No				

T-tests showed a significant p-value (t(132)=-0.028, p=0.018) for the test of inferiority of telesimulation (M=3.834, SD=1.410) to in-person approach (M=3.862, SD=1.182) in the novelty appraisal dimension. Therefore, we retained the hypothesis that telesimulation is non-inferior to the in-person approach based on the novelty emotional appraisal dimension.

The following variable tested for non-inferiority was the dimension of Predictability. The normality and equality of variance test indicated that predictability was non-normal (W=0.9276, p<0.001). However, the groups displayed equal variances (F=2.0015, p=0.16). Therefore the non-parametric method was chosen as the alternative to the traditional t-tests. The results of the non-inferiority testing for predictability are available in the following table:

Table 117: Test for non-inferiority for Telesimulation – Predictability

Predictability		**Telesimulation > In-person - 0.451**					
Comparison	Count	Mean	Std Dev of Data	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	3.742	1.144	0.145	2.000	3.451	4.033
In-person	72	3.472	1.300	0.153	1.994	3.167	3.778
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	0.26971	0.21317	1.978	-0.152	0.691
Unequal		131.9	0.26971	0.21114	1.978	-0.148	0.687
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.451	0.270	0.213	3.381	132	0.000	Yes	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.451	0.270	0.211	3.413	131.93	0.000	Yes	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.451			3.6906		1E-04	Yes	
Shapiro-Wilk	0.9276	5E-04	Yes				
Modified-Levene	2.0015	0.16	No				

The result obtained from the Mann-Whitney test indicates support for rejecting the null hypothesis ($z=3.6909$, $p<0.001$) that the predictability of telesimulation ($M=3.742$, $SD=1.144$) is inferior to in-person ($M=3.472$, $SD=1.300$). We observed a p-value significative at a 1% level, which supported the acceptance of the non-inferiority alternative for the other t-tests. We then conclude that the telesimulation approach is non-inferior to the in-person approach on the predictability dimension.

Following the process, the Situational Control dimension was tested for normality assumption and equal variances. The test indicates that the variable Situational Control during the simulation period can assume normality ($W=0.9759$, $p=0.181$) with equal variances between the telesimulation and in-person simulation not violated ($F=0.0817$, $p=0.776$) indicating the usage of the parametric t-test with equal variance approach for the non-inferiority test.

Table 118: Test for non-inferiority for Telesimulation – Situational Control

Situational Control		**Telesimulation > In-person - 0.588**					
Comparison	Count	Mean	Std Dev of Data	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	4.146	1.058	0.134	2.000	3.877	4.415
In-person	72	4.521	1.106	0.130	1.994	4.261	4.781
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-0.3749	0.1879	1.978	-0.7465	-0.003
Unequal		130.5	-0.3749	0.18727	1.978	-0.7453	-0.004
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.588	-0.375	0.188	1.134	132	0.129	No	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.588	-0.375	0.187	1.138	130.52	0.129	No	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.588			0.8289		0.204	No	
Shapiro-Wilk	0.9759	0.181	No				
Modified-Levene	0.0817	0.776	No				

The result obtained from the parametric t-test failed to reject the null hypothesis($t(132)=1.138$, $p=0.129$) of the inferiority of situational control in the telesimulation ($M=4.146$, $SD=1.058$) compared to the in-person approach ($M=4.521$, $SD=1.106$).

Tests were also insignificant for t-tests using Aspin-Welch and the non-parametric approaches, which supported the acceptance of the inferiority as stated by the null hypothesis. Since we failed to reject the null hypothesis, we maintain the assumption that telesimulation is inferior to the in-person approach based on the Situational Control aspect of the emotional appraisal process.

The last dimension to be tested was the dimension of Importance. We found that the Importance variable was non-normal ($W=0.894$, $p<0.001$), with variances between the telesimulation and in-person simulation being considered equal ($F=0.0914$, $p=0.763$). These results indicate that the recommended test is the non-parametric version.

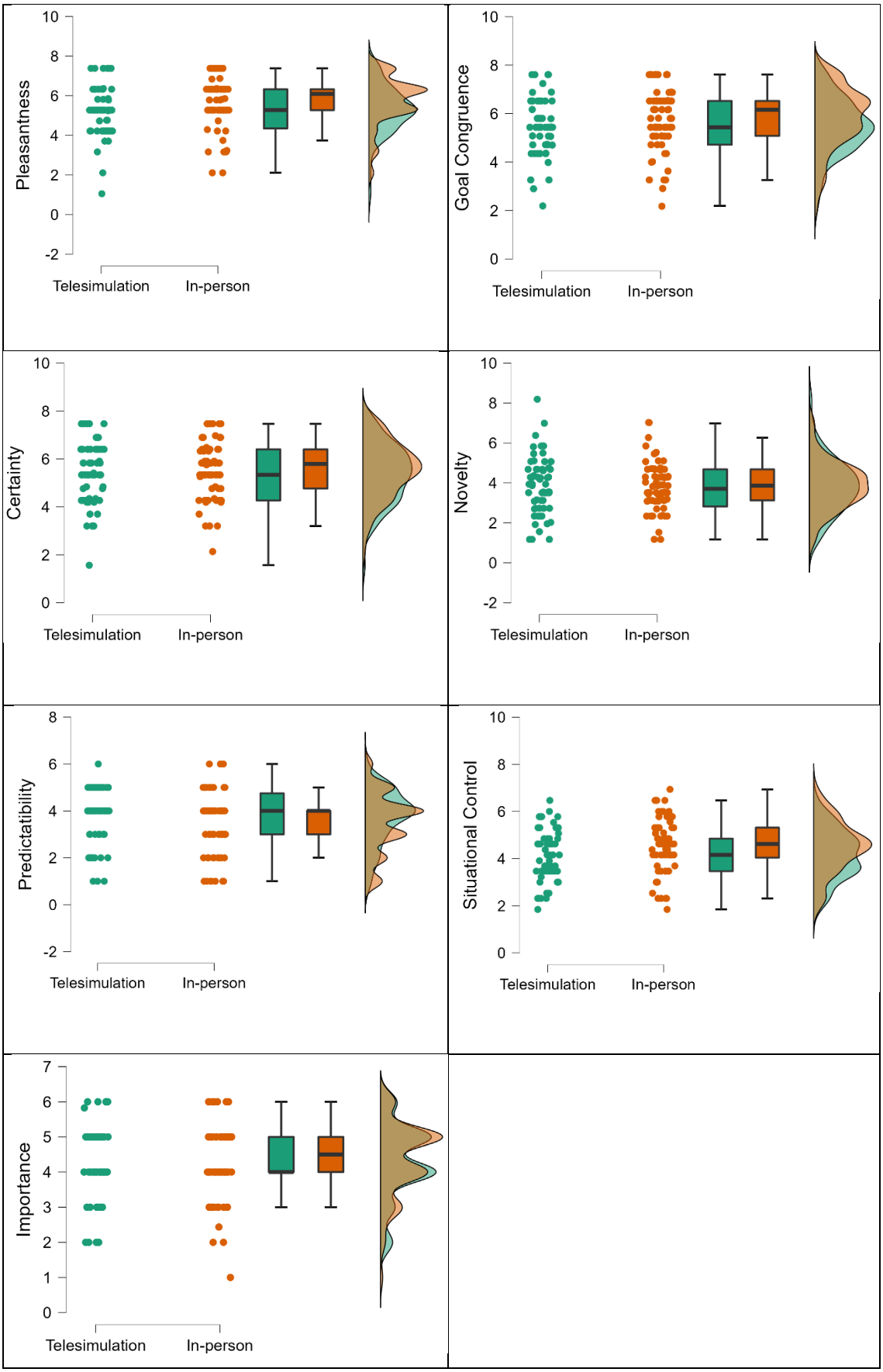
Table 119: Test for non-inferiority for Telesimulation – Importance

Importance		**Telesimulation > In-person - 0.564**					
Comparison	Count	Mean	Std Dev of Data	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	4.304	1.061	0.135	2.000	4.034	4.573
In-person	72	4.339	1.049	0.124	1.994	4.093	4.586
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-0.0358	0.1826	1.978	-0.3971	0.326
Unequal		128.6	-0.0358	0.1828	1.979	-0.3975	0.326
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.564	-0.036	0.183	2.891	132	0.002	Yes	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.564	-0.036	0.183	2.889	128.64	0.002	Yes	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.564			2.5681		0.005	Yes	
Shapiro-Wilk	0.894	2E-05					
Modified-Levene	0.0914	0.763	No				

The Mann-Whitney test was significant (z=2.5681, p=0.005), supporting the rejection of the inferiority of telesimulation (M=4.304, SD=1.061) compared to in-person (M=4.339, SD=1.049) regarding the perception of importance. We also observed significant results from the parametric t-test at a 1% level, supporting the acceptance of the non-inferiority alternative. Given the non-inferiority tests, we can determine that telesimulation is non-inferior to the in-person approach regarding the perceived importance of the emotional appraisal process.

Finally, we generated a raincloud plot with the density distribution, boxplots, and scatter points for telesimulation and in-person for the seven dimensions of the emotional appraisal process available in Figure 82:

Figure 82: Raincloud plots for the variables of Emotion Appraisal



In conclusion, we present the summary of results for hypothesis 6, with the hypothesis, the statistic test and -value, and the decision based on that statistic:

Table 120: Summary of Results for Hypothesis 6

Hypothesis	Statistic test result (p-value)	Decision
H6 (**null**): The scores of emotional appraisal process (Pleasantness, Goal Congruence, Certainty, Novelty, Predictability, Situational Control, Importance and Fairness) scores in telesimulation **are inferior** to in-person simulation	Pleasantness (z=0.622, p=0.267) Goal Congruence (z=1.8309, p=0.034) Novelty ((t(132)=-0.028, p=0.018) Predictability (z=3.6909, p<0.001) Situational Control ((t(132)=1.138, p=0.129) Importance - (z=2.5681, p=0.005)	Partially Reject the hypothesis (Rejected the null hypothesis for four sub hypothesis – Goal congruence, novelty, predictability and importance) Two dimensions, Situational Control and Pleasantness, - null hypothesis was retained

5.9.8. Hypothesis 7

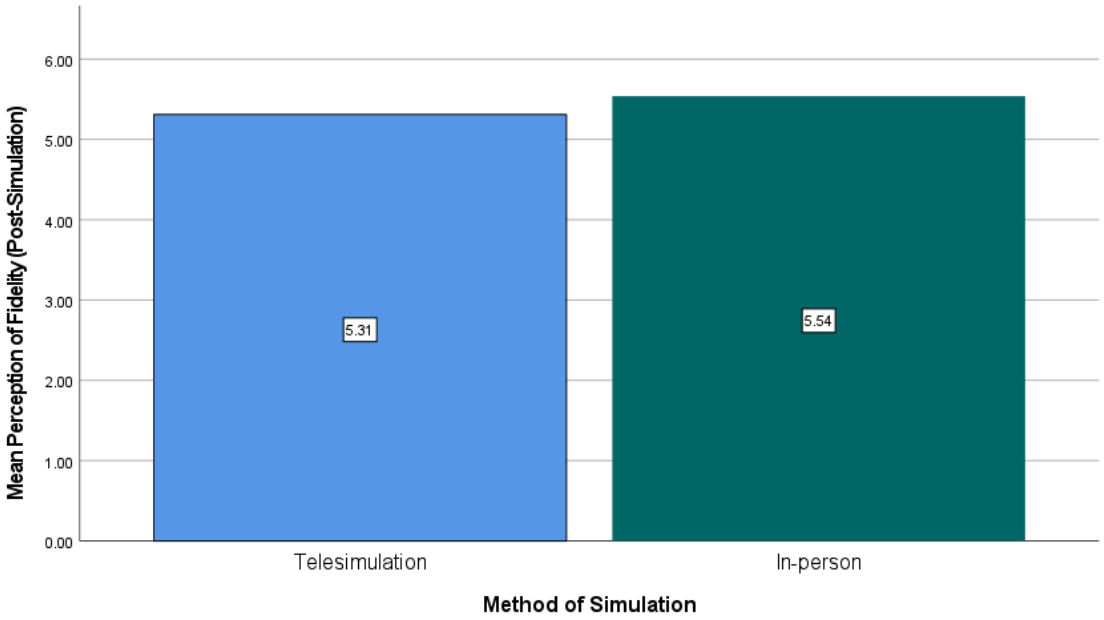
The seventh hypothesis used a nonequivalent group design to compare the perception of fidelity/realism and sense of presence between in-person simulation and telesimulation. In this hypothesis, we also employed a non-inferiority approach. The objective of this hypothesis was to test whether telesimulation is non-inferior to in-person regarding the perception of realism, fidelity, and sense of presence. Details of the hypothesis, including the tests performed and additional information, are presented in Table 121:

Table 121: Summary of included variables and testing procedures for Hypothesis 7

Research Question: Is the perception of realism, fidelity, and sense of presence in telesimulation not inferior to in-person simulation?
Objective: Demonstrate that the perception of fidelity/realism in telesimulation is not inferior to in-person simulation.
Hypothesis 7
H7 (**null):** The perception of fidelity/realism and sense of presence in telesimulation **is inferior to** in-person simulation
Alternative Hypothesis - The perception of fidelity/realism and sense of presence in telesimulation **is not inferior to** in-person simulation
H7a: Perception of fidelity/realism in telesimulation **is inferior** to in-person simulation
Alternative Hypothesis - Perception of fidelity/realism in telesimulation **is not inferior** to in-person simulation
H7b: Sense of presence in telesimulation **is inferior** to in-person simulation
Alternative Hypothesis - Sense of presence in telesimulation **is not inferior** to in-person simulation
Experimental Design: Non-equivalent group design
Independent variable (s): Method of simulation (in person vs telesimulation)
Dependent variable(s): Perception of Fidelity Scale and Sense of Presence
Statistical Analysis to test the hypothesis: Two Sample T-Test (Equal-Variance T-Test for Non-Inferiority, Aspin-Welch Unequal-Variance T-Test for Non-Inferiority and Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority)
Rejection rule: Reject if p-value is less than 0.05 level of significance
One-tailed / two-tailed or non-inferiority: non-inferiority (one-tailed)

For the seventh hypothesis, we compared the mean perception of fidelity between telesimulation and in-person approaches. Results are available in Figure 83:

Figure 83: Perception of Fidelity – Telesimulation vs. In-person



When comparing numerically, the mean plot of the perception of the fidelity reveals that responses evoked from the telesimulation approach are slightly lower than those evoked from the in-person approach.

We conducted a non-inferiority test to verify the statistical difference. First, we tested the normality assumption and equal variances for the perception of fidelity variable and then conducted the non-inferiority tests as displayed in Table 122:

Table 122: Test for non-inferiority for Telesimulation – Perception of Fidelity during Simulation

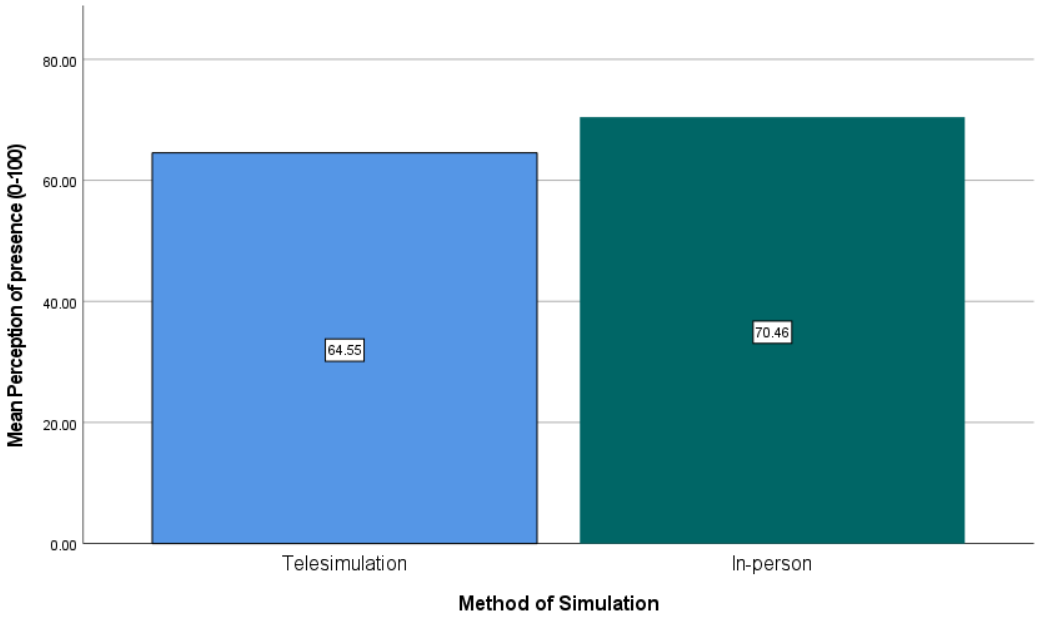
Perception of Fidelity		**NInf Hypothesis: Telesimulation > In-person - 0.72**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	5.311	1.175	0.149	2.000	5.013	5.610
In-person	72	5.539	0.693	0.082	1.994	5.376	5.702
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-0.2277	0.164	1.978	-0.552	0.097
Unequal		95.66	-0.2277	0.170	1.985	-0.566	0.110
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.72	-0.228	0.164	3.000	132	0.002	Yes	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.72	-0.228	0.170	2.893	95.66	0.002	Yes	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.72			3.787		0.000	Yes	
Shapiro-Wilk	0.8519	0	Yes				
Modified-Levene	10.9096	0.00123	Yes				

We found that the perception of fidelity was non-normal (W=0.8519, p<0.001) and that the assumption of equal variances was violated (F=10.9096, p-0.0012). The Mann-Whitney was significant (z=3.7872, p=0.001) for the test of non-inferiority of perception of fidelity, comparing telesimulation (M=5.311, SD=1.175) and in-person (M=5.539, SD=0.693).

The same results were obtained when using the t-test at a 1% level, supporting the acceptance of the non-inferiority alternative, meaning that telesimulation is non-inferior to the in-person approach based on the perceived fidelity of the simulation scenario.

The following variable tested during this hypothesis was the perception of presence. First, we plotted the variable against the telesimulation and in-person, with the results available in Figure 84:

Figure 84: Perception of Presence (0-100) – Telesimulation vs. In-person



By conducting a visual comparison, the in-person approach displayed higher mean results. However, we proceed with the analysis to test the statistical differences using the non-inferiority approach. To complete this process, we tested the variable for normality and equality of variances. We found that the variable had non-normality ($W=0.9470$, $p<0.00995$) with equal variances between telesimulation and in-person ($F=0.0078$, $p=0.92965$), indicating the usage of the non-parametric testing method.

Table 123: Test for non-inferiority for Telesimulation – Perception of presence during Simulation

Perception of Presence		**NInf Hypothesis: Telesimulation > In-person - 9.16**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	64.548	22.722	2.886	2.000	58.778	70.319
In-person	72	70.458	22.619	2.666	1.994	65.143	75.774
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-5.909	3.927	1.978	-13.678	1.858
Unequal		128.89	-5.909	3.929	1.979	-13.683	1.863
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 9.16	-5.910	3.927	0.828	132	0.205	No	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	Df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 9.16	-5.910	3.929	0.827	128.89	0.205	No	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 9.16			0.699		0.2424	No	

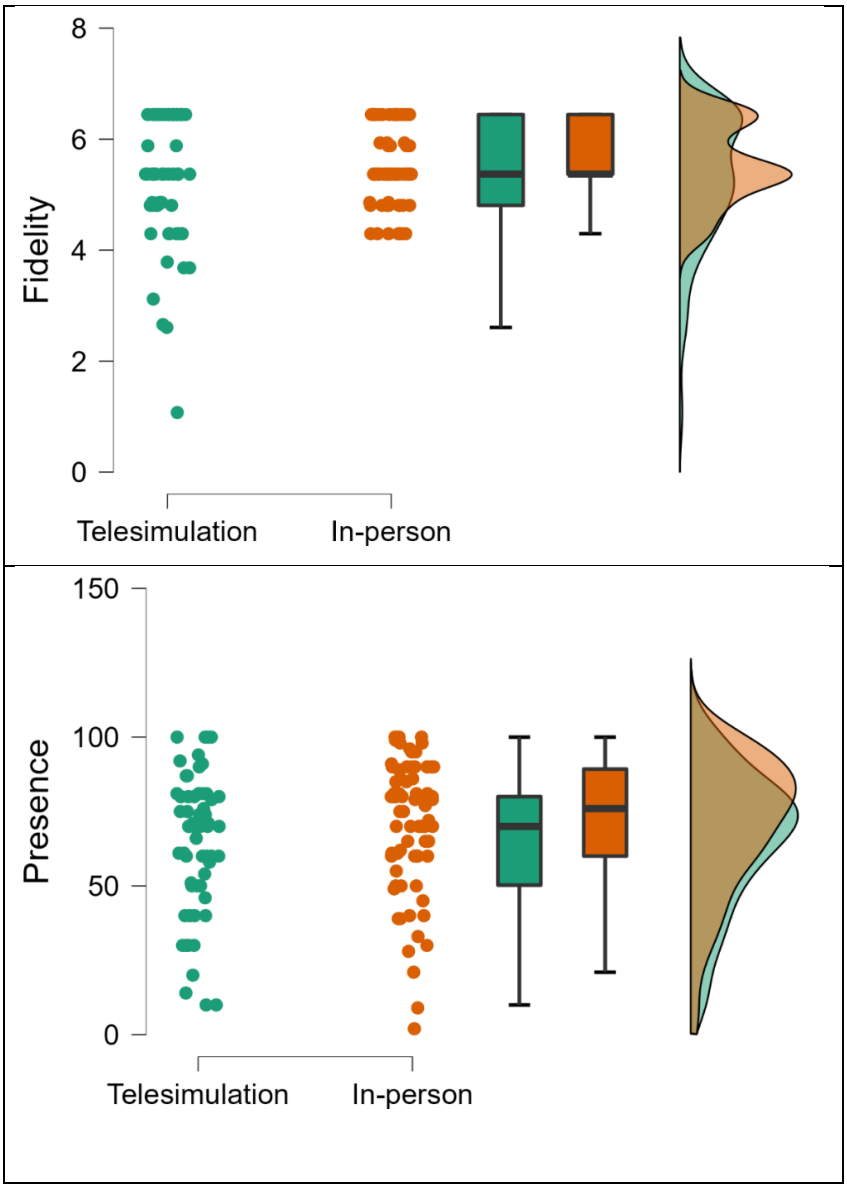
For the Student t-test, location parameter is given by mean difference. For the Mann-Whitney test, location parameter is given by the Hodges-Lehmann estimate.

Shapiro-Wilk	0.9473	0.00995	Yes
Modified-Levene	0.0078	0.92965	No

The Mann-Whitney test results were not significant (z=0.699, p=0.2424) for the non-inferiority of perception of presence, indicating that telesimulation (M=64.548, SD=22.722) was inferior to in-person (M=70.458, SD=22.619). The other parametric t-tests also have not reached a significant result. We retained the null hypothesis that telesimulation is inferior to the in-person approach to the perception of presence during Simulation.

We illustrated the comparisons of this hypothesis with a raincloud plot comparing the density distribution, boxplots, and scatter points for telesimulation and in-person in the dimensions of fidelity and presence:

Figure 85: Raincloud plots for the variables of Fidelity and Sense of Presence



Finally, we summarize the tests performed in this item, highlighting the hypothesis, decisions, and related statistics:

Table 124: Summary of Results for Hypothesis 7

Hypothesis	Statistic test result (p-value)	Decision
H7 (**null):** The perception of fidelity/realism and sense of presence in telesimulation **is inferior to** in-person simulation	Tested in subhypothesis	Partially Reject the hypothesis (One Subhypothesis rejected and one subhypothesis retained)
H7a (**null**): Perception of fidelity/realism in telesimulation **is inferior** to in-person simulation	(z=3.7872, p=0.001)	Reject the hypothesis
H7b(**null**): Sense of presence in telesimulation **is inferior** to in-person simulation	(z=0.699, p=0.2424)	Retain the hypothesis

POST HOC ANALYSIS

Post hoc tests were executed to compare the sense of presence and fidelity between telesimulation and in-person. Tests were carried out to verify if the failure to reject the non-inferiority would be detected as a difference between the groups. The results are available in Table 125:

Table 125: Test of differences in the sense of Presence and Fidelity between in-person and telesimulation

Variable	Test	Statistic	df	p
Sense of Presence (0-100)	Student	-1.505	132	0.135
	Mann-Whitney	1873		0.109
Fidelity	Student	-1.388	132	0.168
	Mann-Whitney	2097.5		0.536

Sense of Presence and Perception of Fidelity variables did not display any differences when comparing the groups of telesimulation and in-person. Again, this finding supports the testing for non-inferiority to compare the two approaches and determine the equality or inferiority of a given method. In this case, the t-tests for comparison failed to reject the differences hypothesis, and one might have concluded that there are no differences between the groups.

The approach of non-inferiority allowed us to determine that the fidelity was non-inferior in telesimulation, but the sense of presence was inferior.

5.9.9. Hypothesis 8

The final hypothesis compared the feedback scores between in-person and telesimulation, using a nonequivalent groups design using a between-subjects comparison. In this hypothesis, we also employed a non-inferiority approach.

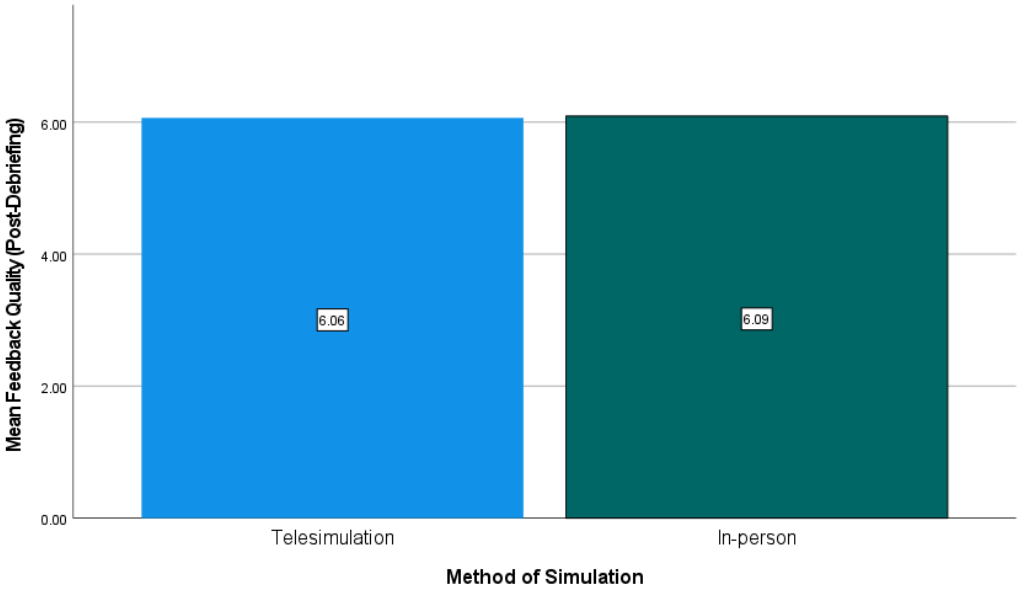
This hypothesis tests whether telesimulation is a non-inferior approach to in-person to provide debriefing after the simulation session. Details of the hypothesis, including the tests performed and additional information, are presented in Table 126:

Table 126: Summary of included variables and testing procedures for Hypothesis 8

Research Question: Is telesimulation not inferior to in-person in debriefing feedback?
Objective: Demonstrate that debriefing conducted by telesimulation is not inferior to in-person simulation.
Hypothesis 8
H8 (**null**): The debriefing feedback score in telesimulation **is inferior** to in-person simulation
Alternative Hypothesis - The debriefing feedback score in telesimulation **is not inferior** to in-person simulation
Experimental Design: Non-equivalent group design
Independent variable (s): Method of simulation (in person vs telesimulation)
Dependent variable(s): Feedback Debriefing Scale
Statistical Analysis to test the hypothesis: Equal-Variance T-Test for Non-Inferiority, Aspin-Welch Unequal-Variance T-Test for Non-Inferiority and Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority
Rejection rule: Reject if p-value is less than 0.05 level of significance
One-tailed / two-tailed or non-inferiority: non-inferiority (one-tailed)

We compared the mean perception of feedback quality for the eighth and final hypothesis, comparing telesimulation and in-person modalities. Results are available in the figure in the sequence:

Figure 86: Feedback quality – Telesimulation vs In-person



The results are numerically close and suggest similar results for both approaches. In the sequence, non-inferiority tests were to determine the non-inferiority of the quality of feedback during the debriefing period between the telesimulation and the in-person. We initially tested for the assumptions of the tests, namely normality and equality of variances:

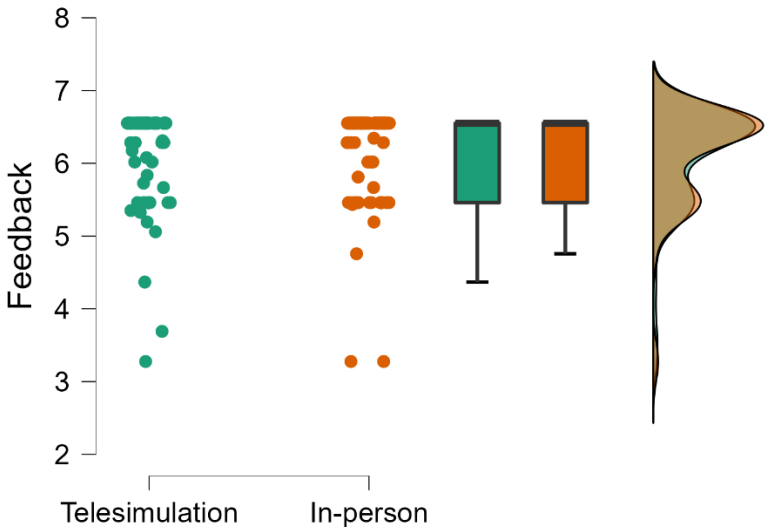
Table 127: Test for non-inferiority for Telesimulation – Feedback Quality - Debriefing

Feedback Quality		**NInf Hypothesis: Telesimulation > In-person - 0.792**					
Comparison	Count	Mean	Std Dev	Std Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Telesimulation	62	6.065	0.716	0.091	2.000	5.883	6.247
In-person	72	6.093	0.695	0.082	1.994	5.930	6.256
Confidence Intervals for the Mean Difference							
Variance Assumption		df	Mean Diff.	Std Error	T*	95% LCL of Diff.	95% UCL of Diff.
Equal		132	-0.0279	0.12205	1.978	-0.269	0.214
Unequal		127.84	-0.0279	0.12232	1.979	-0.270	0.214
Equal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.792	-0.028	0.122	6.261	132	0.000	Yes	
Aspin-Welch Unequal-Variance T-Test for Non-Inferiority							
Alternative Hypothesis	Mean Diff.	Std Error	T-Statistic	df	p	Conclude Non-Inferiority at α = 0.05?	
μT > μC - 0.792	-0.028	0.122	6.247	127.84	0.000	Yes	
Mann-Whitney U or Wilcoxon Rank-Sum Location Difference Test for Non-Inferiority							
Alternative Hypothesis			Z-value		p	Conclude Non-Inferiority at α = 0.05?	
LocT > LocC - 0.792			5.8975		0.000	Yes	
Shapiro-Wilk	0.713	0	Yes				
Modified-Levene	0.0521	0.81979	No				

The normality test results indicated a non-normal distribution (W=0.713; p<0.001). Therefore, we selected the Mann-Whitney U test. The test was significant (z=5.8975, p<0.001) for the non-inferiority test of telesimulation (M=6.065, SD=0.716) to in-person simulation (M=6.093, SD=0.695) for the perception of quality of debriefing feedback, demonstrating that the perception of the quality of feedback during telesimulation was non-inferior to the feedback received during in-person simulation.

In the sequence, for a graphical illustration of the comparisons realized, we present a raincloud plot comparing the density distribution, boxplots, and scatter points for telesimulation and the dimension of feedback quality:

Figure 87 - Raincloud plots for the variables of Feedback



The hypothesis tested, the statistics, and the decision of the eighth and final hypothesis is available in Table 128:

Table 128: Summary of Results for Hypothesis 8

Hypothesis	Statistic test result (p-value)	Decision
H8 (**null**): The debriefing feedback score in telesimulation **is inferior** to in-person simulation	Z=5.8975, p-value<0.001	Reject the hypothesis

Chapter 6. Discussion of Results and Conclusions

This chapter presents the study's conclusions grouped by the results of the hypothesis tests. This chapter also presents the conclusions, implications, study's limitations, and a list of recommendations for future research.

6.1. Summary of Results

This research examined emotional and cognitive responses in telesimulation and in-person simulation in low and medium complexity scenarios, comparing the two groups using natural experiments and a non-inferiority approach. We conducted a series of natural experiments to answer our research question: *Is Telesimulation equally engaging, challenging, and emotion-inducing learning experience as in-person simulation?*

The first two hypotheses demonstrated that either method could generate emotional responses independently. Participants experienced an emotional response when interacting with the robotic mannikins and the team members when collocated and interacting in the same conditions using a remotely facilitated method. We rejected the null hypothesis of no emotional response, and the process and implications will be discussed in further detail in the following subitems related to Hypothesis 1 and 2.

The third study hypothesis tested directly whether the telesimulation was a non-inferior method to elicit emotional responses compared to the in-person simulation, also generally able to establish non-inferiority of telesimulation compared to in-person simulation.

The fourth hypothesis investigated the non-inferiority of telesimulation as a simulation method regarding the mental workload of participants. In this hypothesis, we reject the inferiority of the telesimulation in most of the workload demands, except for the physical demand dimension. The implications and reasons for this inferiority will be discussed as well.

The fifth hypothesis investigated the role of scenario complexity in participants' emotional and cognitive responses, anticipating that higher complexity would lead to more robust emotional and cognitive responses in participants. We have partially rejected the hypothesis, with most tests being statistically significant. However, not all the

comparisons showed that higher complexity is associated with more robust responses in all the dimensions.

In the sixth hypothesis, we tested the non-inferiority of telesimulation concerning the emotional appraisal process in the subdimensions of Pleasantness, Goal Congruence, Certainty, Novelty, Predictability, Situational Control, and Importance. We found that, in most dimensions, non-inferiority was established, suggesting that telesimulation is a non-inferior approach compared to in-person simulation.

The seventh hypothesis investigated the non-inferiority of telesimulation regarding the perception of fidelity, realism, and sense of presence. The null hypothesis was not rejected for one key aspect, indicating that we could not establish non-inferiority of telesimulation for presence. However, we demonstrated the non-inferiority of the perception of fidelity.

The eighth, and last hypothesis of this study, tested the non-inferiority of the quality of feedback during the debriefing, successfully rejecting the hypothesis that telesimulation was inferior to in-person simulation regarding the quality of debriefing.

Considering all the hypotheses and the statistical tests, we could answer our research question, establishing that telesimulation is generally a non-inferior method of medical simulation.

Although some aspects evaluated were not determined to be non-inferior, most post hoc analyses indicated that in-person was not considered superior to telesimulation in most aspects, indicating that telesimulation usage is warranted and non-inferior as a method do conduct healthcare simulations.

In the sequence, we will present discussions, conclusions, and implications for each hypothesis.

6.1.1. Hypothesis 1

The first null hypothesis was that the in-person simulation did not evoke emotional responses during the three moments between baseline and simulation and between simulation and debriefing for the in-person cohort.

We found significant changes and patterns in the three dimensions of emotions, therefore rejecting the null hypothesis and retaining the alternative hypothesis. The same results were observed when we measured the arousal component as measured by the Felt Arousal Scale (FAS). We measured the emotional responses in three moments (baseline,

simulation, and debriefing) using two models: a discrete model of emotions (with eight pairs) classified in three dimensions and a measure of emotional arousal measured in the same three moments.

This hypothesis was established to test the assumption that there is a significant emotional dynamic during simulation and debriefing, using the ANOVA repeated measures test (a particular case of paired t-tests) with post hoc tests to determine which pairs are statistically different when compared.

We tested the model Fraser et al. (2012) proposed to categorize the eight basic emotions using confirmatory factor analysis in two dimensions: tranquility and invigoration. However, findings from the confirmatory factorial analysis indicated a new dimension, which we labeled: cheerfulness given the items that composed its dimension: contended, happy and elated.

We retained the dimensions of tranquility and invigoration proposed by Fraser et al. (2012). However, some instrument items fit better in the cheerfulness dimension, which seemed to include emotional descriptions with positive and pleasant components. However, a different time pattern seemed to indicate more stable and long-lasting emotions composed of the emotions represented by contended, elated and happy items.

We found three diverse mechanisms for each emotional dimension in terms of temporal patterns. The cheerfulness increased between the baseline and debriefing (post hoc test and $p=0.044$). For the dimension of invigoration, we also found a significant emotional response dynamic, indicating that the in-person simulation does evoke changes in their emotional response, with an interesting pattern, as found in prior research (Fraser et al., 2012).

The emotional dimension of invigoration increased during simulation, starting at 0.407(sd=0.91) on the baseline to 1.59 (SD=0.794) after simulation. However, the debriefing represented the hampering of the invigoration, meaning that participants felt less alert and excited than at the end of the simulation ($p>0.001$). However, they had a higher level than their initial baseline component($p<0.001$).

We found a response in the invigoration dimension, contrary to Fraser et al.'s (2019) study that found no differences. The authors mentioned that they were satisfied to observe that the level of invigoration did not decrease, meaning that their excitement and alertness was maintained during the simulation and debriefing.

Our study detected a significant rise in the level of invigoration during the simulation and a lowering during debriefing, being more reassuring and closer to the expected invigoration during a simulation scenario of the literature.

Although to determine the cause of their lack of an emotional response in the invigoration dimension would be speculative and be attributed to a specific scenario or instructor characteristics. Our findings suggest that the three-dimensional model separating invigoration, tranquility, and cheerfulness with a component of emotional longer-term, less transient emotional response needs further investigation in future confirmatory research.

Although in the same pleasantness axis of the pleasantness-arousal model, a change of a discrete emotion from happy to sad may not be as quickly observable or affected over short periods as emotions such as alert or excitement, for instance. This differentiation can be instrumental for future study of emotional responses.

As it represents the other spectrum, the tranquility and invigoration dimensions had an inverse pattern. Participants' tranquility diminished after the simulation compared to the baseline, meaning that participants generally felt less calm, serene, and relaxed ($p<0.001$). However, their tranquility increased during debriefing, demonstrating a lower activation during the debriefing and positive effects of debriefing on emotional response.

These results were in line with Fraser & McLaughlin (2018) when it came to the hampering of tranquility when participants completed the simulation scenario. However, in our sample, we detected tranquility after the debriefing was higher in our sample, instead of returning to the baseline levels. Although needing further exploration, these results may be due to differences in feedback quality and other factors in our sample.

Interestingly, this effect was also associated with the perception of feedback quality and the tranquility dimension with a positive correlation (spearman rho=0.256; p=0.001), indicating that a higher perception of feedback quality was associated with higher tranquility. A similar effect was also found concerning the Cheerfulness dimensions (spearman rho=0.226; p=0.004). These findings highlight the importance of debriefing to deactivate and generate positive emotional responses, especially after stressful scenarios.

Building upon Fraser's et al. (2018) model, we also investigated participants' arousal during in-person simulation. That construct was not part of their research; it is crucial to understand the emotional responses by separating the pleasantness and arousal

components. We also found statistically significant results indicating a clear pattern of arousal dynamics between simulation and debriefing ($p<0.001$).

The pattern was compatible with the observed in the invigoration dimension, representing a statistically significant higher arousal in simulation (M=4.194; SD=1.285) compared to the baseline (M=3.375; SD=1.156), with significant lower arousal in debriefing (M=3.625; SD=1.272). Again, compared with the discrete model of emotions, these results show a similar pattern but can provide additional information, removing the pleasantness component and precisely measuring the activation level.

In summary, we could discern a clear pattern of emotional responses between the three points, supporting our initial hypothesis and assumption that in-person simulation evoked emotional responses when interacting with robotic mannikins and team members in a collocated environment. This result was an essential assumption for testing the second and third hypotheses.

6.1.2. Hypothesis 2

The second null hypothesis was that the telesimulation did not evoke emotional responses during the three moments of a simulation session, between baseline and simulation, and between simulation and debriefing. This hypothesis was identical to the first hypothesis, except that the second hypothesis tested the proposition in the telesimulation group instead of the in-person cohort.

Likewise, we found significative changes and patterns in the three dimensions of emotions, rejecting the null hypothesis. We also found the same dynamics when we tested the arousal response, agreeing with the first hypothesis results.

This hypothesis was established to test the assumption that there is a significant emotional dynamic during simulation and debriefing in telesimulation, using the ANOVA repeated measures test with post hoc tests to determine which pairs are statistically different when compared.

The rejection of this hypothesis and the rejection of the first hypothesis allowed us to proceed to the third hypothesis. Once we confirmed that both in-person and telesimulation generated emotional responses, we could proceed to test the non-inferiority of telesimulation.

For the telesimulation group, we found a qualitatively similar pattern to each emotional dimension compared to the in-person method. The cheerfulness increased between the baseline, simulation, and debriefing ($p=0.004$).

Post hoc tests indicated a significant change between the debriefing and simulation periods and the baseline (start of the training). We also found a significative emotional response dynamic($p<0.001$). The in-person simulation does evoke changes in their emotional response, with the same pattern observed in the in-person group.

The invigoration dimension increased during simulation, moving from 0.31 to 1.20 after simulation (post hoc comparison $p<0.001$). The debriefing response indicated a lower invigoration, showing that participants felt less alert and excited than at the simulation's end ($p<0.001$). However, their invigoration was higher than their initial baseline component ($p<0.001$).

We also detected a significative response in the invigoration, contrary to Fraser et al. (2019) and in agreement with the in-person group. Our study detected a significant rise in invigoration during the simulation and a lowering during the debriefing, the same pattern as the in-person cohort.

The statistical differences between the two groups were tested during this hypothesis since this was the objective of the third hypothesis when we tested the non-inferiority. However, the visual inspection suggested similarities between the two approaches.

The tranquility dimension, complementing the invigoration dimension, also had an inverse pattern in the telesimulation group. Participants' tranquility diminished after simulation compared to the baseline, meaning that, generally, participants felt less calm, serene, and relaxed ($p<0.001$). However, their level of tranquility increased during debriefing, demonstrating a lower activation during debriefing ($p<0.001$) and positive effects of debriefing on the emotional response (Fraser et al., 2018).

These results align with Fraser & McLaughlin (2018) and our in-person group. Still, in our case, we have detected that tranquility after the debriefing was higher in our sample instead of returning to the baseline levels as in Fraser & McLaughlin (2018). However, it agreed with the in-person group, demonstrating the tranquility induction of the debriefing phase.

Regarding participants' arousal during telesimulation simulation, we found statistically significant results indicating a clear pattern of arousal dynamics between simulation and debriefing, like the pattern encountered in the in-person group.

The pattern was similar to the observed pattern on the invigoration dimension. We found statistically higher arousal in simulation (M=4.40; SD=1.063) compared to the baseline (M=3.27; SD=1.089) ($p<0.001$), with significantly lower arousal in debriefing (M=3.726; SD=1.089).

In summary, as stated in our alternative hypothesis, telesimulation demonstrated the capacity to evoke emotions when interacting with the robotic mannikins and team members.

Moreover, we extracted a similar temporal pattern of emotional responses between the three points, supporting our initial hypothesis and suggesting that telesimulation, more than simply evoking emotional responses, does evoke it similarly to in-person simulation. The third hypothesis presented in the sequence tested this proposition statistically properly.

6.1.3. Hypothesis 3

The third null hypothesis was that telesimulation was an inferior approach to evoke emotional responses in simulation and debriefing. We rejected the null hypothesis and retained the alternative hypothesis that telesimulation is not an inferior treatment compared to in-person simulation regarding emotional responses during simulation.

In summary, telesimulation demonstrated the ability to evoke emotions, as demonstrated in our hypothesis second hypothesis. Moreover, we extracted a similar temporal pattern of emotional responses between the three points, supporting our initial hypothesis and suggesting that telesimulation, besides evoking emotional responses, similar to in-person simulation when evaluated qualitatively by observing the graphic plots.

The third hypothesis presented in the sequence tested this proposition in a statistically acceptable way by employing non-inferiority testing.

The first component of emotions – level of arousal – in both debriefing and simulation was compared between both alternatives. When observed regarding telesimulation, in the second hypothesis, we demonstrated the changes in arousal levels during the simulation and telesimulation. In the third hypothesis, we rejected the non-inferiority hypothesis, retaining the alternative that telesimulation is non-inferior compared to in-person on the arousal response variables.

More specifically, when tested in pairs, we rejected the hypothesis of inferiority for debriefing when utilizing a non-parametric method. In contrast, all three tests rejected the null hypothesis in the arousal response in simulation, retaining the non-inferiority of telesimulation for the arousal during stimulation.

The first dimension of emotions - cheerfulness- had a similar pattern in both methods. The telesimulation had a higher cheerfulness during the debriefing than the in-person simulation.

When testing for non-inferiority, we found mixed results, considering the three mean comparison tests and alternative tests we used. The test is more robust to non-normality and heterogeneity of variance rejected the null hypothesis, while the other two did not.

In the sequence, the invigoration dimension had a similar temporal pattern, with telesimulation having a numerically higher increase during the simulation and retaining the invigoration at a higher level after the debriefing than in the in-person simulation.

Finally, the third dimension – tranquility -representing the items of calm, serene and relaxed, did diminish during the simulation in both modalities, in direct response to the invigoration dimensions as expected. The practical reading of this is that as the participants get more engaged in the simulation scenario, they get less relaxed, calm, serene, and more alert and excited.

However, in the non-inferiority comparisons, we failed to reject the hypothesis that telesimulation was an inferior approach to participants' tranquility. It is important to note that the comparison is not performed between final values. However, the actual response is the difference between two temporal data points (baseline and simulation for a response during simulation and simulation and debriefing for the response during debriefing).

The p-value for the non-parametric test was $p=0.06$, very close to our rejection value, suggesting that in a slightly different sample, the non-inferiority could be attained. Finally, post hoc tests did not detect enough evidence to reject the hypothesis that the mean of the two groups is different.

When considering a group of variables, we can partially reject the hypothesis that telesimulation is an inferior approach to inducing emotions. However, while some specific variables could not be determined to be non-inferior, post hoc analysis demonstrated no statistically significant difference between the groups.

When emotions were analyzed as a whole, including the three dimensions and the arousal components, the response was compatible, displaying a similar pattern of increases and hampering during simulation and debriefing.

6.1.4. Hypothesis 4

The fourth null hypothesis was that the telesimulation was inferior to in-person mental workload, measured by the NASA-TLX instrument. In this hypothesis, the mental workload was decomposed into diverse dimensions to provide a clearer insight into the impacts and in which dimensions the workload was inferior and which it was not.

Although the usage of the NASA TLX instrument is often reported only as the raw score obtained by summing all the dimensions (Hart, 2006), we used the two-dimensional model divided into task demand (divided into mental, temporal, and physical subdimensions) and the related effect of that demand on the subject (perception of performance, effort, and frustration).

Finally, the other adjustment performed during the planning phases was excluding the physical component of the task demand, given that telesimulation does not involve the direct physical demand and interaction with the robotic mannikins. This decision was taken on two grounds: a) a practical ground since it was evident that telesimulation did not involve a physical component comparable to an in-person experience; b) previous research demonstrated that the weight attributed to the physical demand tends to be the least, in the complete NLX instrument.

Separating the physical demand proved essential to differentiate the aspects of both types of simulation and give insights into how the telesimulation is non-inferior to telesimulation when the physical component was excluded from the analysis.

We rejected the hypothesis of the inferiority of the fourth hypothesis. However, the components with a physical component added to their score, suggesting that the physical component was the single differentiator factor between the two alternatives. For instance, the numeric values were identical in the NASA RAW scale (telesimulation=18.03 and in-person=18.03) and numerically higher in demand on the subject (telesimulation=9.98 and in-person=9.75) when the physical dimension was excluded from the analysis.

Again, the impossibility of establishing non-inferiority for the NASA workload scale in the physical dimension, although not surprising, brings a rich opportunity for

discussion. Several studies have determined that physical demand is often the lowest-ranked dimension (less demanded aspect) in mental workload (Hart, 2006; Said et al., 2020; Virtanen et al., 2021). However, it is essential to build the capacity to conduct remotely facilitated simulations to allow the exposure of participants to the physical elements of the simulation.

The use of tangible interfaces (Larnpotang, 2013), even for scenario-based (as opposed to skill-based simulation) simulation, is a frontier that needs to be explored (as tangible elements simulation) for telesimulation. Other human-computer interaction research fields have already substantially established the importance of tangible interfaces, therefore, having some elements that bring some physical demand to the participant.

One apparent contradiction in our choice to consider the higher workload as a "higher the better" outcome needs to be explained and discussed in this section in light of the results.

The mental workload is usually a construct that the objective is to be minimized when testing new interfaces and applications; that is, the less mental workload to use an interface, generally the better (Moray, 2013), and widely understood that high levels of mental workload are associated with poor performance (Yurko et al., 2010).

In the case of testing for the non-inferiority, we approached the problem using a different approach and foci of analysis. A higher mental workload in utilizing a computer interface or a robotic-human interaction usually points to design and interaction problems. However, in a simulation scenario, the objective is often to impose a high mental workload, challenging and imposing complex scenarios –close to the challenging reality in hospital and emergency centers- situations that make the participant work in a more complex situation instead of an easier, less challenging scenario.

That is, by introducing factors that increase the mental workload, such as equipment failures, conflicting information, problematic patients, lack of communication, and complex diagnoses, the overall workload is increased with the clear and deliberate purpose of exposing students to situations that would place higher cognitive demands.

By establishing that telesimulation is non-inferior to in-person mental workload, we could demonstrate that both are "equally" perceived as challenging to participants, therefore attaining a vital objective of the simulation: to evoke emotional and cognitive responses in simulation that are close to real-life.

Future research on human-computer interaction and simulation must explore the role of cognitive load better, not only as an adverse or detrimental construct. When exposing participants to training, the inter-relationship and optimal levels aim to impose a higher cognitive load to prepare them for a future challenging situation.

6.1.5. Hypothesis 5

The fifth null hypothesis was that the scenario complexity was unrelated to emotional and cognitive responses during the simulation and debriefing. We rejected the null hypothesis, retaining the alternative hypothesis that scenario complexity was related to higher emotional and cognitive responses in our sample, considering in-person and telesimulation. We also conducted post hoc tests to verify the emotional appraisal process differences in both groups, with some significant results.

Starting with the emotional responses, it was interesting to observe the pattern of responses between simulation and debriefing for the arousal dimension, with the difference observable in magnitude.

In low complexity scenarios, the response on low complexity was (0.72) while arousal in medium complexity was (1.22). Likewise, the deactivation (diminishing arousal) had a similar pattern, indicating that the debriefing provided a “calming” effect in a more complex scenario. Both variables had statistical differences, indicating that the higher complexity scenarios were associated with more robust arousal responses (higher in the simulation case and smaller (or more negative) in the debriefing phase).

In the sequence, when we tested the difference between the emotional response in the three dimensions of cheerfulness, invigoration, and tranquility, we only found differences in the Tranquility Dimension during simulation ($p=0.014$). All other factors were not significantly different according to our tests.

This finding was not expected but highlighted the importance of separating the emotional response into the discrete emotion models and the arousal component concerning the level of activation or arousal as an essential component of emotions.

Medium complexity scenarios were associated with a higher activation, but not necessarily different emotional patterns, supposedly in the negative-positive spectrum.

This finding suggests that higher complexity scenarios do not impact the pleasantness dimension from a logical standpoint. However, they evoke higher charged emotions as measured by the arousal components.

These findings consolidate and support the usage of more fine-grained emotional models to account for the separation of pleasantness and arousal components, with essential applications to education and the study of emotional responses in human-computer interaction, which need to separate these two components.

The next block of tests concerned the dimensions of mental workload in all its dimensions, as specified in the NASA-TLX instrument, and our approach to removing the physical component to evaluate the workload on five distinct dimensions. Differently from what happened in the emotional responses, the effect of a higher complexity on mental workload was more pronounced. We found a statistical difference between the medium and low complexity scenarios.

The complexity of the scenario was associated with higher task demand (Temporal, Mental and Physical demand). Likewise, the task's demands on the subjects (Frustration, Effort, and Performance) were higher in the higher complexity scenarios. This result is expected and aligned with the previous literature but reveals the discriminant validity of the instruments utilized (when comparing emotions and cognitive load) and validates the expected results by establishing the face validity of the instruments, indicating that the mental workload instrument is measuring what it is supposed to measure.

The final block of variables tested, as a post hoc analysis, revealed that, in most cases, a higher scenario complexity was associated with higher goal congruence, more novelty, more importance, less predictability, higher perception of fidelity, and sense of presence. However, the complexity of the scenario, or more specifically a higher complexity, was not associated with a higher value on the variables of pleasantness, certainty, and situational control and the quality of debriefing feedback received.

These findings represent mixed results in terms of expected results, mainly related to novelty, fidelity, predictability, and presence, suggesting that more complex scenarios had more challenging aspects, having a more immersive component than the lower complexity scenarios. However, we did not find the same effect on other variables.

The complexity did not seem to associate with the pleasantness, quality of debriefing feedback and situational control. These are aligned with expectations; inherently, there is no strong suggestion that higher complexities evoke this appraisal response.

In summary, we reject the null hypothesis that there is no positive association between emotional and cognitive responses, with a significant association, although not

in all components of emotional responses. Likewise, the difference in dimensions related to the appraisal process mainly was aligned with the expected results.

6.1.6. Hypothesis 6

The sixth null hypothesis was that the telesimulation had an inferior appraisal process compared to in-person simulation. Our null hypothesis stated that the telesimulation was deemed less pleasant, less important, and less predictable, with less novelty in the situations and lower situational control. Those items are desirable in general educational activities, human-computer interaction, and medical education.

We rejected the null hypothesis for five of the seven dimensions, retaining the alternative hypothesis that telesimulation is non-inferior to in-person simulation in goal congruence, certainty, novelty, predictability, and importance dimensions. On the other hand, we could not reject the hypothesis of the inferiority of telesimulation to in-person simulation for the dimensions of pleasantness and situational control.

In most emotional appraisal processes, it is essential to establish the non-inferiority of telesimulation on critical dimensions often used in human-computer interaction and educational studies. The dimensions of goal congruence, certainty, novelty, predictability, and importance are fundamental to understanding human-computer interaction (Jeon, 2017). They are demonstrating that these dimensions are non-inferior support the usage of telesimulation and its capacity to provide novel, critical activities congruent with participants' learning goals.

Similar to procedures adopted in the previous hypothesis tests, we also conducted post hoc tests to compare the emotional appraisal process in both groups, testing for both the equality and the hypothesis of a higher mean of in-person approach, with some significant results. For this hypothesis, we have found more means between approaches, confirming that the dimensions of pleasantness, goal congruence, and situational control were lower in telesimulation than in person.

Comparing the approaches to tests for differences (non-inferiority and simple means comparison) we have one result not present in both lists. The Goal congruence was lower in telesimulation but not inferior. Recalling the reasons for adopting non-inferiority testing, this case illustrates the importance and practical usefulness of adopting this approach. Although the goal congruence difference of groups is greater than zero, it is not greater than the margin defined (13%), therefore considered not practically relevant.

Still, the dimensions of pleasantness and situational control were lower in telesimulation compared to in-person. These dimensions are also widely pointed out as critical factors when determining the emotional responses during human-computer interactions and interfaces (Jeon, 2017).

Given the differences between the in-person and telesimulation, we infer those participants find it more enjoyable to interact with in-person manikins instead of working remotely without essential elements, mainly tangible, otherwise present when co-located in the simulation center. Likewise, the situational control is plausibly lower when they are working remotely in comparison since participants of the simulation cannot physically take control of the situation, and act directly on the robotic patient, compared to the remotely facilitated version, so the perception of lower situational control is warranted and justified.

This diminished sense of control is not surprising, given prior results in telesimulation (Patel et al., 2020; Naik, 2020) and telemedicine (Mackintosh et al., 2016; Almathami et a., 2020), indicating barriers to these approaches when interacting with remote patients. The two approaches have a similar distancing between the "patient" and the healthcare provider, and these findings support the differences between in-person and remotely facilitated options.

These limitations seem to be inherent to the remote methods. They should be addressed by either future forms to increase the sense of control and pleasantness by new technologies or to adjust the remote sessions to provide a higher sense of control and pleasantness.

It is essential to discuss the limitations of the scale to measure the emotional appraisal process. Notably, the dimensions related to situational control proved problematic, perhaps to the dubiety of items that asked the participants if they felt someone was controlling the events and if events were beyond their control.

In hindsight, it is now more evident that those questions may have been confusing or ambiguous since the items were generally constructed for general activities to measure how the respondent felt that they had control over a given process. However, in simulation, someone else controls the environment, and the participant does not control the sequence of events, one of the fundamental aspects of situational control.

Future research could tackle this problem by clarifying it to the participant or creating new constructs that can measure perceived control levels without ambiguities related to researching a simulated environment. Therefore, a situational control measure

for the specific simulation case could be developed to capture this vital aspect of the emotional appraisal process in this specific context.

6.1.7. Hypothesis 7

The seventh hypothesis employed a non-inferiority approach to compare in-person and telesimulation. The objective of this hypothesis was to test whether telesimulation was non-inferior to in-person simulation regarding the perception of realism, fidelity, and sense of presence.

The null hypothesis of the inferiority of sense of presence was rejected, allowing us to retain the alternative hypothesis of non-inferiority. Although smaller numerically, the sense of presence is "equivalent" in both methodologies, supporting the continued use of telesimulation, not only in the context of COVID-19 and as a method to improve access to simulation to remote areas and countries with fewer resources, democratizing the healthcare education.

The perception of fidelity was considered non-inferior, and the sense of presence was not considered non-inferior. The post hoc tests found no difference, and therefore the superiority of sense of presence could not be established, furnishing a strong indication that although not statistically possible to establish non-inferiority, there is a reasonable suggestion that both methods have a similar sense of presence, given that we failed to reject the hypothesis for non-inferiority.

Also, this finding of possible inferiority of telesimulation about the sense of presence makes clear that advancements to improve the sense of presence and fidelity need to be envisioned for telesimulation. Also, given our sample size, it can be speculated that a larger sample size would reject the hypothesis of inferiority with more statistical power associated.

Although this is not as extensively researched in telesimulation as in in-person simulation, the sense of presence is crucial in healthcare simulation. A search for HCI methods and technologies to improve the sense of presence is an important endeavor, especially since the development of new mechanisms can easily cross-pollinate the field of telemedicine, a closely related field.

6.1.8. Hypothesis 8

The last hypothesis had its primary objective to test if telesimulation was not an inferior approach to conducting debriefing compared to in-person debriefing. We rejected the null hypothesis, therefore retaining the alternative hypothesis that telesimulation is not inferior to providing feedback after debriefing treatment than in-person simulation.

Although the hypothesis was the last in the sequence, the importance of debriefing is confirmed as a fundamental aspect of healthcare simulation (Fraser & McLaughlin, 2018). When comparing the two approaches, the feedback dimension was non-inferior on telesimulation, achieving an almost identical numerical value between the two alternatives (telesimulation=6.06 vs. in-person=6.09).

These findings indicated that the instructors could deliver a similar quality of feedback using telesimulation compared to the in-person feedback. Connecting this to the earlier tests that indicated the calming effect of debriefing during the debriefing moment, with similar effects between in-person and telesimulation, in line with previous studies (e.g., Roach & Okrainec, 2021; Padhya et al., 2021).

In practical terms, these findings indicate the viability of conducting feedback remotely without co-located members and the visual and emotional cues that are harder to apprehend when a meeting or a discussion is conducted online during healthcare simulation.

Even though the finding was specific to telesimulation, this suggests that other forms of interaction that are mediated by remote technologies can also obtain their objectives in relationship with the perceived quality of feedback, with applications in diverse areas that are mediated by remotely facilitated technologies such as work and educational environments, healthcare applications, among others.

The increasing usage of telemedicine in emotional-intensive activities is another aspect that supports this conjecture of the feasibility and quality of feedback delivered remotely facilitated in similar settings.

Although the objective of this hypothesis was not to evaluate the quality of the feedback objectively but to compare the two methods and determine the non-inferiority of telesimulation, it is relevant to mention that both modalities had values close to 6 on a theoretical scale of 7, indicating very high perception of the quality of the feedback.

6.2. Final Conclusions

The results obtained by the several parametric and non-parametric tests described in both the data analysis and hypothesis testing support the use of telesimulation as a non-inferior approach to in-person healthcare simulation. This research presented findings, validations, and contributions to a highly multidisciplinary field, as demonstrated in the bibliometric analysis, the field of Healthcare Simulation.

We employed an approach centered on Informatics and the framing using the Human-Computer Interface theoretical lens, with a foundation in the affective computing approach, based on the crucial importance of understanding the emotions in both in-person and remotely facilitated human-computer interactions, as in the case of telesimulation.

The present study expands the study of emotional and cognitive responses to computer and robotic interfaces, pioneered by studies such as Piccard (2003), Larkin et al. (2010), Fraser et al. (2012), Baños et al. (2014), among many others, to the healthcare simulation to the specific case of telesimulation.

Although the findings and insights generated from this research are related to telesimulation, the findings support and incentivize further research in at least three other distinct but entangled fields: education, work, organizational dynamics, and telemedicine.

Living during the 2019-2022 period, all of us have transformed our lives by the remoteness of work, personal lives, and education, and medical education was not exempt from it. Most universities, colleges, and schools of all levels have migrated or blended their instructional format with online education.

While this research has effectively demonstrated that remotely facilitated simulations are not inferior to in-person simulations regarding emotional and cognitive responses, including the delivery of feedback during the debriefing sessions, the findings are likely to be reproduced in other forms of remote education, given that although with well-defined particularities, healthcare simulation as a practice and field, is an educational activity.

In retrospect, it is worthy to mention the non-inferiority approach, not in the statistical sense, but in the philosophy of using the test and its assumptions. From the onset of the study, it was clear that the rationale for using telesimulation during COVID-19 was to provide an alternative option to maintain the education of healthcare education in the country and abroad. Therefore, telesimulation would not be expected to be better

than the in-person simulation. Instead, it was tested whether it could effectively be employed without prejudice to learners' educational experience.

Although some cited studies in this thesis have determined the satisfaction and acceptability on the more subjective spectrum of compatibility, a few others have compared more objectively learning outcomes. However, much of the process was not explained, and even less, the direct comparison of the telesimulation and in-person. This research has sought to clarify the process and explain the similarities in cognitive processes, emotional responses, and the viability of debriefing using remotely facilitated procedures.

Another vital aspect also involves using remotely facilitated or conducted activities or procedures in the workplace, given that the remote-working arrangement is a new reality. Recalling that emotions are a crucial construct influencing computer-facilitated work and workplace dynamics, it is essential to understand emotional dynamics further while conducting educational and other online activities, especially technology-mediated realities.

As described in the earlier paragraphs, the physical co-location of participants did not seem to be a prerequisite for the emotional and cognitive response in group activities. These findings have both practical (including new practices and rules for the workplace) and theoretical implications to better understand these dynamics and how they compare to in-person responses.

Considering how healthcare simulation closely mimics the settings experiences in real hospital and field settings, we understand that the responses verified in remotely facilitated simulation may be strongly correlated with the responses found in telemedicine and other types of remote delivery of healthcare.

Considering the rise of available options for convenient, all-day available healthcare services possible by telemedicine, understanding the emotions and cognitive process in telesimulation suggests two things: a) telemedicine may have comparable responses to in-person healthcare services and be an equivalent or at least non-inferior approach.

In this regard, although not found as a common practice in the literature review, the use of telesimulation could also be envisaged to simulate telemedicine scenarios and not only replace traditional in-person scenarios, which may require new adaptations, technologies, and computer-supported environments to increase the ecological validity of the educational scenarios for telemedicine.

Also, the “non-findings” of this research, or hypotheses that were not rejected, that is, we could not determine the non-inferiority of telesimulation, brought some thought-provoking that should be addressed and start a conversation about the current limitations of telesimulation as a teaching approach.

The sense of presence, one of the appraisal dimensions not established to be non-inferior, warrants an essential discussion and consideration. The incapacity to demonstrate the non-inferiority of this dimension brings us to question how a remote facilitation simulation can evoke a similar sense of presence compared to an in-person simulation. Additional technologies or methodologies can improve the sense of presence, a widely studied aspect of virtual and remote applications.

However, it is crucial to note that the sense of presence was very similar between the two approaches. Our tests indicated that we could not establish a statistical difference between the two approaches.

6.3. Limitations and recommendations for future research

The present research, like any scholarly endeavor, had limitations due to the scope and limited term of doctoral research, the historical time in which it was conducted, the difficulty of generalizing results and findings, the challenge of access to participants, sample sizes, and availability of technologies to collect data.

As discussed in the personal statements, in the first chapter of this thesis, before COVID-19, the primary research objective was to employ bodily sensors to measure participants' emotions in healthcare simulation. With the surge of COVID-19 cases and the start of the conduction of telesimulation, remote data collection data was proven too complex and not feasible considering logistics, IRB approval, and participant recruitment during 2020/2021.

The COVID-19 context was simultaneously a limitation and a unique opportunity to get insights into a phenomenon that changed how healthcare education was conducted in the United States and the world.

The data collection was performed in two simulation centers in specific settings. More specifically, the collection was concentrated in one of the simulation centers, diminishing the representativeness of participants. Although efforts have been made to sample representative classes within the available simulation sessions and scenarios, the sampling procedure was not designed to be representative.

Therefore, the conclusions and findings may not generalize to other locations and conditions. We urge future researchers to replicate the research in additional simulation centers and a more diverse range of cases and situations, including more complex cases.

Although our samples were similar in all possible cofounding variables identified in the design and samples, such as experience using simulation, class standing, gender, and experience in clinical settings, we urge future studies to conduct large-scale randomized clinical trials.

This work could be seen as a pilot endeavor to establish and validate measures and combine them into the three critical points of a simulation (baseline, simulation, and debriefing) that call for larger-scale studies.

Another significant limitation is that the research design option does not measure objective learning outcomes, one of the most obvious ways to compare two educational methods. Although that was envisioned in this thesis's earlier stages, this proved not feasible for a series of factors.

First, we needed to consider the thesis approach of using affective computer assumptions to understand emotions in the user-robotic and remotely facilitated technologies. The use of learning outcomes could provide an additional load on participants considering the number of measures already collected.

Secondly, the scoping review demonstrated that there is already sufficient evidence demonstrating the effectiveness and satisfaction of participants with telesimulation. However, learning outcomes could be an exciting aspect and will be explored should this research continues as a multidisciplinary team effort, with simulationists and subject matter experts on learning and simulation.

We found a few practical difficulties. The reduced number of participants in each simulation session was a considerable contributor to the reduced sample size of this research. Although we had reached statistical power to reject the study hypotheses, a larger sample would be essential to perform analysis and crosstabulations between different groups, such as age groups, experience with simulation, class standing, and even specific scenario characteristics.

Another practical difficulty was related to the technologies available to collect data, including using sensors and the surveys in this study. Given the COVID-19 precautions and recommendations during the data collection period, it was not feasible to introduce physiological sensors, despite the researcher being the author's coordinator of the SensorLab (sensorlab.arizona.edu). SensorLab is a state-of-the-art facility at The

University of Arizona – Health Sciences with access to various sensors that can measure participants' bodily responses.

Therefore, researchers are urged to introduce sensor technologies that can measure heart rate, electroencephalogram (EEG), motion and activity level, eye-tracking, and other signals to help understand emotional and cognitive load in future research. The addition of sensors can happen as either a follow-up to this research or independently by other researchers.

However, the usage of the sensors must be complementary, not a replacement to the surveys validated and employed in this study. Using the sensors without the accompanying psychometric measurement may be too speculative, and a combination is necessary to validate and calibrate the correlation of signals and the complex emotional and cognitive responses.

Therefore, we see this study as a precursor to future research by separating key moments for data collection and defining and testing data collection instruments within simulation in-person and telesimulation.

In that regard, this study, by conducting an empirical study in telesimulation and by executing a scoping review, points out in the direction of experimental studies in telesimulation, departing from studying solely the feasibility and user satisfaction with telesimulation to explore more complex phenomena on both telesimulation and in-person healthcare simulation.

Finally, future research should build upon and as a field to make it more common to use experimental and quasi-experimental designs, focusing on statistical testing of constructs, employing and reporting validity, instrument structure, and other consistency, validity, and reliability measures. The proper usage of superiority, equality, or non-inferiority testing, including a careful definition of an adequate margin for testing, should be preferred instead of simply employing t-tests or ANOVA to test equality or superiority.

REFERENCES

Abeysekera, L., & Dawson, P. (2014). Motivation and cognitive load in the flipped classroom: Definition, rationale and a call for research. Higher Education Research & Development, 34(1), 1-14. https://doi.org/10.1080/07294360.2014.934336

Abrahamson, S., Wallace, P., & Hansen, S. (1980). Simulation in medical education: Static and interactive manikins. *Programmed Learning and Educational Technology*, *17*(3), 148-153. https://doi.org/10.1080/0033039800170303

Adamson, K. (2015). A systematic review of the literature related to the NLN/Jeffries simulation framework. Nursing Education Perspectives, 36(5), 281-291. https://doi.org/10.5480/15-1655

Ahmed, F. (2016). Using structured analysis and design technique (SADT) for simulation conceptual modelling. Doctoral dissertation, Loughborough University.

Ahn, J., & Clegg, T. (2018). Human-computer interaction and education. The Wiley Handbook of Human Computer Interaction, 2, 821-830. https://doi.org/10.1002/9781118976005.ch38

Ainley, M., & Ainley, J. (2011). Student engagement with science in early adolescence: The contribution of enjoyment to students' continuing interest in learning about science. Contemporary Educational Psychology, 36(1), 4-12. https://doi.org/10.1016/j.cedpsych.2010.08.001

Al-Elq, A. (2010). Simulation-based medical teaching and learning. Journal of Family and Community Medicine, 17(1), 35. https://doi.org/10.4103/1319-1683.68787

Ali, W. (2020). Online and remote learning in higher education institutes: A necessity in light of COVID-19 pandemic. Higher Education Studies, 10(3), 16. https://doi.org/10.5539/hes.v10n3p16

Akaike, M., Fukutomi, M., Nagamune, M., Fujimoto, A., Tsuji, A., Ishida, K., & Iwata, T. (2012). Simulation-based medical education in clinical skills laboratory. The Journal of Medical Investigation, 59(1, 2), 28-35.

Almousa, O., Zhang, R., Dimma, M., Yao, J., Allen, A., Chen, L., Heidari, P., & Qayumi, K. (2021). undefined. Simulation & Gaming, 52(5), 614-634. https://doi.org/10.1177/10468781211008258

Ananga, P. (2020). Pedagogical considerations of E-learning in education for development in the face of COVID-19. International Journal of Technology in Education and Science, 4(4), 310-321. https://doi.org/10.46328/ijtes.v4i4.123

Anscombe, N. (2015). A different view: In Europe, virtual and immersive technologies are starting to change the landscape of medical simulation. *IEEE Pulse*, *6*(4), 14-19. https://doi.org/10.1109/mpul.2015.2428295

APA Dictionary of Psychology. (2021). American Psychological Association (APA). https://dictionary.apa.org/emotion

Arguel, A., & Jamet, E. (2009). Using video and static pictures to improve learning of procedural contents. Computers in Human Behavior, 25(2), 354-359. https://doi.org/10.1016/j.chb.2008.12.014

Aria, M., & Cuccurullo, C. (2017). Bibliometrix : An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Arias-Aranda, D. (2007). Simulating reality for teaching strategic management. Innovations in Education and Teaching International, 44(3), 273-286. https://doi.org/10.1080/14703290701486662

Aricò, P., Borghini, G., Di Flumeri, G., Colosimo, A., Bonelli, S., Golfetti, A., Pozzi, S., Imbert, J., Granger, G., Benhacene, R., & Babiloni, F. (2016). Adaptive automation triggered by EEG-based mental workload index: A passive brain-computer interface application in realistic air traffic control environment. Frontiers in Human Neuroscience, 10. https://doi.org/10.3389/fnhum.2016.00539

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. International Journal of Social Research Methodology, 8(1), 19-32. https://doi.org/10.1080/1364557032000119616

Armitage, R., & Nellums, L. B. (2020). COVID-19 and the consequences of isolating the elderly. The Lancet Public Health, 5(5), e256. https://doi.org/10.1016/s2468-2667(20)30061-x

Aromataris, E., & Munn, Z. (2020). *Jbi manual for evidence synthesis - Jbi manual for evidence synthesis - Jbi global Wiki*. Digital Object Identifier System. https://doi.org/10.46658/JBIMES-20-01

Arnold, M. B. (1960). Emotion and personality. Vol. I. Psychological aspects.

Artino, A. R. (2009). undefined. Journal of Computing in Higher Education, 21(2), 146-166. https://doi.org/10.1007/s12528-009-9020-9

Baily, L. W. (2019). History of simulation. Comprehensive Healthcare Simulation, 3-11. https://doi.org/10.1007/978-3-030-15378-6_1

Baker, C. (2017). Quantitative research designs: Experimental, quasi-experimental, and descriptive. *Evidence-based practice: An integrative approach to research, administration, and practice*, 155-183.

Bales, R. F. (2017). Social interaction systems: Theory and measurement. Routledge.

Barré, J., Michelet, D., Truchot, J., Jolivet, E., Recanzone, T., Stiti, S., Tesnière, A., & Pourcher, G. (2019). Virtual reality single-port sleeve gastrectomy training decreases physical and mental workload in novice surgeons: An exploratory study. Obesity Surgery, 29(4), 1309-1316. https://doi.org/10.1007/s11695-018-03680-9

Barry Issenberg, S., Mcgaghie, W. C., Petrusa, E. R., Lee Gordon, D., & Scalese, R. J. (2005). undefined. Medical Teacher, 27(1), 10-28. https://doi.org/10.1080/01421590500046924

Barto, A. G. (2013). Intrinsic motivation and reinforcement learning. In Intrinsically motivated learning in natural and artificial systems (pp. 17-47). Springer, Berlin, Heidelberg.

Bediou, B., & Scherer, K. R. (2014). Egocentric fairness perception: Emotional reactions and individual differences in overt responses. PLoS ONE, 9(2), e88432. https://doi.org/10.1371/journal.pone.0088432

Berg, H., & Steinsbekk, A. (2020). Is individual practice in an immersive and interactive virtual reality application non-inferior to practicing with traditional equipment in learning systematic clinical observation? A randomized controlled trial. *BMC Medical Education*, *20*(1). https://doi.org/10.1186/s12909-020-02030-7

Blackstock, F. C., & Jull, G. A. (2007). High-fidelity patient simulation in physiotherapy education. Australian Journal of Physiotherapy, 53(1), 3-5. https://doi.org/10.1016/s0004-9514(07)70056-9

Bogo, M., Katz, E., Regehr, C., Logie, C., Mylopoulos, M., & Tufford, L. (2013). Toward understanding meta-competence: An analysis of students' reflection on their simulated interviews. Social Work Education, 32(2), 259-273. https://doi.org/10.1080/02615479.2012.738662

Bogossian, F. E., Cant, R. P., Ballard, E. L., Cooper, S. J., Levett-Jones, T. L., McKenna, L. G., Ng, L. C., & Seaton, P. C. (2019). Locating “gold standard” evidence for simulation as a substitute for clinical practice in prelicensure health professional education: A systematic review. *Journal of Clinical Nursing*, *28*(21-22), 3759-3775. https://doi.org/10.1111/jocn.14965

Bornmann, L., & Daniel, H. (2007). What do we know about theh index? *Journal of the American Society for Information Science and Technology*, *58*(9), 1381-1385. https://doi.org/10.1002/asi.20609

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H. (2008). Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics*, *8*, 93-102. https://doi.org/10.3354/esep00084

Bradley, M. M., Codispoti, M., Sabatinelli, D., & Lang, P. J. (2001). Emotion and motivation II: Sex differences in picture processing. Emotion, 1(3), 300-319. https://doi.org/10.1037/1528-3542.1.3.300

Bradley, P. (2006). The history of simulation in medical education and possible future directions. Medical Education, 40(3), 254-262. https://doi.org/10.1111/j.1365-2929.2006.02394.x

Bradley, C. S., Johnson, B. K., Dreifuerst, K. T., White, P., Conde, S. K., Meakim, C. H., Curry-Lourenco, K., & Childress, R. M. (2019). Regulation of simulation use in United States Prelicensure nursing programs. *Clinical Simulation in Nursing*, *33*, 17-25. https://doi.org/10.1016/j.ecns.2019.04.004

Brave, S., & Nass, C. (2007). Emotion in human-computer interaction. In The human-computer interaction handbook (pp. 103-118). CRC Press.

Browner, W. S., Hulley, S. B., & Newman, T. B. (2001). Getting ready to estimate sample size: hypotheses and underlying principles. *Designing clinical research. Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB, eds*, *121*(3), 498-499. https://doi.org/10.1016/s0022-3476(05)81819-0

Budd, J. M. (1988). A bibliometric analysis of higher education literature. *Research in Higher Education*, *28*(2), 180-190. https://doi.org/10.1007/bf00992890

Byrne, B. M., & Van de Vijver, F. J. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of Nonequivalence. *International Journal of Testing*, *10*(2), 107-132. https://doi.org/10.1080/15305051003637306

Campbell, J. (2005). Précis of reference and consciousness. Philosophical Studies, 126(1), 103-114. https://doi.org/10.1007/s11098-005-2138-3

Cannon, W. B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. The American Journal of Psychology, 39(1/4), 106. https://doi.org/10.2307/1415404

Cant, R., Cooper, S., Sussex, R., & Bogossian, F. (2019). What's in a name? Clarifying the nomenclature of virtual simulation. *Clinical Simulation in Nursing*, *27*, 26-30. https://doi.org/10.1016/j.ecns.2018.11.003

Cantrell, M. A. (2008). The importance of debriefing in clinical simulations. Clinical Simulation in Nursing, 4(2), e19-e23. https://doi.org/10.1016/j.ecns.2008.06.006

Cantrell, M. A., Franklin, A., Leighton, K., & Carlson, A. (2017). The evidence in simulation-based learning experiences in nursing education and practice: An umbrella review. *Clinical Simulation in Nursing*, *13*(12), 634-667. https://doi.org/10.1016/j.ecns.2017.08.004

Carrizo Sainero, G. C. (2002). Las fuentes de información: presente y futuro. Algunos apuntes metodológicos. *Cuadernos de documentación multimedia*, *10*(19). https://doi.org/10.5209/rev_dcin.2011.v34.36446

CarrYanchus, N. J. (2006). *Development and validation of a self-report cognitive appraisal scale*. Doctoral dissertation, uga.

Carswell, C. M., Clarke, D., & Seales, W. B. (2005). Assessing mental workload during Laparoscopic surgery. Surgical Innovation, 12(1), 80-90. https://doi.org/10.1177/155335060501200112

Cen, L., Wu, F., Yu, Z., & Hu, F. (2016). A real-time speech emotion recognition system and its application in online learning. In Emotions, technology, design, and learning (pp. 27-46). Academic Press.

Chanel, G., Kierkels, J. J., Soleymani, M., & Pun, T. (2009). Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, *67*(8), 607-627. https://doi.org/10.1016/j.ijhcs.2009.03.005

Charland, L. C. (2008). The heat of emotion: Valence and the demarcation problem. Journal of consciousness studies, 12(8), 82-102. https://doi.org/10.1075/ceb.4.02cha

Cheung, K. L., Hors-Fraile, S., & De Vries, H. (2019). How to use the integrated-change model to design digital health programs. Digital Health, 143-157. https://doi.org/10.1016/b978-0-12-820077-3.00008-0

Chick, R., Adams, A., Peace, K., Kemp Bohan, P., Schwantes, I., Clifton, G., Vicente, D., Propper, B., Newhook, T., Grubbs, E., Bednarski, B., & Vreeland, T. (2021). Using the flipped classroom model in surgical education: Efficacy and trainee perception. Journal of Surgical Education, 78(6), 1803-1807. https://doi.org/10.1016/j.jsurg.2021.05.008

Chin, K. E., Kwon, D., Gan, Q., Ramalingam, P. X., Wistuba, I. I., Prieto, V. G., & Aung, P. P. (2020). undefined. Archives of Pathology & Laboratory Medicine, 145(1), 22-31. https://doi.org/10.5858/arpa.2020-0467-sa

Christensen, M. D., Rieger, K., Tan, S., Dieckmann, P., Østergaard, D., & Watterson, L. M. (2015). undefined. Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare, 10(6), 352-359. https://doi.org/10.1097/sih.0000000000000123

Cioffi, J. (2001). Clinical simulations: Development and validation. Nurse Education Today, 21(6), 477-486. https://doi.org/10.1054/nedt.2001.0584

Clapper, T. C. (2010). Beyond Knowles: What those conducting simulation need to know about adult learning theory. Clinical Simulation in Nursing, 6(1), e7-e14. https://doi.org/10.1016/j.ecns.2009.07.003

Cleveland-Innes, M., & Campbell, P. (2012). Emotional presence, learning, and the online learning environment. The International Review of Research in Open and Distributed Learning, 13(4), 269. https://doi.org/10.19173/irrodl.v13i4.1234

Clore, G. L., Ortony, A., & Foss, M. A. (1987). The psychological foundations of the affective lexicon. Journal of personality and social psychology, 53(4), 751.

Coffey, C. S., MacDonald, B. V., Shahrvini, B., Baxter, S. L., & Lander, L. (2020). Student perspectives on remote medical education in clinical core clerkships during the COVID-19 pandemic. Medical Science Educator, 30(4), 1577-1584. https://doi.org/10.1007/s40670-020-01114-9

Cohen, A. S., Minor, K. S., Baillie, L. E., & Dahir, A. M. (2008). Clarifying the linguistic signature: Measuring personality from natural speech. Journal of personality assessment, 90(6), 559-563.

Colligan, L., Potts, H. W., Finn, C. T., & Sinkin, R. A. (2015). undefined. International Journal of Medical Informatics, 84(7), 469-476. https://doi.org/10.1016/j.ijmedinf.2015.03.003

Colquhoun, H. L., Jesus, T. S., O'Brien, K. K., Tricco, A. C., Chui, A., Zarin, W., Lillie, E., Hitzig, S. L., & Straus, S. (2017). Study protocol for a scoping review on rehabilitation scoping reviews. Clinical Rehabilitation, 31(9), 1249-1256. https://doi.org/10.1177/0269215516688514

Conway, J. M., & Lance, C. E. (2010). What reviewers should expect from authors regarding common method bias in organizational research. *Journal of Business and Psychology*, *25*(3), 325-334. https://doi.org/10.1007/s10869-010-9181-6

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues in field settings*. Boston, MA: Houghton Mifflin.

Corter, J. E., Esche, S. K., Chassapis, C., Ma, J., & Nickerson, J. V. (2011). Process and learning outcomes from remotely-operated, simulated, and hands-on student

laboratories. Computers & Education, 57(3), 2054-2067. https://doi.org/10.1016/j.compedu.2011.04.009

Cosnita, R. M., Cimpean, A. M., & Raica, M. (2020). E-learning and E-assessement: Two big challenges of medical education management in Romania. Mednarodno inovativno poslovanje = Journal of Innovative Business and Management, 12(1), 61-71. https://doi.org/10.32015/jibm/2020-12-1-8

Craig, P., Cooper, C., Gunnell, D., Haw, S., Lawson, K., Macintyre, S., Ogilvie, D., Petticrew, M., Reeves, B., Sutton, M., & Thompson, S. (2012). Using natural experiments to evaluate population health interventions: New medical research council guidance. *Journal of Epidemiology and Community Health*, *66*(12), 1182-1186. https://doi.org/10.1136/jech-2011-200375

D'agostino, R. B., Belanger, A., & D'agostino, R. B. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, *44*(4), 316-321. https://doi.org/10.1080/00031305.1990.10475751

D'Agostino, R. B., Massaro, J. M., & Sullivan, L. M. (2003). Non-inferiority trials: Design concepts and issues - the encounters of academic consultants in statistics. *Statistics in Medicine*, *22*(2), 169-186. https://doi.org/10.1002/sim.1425

Damasio, A. R. (2001). Emotion and the human brain. Annals of the New York Academy of Sciences, 935(1), 101-106. https://doi.org/10.1111/j.1749-6632.2001.tb03475.x

Datta, R., Upadhyay, K., & Jaideep, C. (2012). Simulation and its role in medical education. Medical Journal Armed Forces India, 68(2), 167-172. https://doi.org/10.1016/s0377-1237(12)60040-9

De Vocht, F., Katikireddi, S. V., McQuire, C., Tilling, K., Hickman, M., & Craig, P. (2021). Conceptualising natural and quasi experiments in public health. *BMC Medical Research Methodology*, *21*(1). https://doi.org/10.1186/s12874-021-01224-x

Decker, S., Sportsman, S., Puetz, L., & Billings, L. (2008). The evolution of simulation and its contribution to competency. The Journal of Continuing Education in Nursing, 39(2), 74-80. https://doi.org/10.3928/00220124-20080201-06

Dean, B. (2016). Natural and quasi-natural experiments to evaluate cybersecurity policies. *Journal of International Affairs*, *70*(1), 139-160.

Dede, C. (2007). Transforming education for the 21st century: New pedagogies that help all students attain sophisticated learning outcomes. Commissioned by the NCSU Friday Institute, February. https://doi.org/10.5772/intechopen.96998

Definition of research. (n.d.). Dictionary by Merriam-Webster: America's most-trusted online dictionary. https://www.merriam-webster.com/dictionary/research

Dewaele, J., & Alfawzan, M. (2018). Does the effect of enjoyment outweigh that of anxiety in foreign language performance? Studies in Second Language Learning and Teaching, 8(1), 21-45. https://doi.org/10.14746/ssllt.2018.8.1.2

Dhawan, S. (2020). Online learning: A Panacea in the time of COVID-19 crisis. Journal of Educational Technology Systems, 49(1), 5-22. https://doi.org/10.1177/0047239520934018

Diaconis, P. (2006). Theories of data analysis: From magical thinking through classical statistics. *Exploring Data Tables, Trends, and Shapes*, 1-36. https://doi.org/10.1002/9781118150702.ch1

Diaz, M. C., & Walsh, B. M. (2021). Telesimulation-based education during COVID-19. *The Clinical Teacher*, *18*(2), 121-125. https://doi.org/10.1111/tct.13273

Dilmegani, C. (2021). What is Affective Computing / Emotion AI in 2021: In-depth Guide?

Dirkx, J. M. (2008). The meaning and role of emotions in adult learning. New Directions for Adult and Continuing Education, 2008(120), 7-18. https://doi.org/10.1002/ace.311

DistillerSR. (2022, February 23). *Evidence Partners*. https://www.evidencepartners.com

Dolcos, F., Katsumi, Y., Moore, M., Berggren, N., De Gelder, B., Derakshan, N., Hamm, A. O., Koster, E. H., Ladouceur, C. D., Okon-Singer, H., Pegna, A. J., Richter, T., Schweizer, S., Van den Stock, J., Ventura-Bort, C., Weymar, M., & Dolcos, S. (2020). Neural correlates of emotion-attention interactions: From perception, learning, and memory to social cognition, individual differences, and training interventions. Neuroscience & Biobehavioral Reviews, 108, 559-601. https://doi.org/10.1016/j.neubiorev.2019.08.017

Doughlas, C. (2008). The role of emotions in online travel: Towards the development of a Website Emotional Design (WED) model for romantic travel consumers. Doctoral dissertation, Purdue University.

Doyle, D. J. (2002). Simulation in medical education: Focus on anesthesiology. Medical Education Online, 7(1), 4544. https://doi.org/10.3402/meo.v7i.4544

Dreifuerst, K. T. (2009). Debriefing simulation using the DML strategy to foster meaningful learning. Clinical Simulation in Nursing, 5(3), e137. https://doi.org/10.1016/j.ecns.2009.04.029

Dreifuerst, K. T. (2009). Debriefing simulation using the DML strategy to foster meaningful learning. Clinical Simulation in Nursing, 5(3), e137. https://doi.org/10.1016/j.ecns.2009.04.029

Dubois, P., Rouland, J., Chaillou, C., Karpf, S., Meseure, P., & Duquenoy, F. (n.d.). A training simulator for retinal laser photocoagulation. *Proceedings of 17th International Conference of the Engineering in Medicine and Biology Society*. https://doi.org/10.1109/iembs.1995.575359

Duffy, M. C., Lajoie, S., & Lachapelle, K. (2016). Measuring emotions in medical education: Methodological and technological advances within authentic medical learning environments. Advances in Medical Education, 181-213. https://doi.org/10.1007/978-3-319-08275-2_10

Ekman, P. (1999). Basic emotions. Handbook of Cognition and Emotion, 45-60. https://doi.org/10.1002/0470013494.ch3

Ekman, P., & Friesen, W. V. (1978). Facial action coding system: Investigator's guide. Consulting Psychologists Press.

Elliot, L. R., Coovert, m. D., & Schiflett, S. G. (2017). A Psychological Fidelity Approach to Simulation-Based Training: Theory, Research and Principles. In Scaled Worlds: Development, Validation and Applications (pp. 85-109). Routledge.

Ericsson, K. A. (2015). Acquisition and maintenance of medical expertise: a perspective from the expert-performance approach with deliberate practice. Academic Medicine, 90(11), 1471-1486. https://doi.org/10.1097/acm.0000000000000939

Evans, D. J., Bay, B. H., Wilson, T. D., Smith, C. F., Lachman, N., & Pawlina, W. (2020). Going virtual to support anatomy education: A stopgap in the midst of the covid-19 pandemic. Anatomical Sciences Education, 13(3), 279-283. https://doi.org/10.1002/ase.1963

Farugia, P., Petrisor, B. A., Farrokhyar, F., & Bhandari, M. (2010). Research questions, hypotheses, and objectives. *Canadian journal of surgery*, *53*(4), 278.

Ferber, P. S. (2002). Adult learning theory and simulations-Designing simulations to eduate lawyers. Clinical L. Rev, 9, 417. https://doi.org/10.1201/9780203910276-10

Ferreira-González, I. (2014). Basis for the interpretation of Noninferiority studies: Considering the ROCKET–AF, RE-LY, and Aristotle studies. *Revista Española de Cardiología (English Edition)*, *67*(6), 432-435. https://doi.org/10.1016/j.rec.2013.10.018

Fichten, C., Havel, A., Wileman, S., Jorgensen, M., Arcuri, R., & Ruffolo, O. (2021). Digital tools faculty expected students to use during the COVID-19 pandemic in 2021: Problems and solutions for future hybrid and blended courses. Journal of Education and Training Studies, 9(8), 24. https://doi.org/10.11114/jets.v9i8.5310

Finan, E., Bismilla, Z., Whyte, H. E., LeBlanc, V., & McNamara, P. J. (2011). High-fidelity simulator technology may not be superior to traditional low-fidelity equipment for neonatal resuscitation training. Journal of Perinatology, 32(4), 287-292. https://doi.org/10.1038/jp.2011.96

Flanagan, B., Nestel, D., & Joseph, M. (2004). Making patient safety the focus: Crisis resource management in the undergraduate curriculum. Medical Education, 38(1), 56-66. https://doi.org/10.1111/j.1365-2923.2004.01701.x

Foronda, C. L., Fernandez-Burgos, M., Nadeau, C., Kelley, C. N., & Henry, M. N. (2020). Virtual simulation in nursing education: A systematic review spanning 1996 to 2018. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, *15*(1), 46-54. https://doi.org/10.1097/sih.0000000000000411

Fraser, K., & McLaughlin, K. (2019). Temporal pattern of emotions and cognitive load during simulation training and debriefing. Medical Teacher, 41(2), 184-189. https://doi.org/10.1080/0142159x.2018.1459531

Fraser, K., Ma, I., Teteris, E., Baxter, H., Wright, B., & McLaughlin, K. (2012). Emotion, cognitive load and learning outcomes during simulation training. *Medical Education*, *46*(11), 1055-1062. https://doi.org/10.1111/j.1365-2923.2012.04355.x

Fray, L., & Gore, J. (2018). Why people choose teaching: A scoping review of empirical studies, 2007–2016. Teaching and Teacher Education, 75, 153-163. https://doi.org/10.1016/j.tate.2018.06.009

Frijda, N. H. (2009). Emotion experience and its varieties. Emotion Review, 1(3), 264-271. https://doi.org/10.1177/1754073909103595

Frijda, N. H., & Mesquita, B. (2000). Beliefs through emotions. Emotions and Beliefs, 45-77. https://doi.org/10.1017/cbo9780511659904.003

Frijda, N. H., Kuipers, P., & Ter Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. Journal of Personality and Social Psychology, 57(2), 212-228. https://doi.org/10.1037/0022-3514.57.2.212

Gardner, R. (2013). Introduction to debriefing. In Seminars in perinatology (pp. 166-174). WB Saunders.

George, D., & Mallery, P. (2018). Descriptive statistics. In *IBM SPSS Statistics 25 Step by Step* (pp. 126-134). Routledge.

Glänzel, W. (1996). A bibliometric approach to social sciences. National research performances in 6 selected social science areas, 1990–1992. *Scientometrics*, *35*(3), 291-307. https://doi.org/10.1007/bf02016902

Goetz, T., Zirngibl, A., Pekrun, R., & Hall, N. (2003). Emotions, learning and achievement from an educational-psychological perspective (pp. 9-28).

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. Health Information & Libraries Journal, 26(2), 91-108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Green, J., Wyllie, A., & Jackson, D. (2014). Virtual worlds: A new frontier for nurse education? Collegian, 21(2), 135-141. https://doi.org/10.1016/j.colegn.2013.11.004

Gross, P. L., & Gross, E. M. (1927). College libraries and chemical education. *Science*, *66*(1713), 385-389. https://doi.org/10.1126/science.66.1713.385

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1), 1-21. https://doi.org/10.1080/00401706.1969.10490657

Guedes, V. L., & Borschiver, S. (2018). Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. *Encontro Nacional de Ciência da Informação*, *6*(1), 18. https://doi.org/10.18225/9788570131485.cap14

Gutierrez-Barreto, S. E., Argueta-Muñoz, F. D., Ramirez-Arias, J. D., Scherer-Castanedo, E., Hernández-Gutiérrez, L. S., & Olvera-Cortés, H. E. (2021). Implementation barriers in Telesimulation as an educational strategy: An interpretative description. Cureus. https://doi.org/10.7759/cureus.17852

Harris, A. D., McGregor, J. C., Perencevich, E. N., Furuno, J. P., Zhu, J., Peterson, D. E., & Finkelstein, J. (2006). The use and interpretation of quasi-experimental studies in medical informatics. *Journal of the American Medical Informatics Association*, *13*(1), 16-23. https://doi.org/10.1197/jamia.m1749

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task load index): Results of empirical and theoretical research. *Advances in Psychology*, 139-183. https://doi.org/10.1016/s0166-4115(08)62386-9

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904-908. https://doi.org/10.1177/154193120605000909

HARVARD. (n.d.). *Designing a simulation session*. Instructional Moves. https://instructionalmoves.gse.harvard.edu/designing-simulation-session

Heffernan, R., Brumpton, K., Randles, D., & Pinidiyapathirage, J. (2021). Acceptability, technological feasibility and educational value of remotely facilitated simulation based training: A scoping review. Medical Education Online, 26(1). https://doi.org/10.1080/10872981.2021.1972506

Hecimovich, M., & Volet, S. (2011). Development of professional confidence in health education: Research evidence of the impact of guided practice into the profession. Health Education.

Helle, L., & Säljö, R. (2012). Collaborating with digital tools and peers in medical education: Cases and simulations as interventions in learning. Instructional Science, 40(5), 737-744. https://doi.org/10.1007/s11251-012-9216-7

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklade, A. L., & Christ, R. E. (1992). Comparison of four subjective workload rating scales. Human Factors: The Journal of the Human Factors and Ergonomics Society, 34(4), 429-439. https://doi.org/10.1177/001872089203400405

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85-126. https://doi.org/10.1023/b:aire.0000045502.10941.a9

Hohmann, G. W. (1966). Some effects of spinal cord lesions on experienced emotional feelings. Psychophysiology, 3(2), 143-156. https://doi.org/10.1111/j.1469-8986.1966.tb02690.x

Honey, M. A., & Hilton, M. L. (2011). Learning science through computer games. National Academies Press, Washington, DC.

Hoonakker, P., Carayon, P., Gurses, A. P., Brown, R., Khunlertkit, A., McGuire, K., & Walker, J. M. (2011). Measuring workload of ICU nurses with a questionnaire survey: The NASA task load index (TLX). IIE Transactions on Healthcare Systems Engineering, 1(2), 131-143. https://doi.org/10.1080/19488300.2011.609524

Hooper, S., & Rieber, L. P. (1995). Teaching with technology. Teaching: Theory into practice, 2013, 154-170.

Horbar, J. D. (1985). A computer simulation of medical decision stategy performance. *Computers and Biomedical Research*, *18*(6), 563-575. https://doi.org/10.1016/0010-4809(85)90031-x

Hosany, S. (2011). Appraisal determinants of tourist emotional responses. Journal of Travel Research, 51(3), 303-314. https://doi.org/10.1177/0047287511410320

Hosany, S. (2011). Appraisal determinants of tourist emotional responses. *Journal of Travel Research*, *51*(3), 303-314. https://doi.org/10.1177/0047287511410320

Holmstrom, A. J. (2015). Emotional appraisal/reappraisal in social support. The International Encyclopedia of Interpersonal Communication, 1-10.

Houlden, S., & Veletsianos, G. (2020). Coronavirus pushes universities to switch to online classes–but are they ready. The Conversation, 12.

Houlden, S., & Veletsianos, G. (2020). The problem with flexible learning: Neoliberalism, freedom, and learner subjectivities. Learning, Media and Technology, 46(2), 144-155. https://doi.org/10.1080/17439884.2020.1833920

Hsu, T. (2017). Effects of gender and different augmented reality learning systems on English vocabulary learning of elementary school students. Universal Access in the Information Society, 18(2), 315-325. https://doi.org/10.1007/s10209-017-0593-1

Hu, M., Wattchow, D., & De Fontgalland, D. (2018). From ancient to avant-garde: A review of traditional and modern multimodal approaches to surgical anatomy education. ANZ Journal of Surgery, 88(3), 146-151. https://doi.org/10.1111/ans.14189

Hudlicka, E. (2017). Computational modeling of cognition–emotion interactions: Theoretical and practical relevance for behavioral healthcare. In Emotions and affect in human factors and human-computer interaction (pp. 383-436). Academia Press.

Hung, M., Chang, I., & Hwang, H. (2011). Exploring academic teachers' continuance toward the web-based learning system: The role of causal attributions. Computers & Education, 57(2), 1530-1543. https://doi.org/10.1016/j.compedu.2011.02.001

Ikeyama, T., Shimizu, N., & Ohta, K. (2012). Low-cost and ready-to-Go remote-facilitated simulation-based learning. Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare, 7(1), 35-39. https://doi.org/10.1097/sih.0b013e31822eacae

Isman, A., & Dabaj, F. (2014). Communication barriers in distance education. In Society for Information Technology & Teacher Education International Conference (pp. 491-496). Association for the Advancement of Computing in Education (AACE).

Israel, L. (2020). The cognitive emotion process. Doctoral dissertation, lmu.

Issenberg, B., Gordon, M. S., Gordon, D. L., Safford, R. E., & Hart, I. R. (2001). undefined. Medical Teacher, 23(1), 16-23. https://doi.org/10.1080/01421590020007324

James Hung, H. M., Wang, S., Tsong, Y., Lawrence, J., & O'Neil, R. T. (2003). Some fundamental issues with non-inferiority testing in active-controlled trials. *Statistics in Medicine*, *22*(2), 213-225. https://doi.org/10.1002/sim.1315

Jeffery, K. A., & Bauer, C. F. (2020). Students' responses to emergency remote online teaching reveal critical factors for all teaching. Journal of Chemical Education, 97(9), 2472-2485. https://doi.org/10.1021/acs.jchemed.0c00736

Jeffries, P. (2005). Guest editor: Developing evidenced-based teaching strategies and practices when using simulations. Clinical Simulation in Nursing, 2(1), e1-e2. https://doi.org/10.1016/j.ecns.2009.05.014

Jerritta, S., Murugappan, M., Nagarajan, R., & Wan, K. (2011). Physiological signals based human emotion recognition: a review. In 2011 IEEE 7th international colloquium on signal processing and its applications (pp. 410-415). IEEE.

Jhangiani, R. S., Chiang, I. A., & Price, P. C. (2015). *Research methods in psychology*-2nd Canadian Edition. BC Campus.

Jhangiani, R. S., Chiang, I. C., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology*. Kwantlen Polytechnic University.

Jones, M. V. (2003). Controlling emotions in sport. The Sport Psychologist, 17(4), 471-486. https://doi.org/10.1123/tsp.17.4.471

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., & Mack, S. (Eds.). (2000). Principles of neural science (Vol. 4, pp. 1227-1246). New York: McGraw-hill.

Kang, H. (2004). *The prevention and handling of the missing data*. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402. https://doi.org/10.4097/kjae.2013.64.5.402

Karimi, L., & Meyer, D. (2019). An evaluation of common method variance-bias in psychology. *International Journal of Psychological Studies*, *11*(3), 83. https://doi.org/10.5539/ijps.v11n3p83

Kaur, N., Dwivedi, D., Arora, J., & Gandhi, A. (2020). undefined. National Journal of Physiology, Pharmacy and Pharmacology, 10(7), 1. https://doi.org/10.5455/njppp.2020.10.04096202028042020

Keskitalo, T., & Ruokamo, H. (2021). Exploring learners' emotions and emotional profiles in simulation-based medical education. Australasian Journal of Educational Technology, 15-26. https://doi.org/10.14742/ajet.5761

Khalaila, R. (2014). Simulation in nursing education: An evaluation of students' outcomes at their first clinical practice combined with simulations. Nurse Education Today, 34(2), 252-258. https://doi.org/10.1016/j.nedt.2013.08.015

Kim, H. (2013). Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, *38*(1), 52. https://doi.org/10.5395/rde.2013.38.1.52

Klasen, M., & Sopka, S. (2021). Demonstrating equivalence and non-inferiority of medical education concepts. *Medical Education*, *55*(4), 455-461. https://doi.org/10.1371/journal.pone.0254923

Kneebone, R. (2003). Simulation in surgical training: Educational issues and practical implications. Medical Education, 37(3), 267-277. https://doi.org/10.1046/j.1365-2923.2003.01440.x

Koedinger, K. R., McLaughlin, E. A., & Heffernan, N. T. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research*, *43*(4), 489-510. https://doi.org/10.2190/ec.43.4.d

Kononowicz, A. A., Woodham, L. A., Edelbring, S., Stathakarou, N., Davies, D., Saxena, N., Tudor Car, L., Carlstedt-Duke, J., Car, J., & Zary, N. (2019). Virtual patient simulations in health professions education: Systematic review and meta-analysis by the digital health education collaboration. *Journal of Medical Internet Research*, *21*(7), e14676. https://doi.org/10.2196/14676

Koster, R. W., Beenen, L. F., Van der Boom, E. B., Spijkerboer, A. M., Tepaske, R., Van der Wal, A. C., Beesems, S. G., & Tijssen, J. G. (2017). Safety of mechanical chest compression devices AutoPulse and LUCAS in cardiac arrest: A randomized clinical trial for non-inferiority. *European Heart Journal*, *38*(40), 3006-3013. https://doi.org/10.1093/eurheartj/ehx318

Kothari, C. R. (2004). *Research methodology: Methods and techniques*. New Age International.

Krogh, L. Q., Bjørnshave, K., Vestergaard, L. D., Sharma, M. B., Rasmussen, S. E., Nielsen, H. V., Thim, T., & Løfgren, B. (2015). E-learning in pediatric basic life support: A randomized controlled non-inferiority study. *Resuscitation*, *90*, 7-12. https://doi.org/10.1016/j.resuscitation.2015.01.030

Kuiper, R. A., & Pesut, D. J. (2004). Promoting cognitive and metacognitive reflective reasoning skills in nursing practice: Self-regulated learning theory. Journal of Advanced Nursing, 45(4), 381-391. https://doi.org/10.1046/j.1365-2648.2003.02921.x

Kurup, V., Matei, V., & Ray, J. (2017). Role of in-situ simulation for training in healthcare. Current Opinion in Anaesthesiology, 30(6), 755-760. https://doi.org/10.1097/aco.0000000000000514

Lamé, G., & Dixon-Woods, M. (2020). Using clinical simulation to study how to improve quality and safety in healthcare. BMJ Simulation and Technology Enhanced Learning, 6(2), 87-94. https://doi.org/10.1136/bmjstel-2018-000370

Lang, P. J., & Davis, M. (2006). Emotion, motivation, and the brain: Reflex foundations in animal and human research. Understanding Emotions, 3-29. https://doi.org/10.1016/s0079-6123(06)56001-7

Lasater, K. (2007). High-fidelity simulation and the development of clinical judgment: Students' experiences. Journal of Nursing Education, 46(6), 269-276. https://doi.org/10.3928/01484834-20070601-06

Lateef, F. (2010). Simulation-based learning: Just like the real thing. Journal of Emergencies, Trauma, and Shock, 3(4), 348. https://doi.org/10.4103/0974-2700.70743

Leatherdale, S. T. (2018). Natural experiment methodology for research: A review of how different methods can support real-world research. *International Journal of Social Research Methodology*, *22*(1), 19-35. https://doi.org/10.1080/13645579.2018.1488449

Lewin, K. (1951). Field theory in social science: selected theoretical papers (Edited by Dorwin Cartwright.).

Lewis, S. E., & Lewis, J. E. (2005). The same or not the same: Equivalence as an issue in educational research. *Journal of Chemical Education*, *82*(9), 1408. https://doi.org/10.1021/ed082p1408

Lewis, R., Strachan, A., & Smith, M. M. (2012). Is high fidelity simulation the most effective method for the development of non-technical skills in nursing? A review of the current evidence. The open nursing journal, 6, 82.

Liang, T., & Liu, Y. (2018). Research landscape of business intelligence and big data analytics: A Bibliometrics study. *Expert Systems with Applications*, *111*, 2-10. https://doi.org/10.1016/j.eswa.2018.05.018

Limentani, G. B., Ringo, M. C., Ye, F., Bergquist, M. L., & MCSorley, E. O. (2005). Beyond the T-TesT: Statistical equivalence testing. *Analytical Chemistry*, *77*(11), 221 A-226 A. https://doi.org/10.1021/ac053390m

Lin, K., Travlos, D. V., Wadelin, J. W., & Vlasses, P. H. (2011). Simulation and introductory pharmacy practice experiences. American journal of pharmaceutical education, 75(10).

Linden, L. L., QUARTERMAN, J. S., & WHINSTON, A. B. (2013). Improving Internet Security Through Social Information and Social Comparison: A Field Quasi-Experiment. *Workshop on the Economics of Information Security 12th WEIS*, *11*(12), 1-43. https://ink.library.smu.edu.sg/sis_research/2566

Liu, D., Santhanam, R., & Webster, J. (2017). Toward meaningful engagement: A framework for design and research of Gamified information systems. MIS Quarterly, 41(4), 1011-1034. https://doi.org/10.25300/misq/2017/41.4.01

Lotka, A. J. (1929). Biometric functions in a population growing in accordance with a prescribed law. *Proceedings of the National Academy of Sciences*, *15*(10), 793-798. https://doi.org/10.1073/pnas.15.10.793

MacInnis, D. J., & De Mello, G. E. (2005). The concept of hope and its relevance to product evaluation and choice. Journal of Marketing, 69(1), 1-14. https://doi.org/10.1509/jmkg.69.1.1.55513

MacKenzie, S. B., & Podsakoff, P. M. (2012). Common method bias in marketing: Causes, mechanisms, and procedural remedies. *Journal of Retailing*, *88*(4), 542-555. https://doi.org/10.1016/j.jretai.2012.08.001

Mangold, K., & Adler, M. (2019). Designing quantitative research studies. In *Healthcare simulation research* (pp. 169-174). Springer, Cham.

Maran, N. J., & Glavin, R. J. (2003). Low-to high-fidelity simulation–a continuum of medical education?. Medical education, 37, 22-28.

Mazer, J. P., McKenna-Buchanan, T. P., Quinlan, M. M., & Titsworth, S. (2014). The dark side of emotion in the classroom: Emotional processes as mediators of teacher communication behaviors and student negative emotions. Communication Education, 63(3), 149-168. https://doi.org/10.1080/03634523.2014.904047

McConnell, M. M., & Eva, K. W. (2012). The role of emotion in the learning and transfer of clinical skills and knowledge. Academic Medicine, 87(10), 1316-1322. https://doi.org/10.1097/acm.0b013e3182675af2

McCoy, C. E., Sayegh, J., Alrabah, R., & Yarris, L. M. (2017). Telesimulation: An innovative tool for health professions education. AEM Education and Training, 1(2), 132-136. https://doi.org/10.1002/aet2.10015

McCoy, S., Everard, A., Polak, P., & Galletta, D. F. (2008). An experimental study of antecedents and consequences of online ad intrusiveness. International Journal of Human-Computer Interaction, 24(7), 672-699. https://doi.org/10.1080/10447310802335664

McCoy, C. E., Sayegh, J., Alrabah, R., & Yarris, L. M. (2017). Telesimulation: An innovative tool for health professions education. *AEM Education and Training*, *1*(2), 132-136. https://doi.org/10.1002/aet2.10015

McGaghie, W. C., Issenberg, S. B., Cohen, E. R., Barsuk, J. H., & Wayne, D. B. (2011). Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Academic Medicine, 86(6), 706-711. https://doi.org/10.1097/acm.0b013e318217e119

Mehrabian, A., & Bernath, M. S. (1991). Factorial composition of commonly used self-report depression inventories: Relationships with basic dimensions of temperament. Journal of Research in Personality, 25(3), 262-275. https://doi.org/10.1016/0092-6566(91)90019-m

Mehrabian, A., & O'Reilly, E. (1980). Analysis of personality measures in terms of basic dimensions of temperament. Journal of Personality and Social Psychology, 38(3), 492-503. https://doi.org/10.1037/0022-3514.38.3.492

Meltzer, J. A., Baird, A. J., Steele, R. D., & Harvey, S. J. (2018). Computer-based treatment of poststroke language disorders: A non-inferiority study of telerehabilitation compared to in-person service delivery. *Aphasiology*, *32*(3), 290-311. https://doi.org/10.1080/02687038.2017.1355440

Miloff, A., Lindner, P., Dafgård, P., Deak, S., Garke, M., Hamilton, W., Heinsoo, J., Kristoffersson, G., Rafi, J., Sindemark, K., Sjölund, J., Zenger, M., Reuterskiöld, L., Andersson, G., & Carlbring, P. (2019). Automated virtual reality exposure therapy for spider phobia vs. in-vivo one-session treatment: A randomized non-inferiority trial. *Behaviour Research and Therapy*, *118*, 130-140. https://doi.org/10.1016/j.brat.2019.04.004

Mukhtar, K., Javed, K., Arooj, M., & Sethi, A. (2020). Advantages, limitations and recommendations for online learning during COVID-19 pandemic era. Pakistan Journal of Medical Sciences, 36(COVID19-S4). https://doi.org/10.12669/pjms.36.covid19-s4.2785

Munshi, F., Lababidi, H., & Alyousef, S. (2015). undefined. Journal of Taibah University Medical Sciences, 10(1), 12-15. https://doi.org/10.1016/j.jtumed.2015.01.008

Murray, B. A. (2014). The use of high-fidelity simulation in psychiatric and mental health nursing clinical education. International Journal of Health Sciences Education, 2(1), 3. https://doi.org/10.1097/00024776-201403000-00012

Murray, E. G., Neumann, D. L., Moffitt, R. L., & Thomas, P. R. (2016). The effects of the presence of others during a rowing exercise in a virtual reality environment. *Psychology of Sport and Exercise*, *22*, 328-336. https://doi.org/10.1016/j.psychsport.2015.09.007

Nel, C. P., Van Zyl, G. J., & Labuschagne, M. J. (2021). Simulation in plastic surgery: Features and uses that lead to effective learning. *African Journal of Health Professions Education*, *13*(1), 29. https://doi.org/10.7196/ajhpe.2021.v13i1.1181

Nestel, D., & Tierney, T. (2007). Role-play for medical students learning about communication: Guidelines for maximising benefits. BMC Medical Education, 7(1). https://doi.org/10.1186/1472-6920-7-3

Niedermeier, M., Einwanger, J., Hartl, A., & Kopp, M. (2017). Affective responses in mountain hiking—A randomized crossover trial focusing on differences between indoor and outdoor activity. *PLOS ONE*, *12*(5), e0177719. https://doi.org/10.1371/journal.pone.0177719

Nomura, O., Onishi, H., & Kato, H. (2017). Medical students can teach communication skills – a mixed methods study of cross-year peer tutoring. *BMC Medical Education*, *17*(1). https://doi.org/10.1186/s12909-017-0939-7

Norman, G., Dore, K., & Grierson, L. (2012). The minimal relationship between simulation fidelity and transfer of learning. Medical Education, 46(7), 636-647. https://doi.org/10.1111/j.1365-2923.2012.04243.x

Norman, S. B. (2003). The effect of training managers to gain employee trust on employee work-related attitudes and emotional well-being.

Noteborn, G., Bohle Carbonell, K., Dailey-Hebert, A., & Gijselaers, W. (2012). The role of emotions and task significance in virtual education. The Internet and Higher Education, 15(3), 176-183. https://doi.org/10.1016/j.iheduc.2012.03.002

Oh, S., Mishler, O., Yang, J. S., & Barnes, C. (2021). Effectiveness of remote simulation-based learning for periodontal instrumentation: A non-inferiority study. *Journal of Dental Education*. https://doi.org/10.1002/jdd.12820

Okuda, Y., Bryson, E. O., DeMaria, S., Jacobson, L., Quinones, J., Shen, B., & Levine, A. I. (2009). The utility of simulation in medical education: What is the evidence? Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine, 76(4), 330-343. https://doi.org/10.1002/msj.20127

Ontrup, G., Vogel, M., Wolf, O. T., Zahn, P. K., Kluge, A., & Hagemann, V. (2020). Does simulation-based training in medical education need additional stressors? An experimental study. *Ergonomics*, *63*(1), 80-90. https://doi.org/10.1080/00140139.2019.1677948

Ossom Williamson, P., & Minter, C. I. (2019). Exploring PubMed as a reliable resource for scholarly communications services. *Journal of the Medical Library Association*, *107*(1). https://doi.org/10.5195/jmla.2019.433

Oviatt, S. (2006). Human-centered design meets cognitive load theory. Proceedings of the 14th annual ACM international conference on Multimedia - MULTIMEDIA '06. https://doi.org/10.1145/1180639.1180831

Padilha, J. M., Machado, P. P., Ribeiro, A., Ramos, J., & Costa, P. (2019). Clinical virtual simulation in nursing education: Randomized controlled trial. *Journal of Medical Internet Research*, *21*(3), e11529. https://doi.org/10.2196/11529

Pai, M., McCulloch, M., Gorman, J. D., Pai, N., Enanoria, W., Kennedy, G., Tharyan, P., & Colford Jr, J. M. (2004). Systematic reviews and meta-analyses: an illustrated, step-by-step guide. *The National Medical Journal of India*, *17*(2), 86-95. https://pubmed.ncbi.nlm.nih.gov/15141602/

Papanagnou, D. (2017). Telesimulation: A paradigm shift for simulation education. AEM Education and Training, 1(2), 137-139. https://doi.org/10.1002/aet2.10032

Papanagnou, D. (2017). Telesimulation: A paradigm shift for simulation education. *AEM Education and Training*, *1*(2), 137-139. https://doi.org/10.1002/aet2.10032

Paskins, Z., & Peile, E. (2010). Final year medical students' views on simulation-based teaching: A comparison with the best evidence medical education systematic review. Medical Teacher, 32(7), 569-577. https://doi.org/10.3109/01421590903544710

Pather, N., Blyth, P., Chapman, J. A., Dayal, M. R., Flack, N. A., Fogg, Q. A., Green, R. A., Hulme, A. K., Johnson, I. P., Meyer, A. J., Morley, J. W., Shortland, P. J., Štrkalj, G., Štrkalj, M., Valter, K., Webb, A. L., Woodley, S. J., & Lazarus, M. D. (2020). undefined. Anatomical Sciences Education, 13(3), 284-300. https://doi.org/10.1002/ase.1968

Pekrun, R., & Linnenbrink-Garcia, L. (2014). Introduction to emotions in education. Routledge.

Pekrun, R., & Perry, R. P. (2014). Control-value theory of achievement emotions. In International handbook of emotions in education (pp. 130-159). Routledge.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. Educational Psychologist, 37(2), 91-105. https://doi.org/10.1207/s15326985ep3702_4

Peters, M., Godfrey, C., McInerney, P., Munn, Z., Trico, A., & Khalil, H. (2020). Chapter 11: Scoping reviews. JBI Manual for Evidence Synthesis. https://doi.org/10.46658/jbimes-20-12

Peters, M. D., Godfrey, C., McInerney, P., Munn, Z., Tricco, A. C., & Khalil, H. (2020). Chapter 11: Scoping Reviews (2020 version). In *Aromataris E, Munn Z (Editors). JBI Manual for Evidence Synthesis*. JBI.

Pham, M. T., Rajić, A., Greig, J. D., Sargeant, J. M., Papadopoulos, A., & McEwen, S. A. (2014). A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. Research Synthesis Methods, 5(4), 371-385. https://doi.org/10.1002/jrsm.1123

Picard, R. W. (1999). Affective Computing for HCI. In HCI (pp. 829-833).

Plass, J. L., & Kaplan, U. (2016). Emotional design in digital media for learning. In Emotions, technology, design, and learning (pp. 131-161). Academia Press.

Podsakoff, P. M., MacKenzie, S. B., Lee, J., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879-903. https://doi.org/10.1037/0021-9010.88.5.879

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98-125. https://doi.org/10.1016/j.inffus.2017.02.003

Porter, A. L., Kongthon, A., & Lu, J. (. (2002). Research profiling: Improving the literature review. *Scientometrics*, *53*(3), 351-370. https://doi.org/10.1023/a:1014873029258

Postareff, L., Mattsson, M., Lindblom-Ylänne, S., & Hailikari, T. (2017). The complex relationship between emotions, approaches to learning, study success and study progress during the transition to university. Higher Education, 73(3), 441-457. https://doi.org/10.1007/s10734-016-0096-7

Pryor, R. R., Colburn, D., Crill, M. T., Hostler, D. P., & Suyama, J. (2012). Fitness characteristics of a suburban special weapons and tactics team. Journal of Strength and Conditioning Research, 26(3), 752-757. https://doi.org/10.1519/jsc.0b013e318225f177

Rafaeli, A., & Sutton, R. I. (1989). The expression of emotion in organizational life. Research in organizational behavior, 11(1), 1-42.

Rall, M., Manser, T., & Howard, S. K. (2000). Key elements of debriefing for simulator training. European Journal of Anaesthesiology, 17(8), 516-517. https://doi.org/10.1097/00003643-200008000-00011

Rashid, P., & Gianduzzo, T. R. (2015). Urology technical and non-technical skills development: The emerging role of simulation. BJU International, 117, 9-16. https://doi.org/10.1111/bju.13259

Ray, J. M., Wong, A. H., Yang, T. J., Buck, S., Joseph, M., Bonz, J. W., Auerbach, M. A., Couturier, K., Tomassoni, A. J., Schwartz, M. L., & Evans, L. V. (2021). Virtual Telesimulation for medical students during the COVID-19 pandemic. *Academic Medicine*, *96*(10), 1431-1435. https://doi.org/10.1097/acm.0000000000004129

Reisenzein, R., Horstmann, G., & Schützwohl, A. (2019). The cognitive-evolutionary model of surprise: A review of the evidence. Topics in Cognitive Science, 11(1), 50-74. https://doi.org/10.1111/tops.12292

Remington, N. A., Fabrigar, L. R., & Visser, P. S. (2000). Reexamining the circumplex model of affect. Journal of Personality and Social Psychology, 79(2), 286-300. https://doi.org/10.1037/0022-3514.79.2.286

Renouf, T., & Pollard, M. (2017). Emergency medicine in remote regions. Cureus, 9(6). https://doi.org/10.7759/cureus.774

Rickles, N. M., Tieu, P., Myers, L., Galal, S., & Chung, V. (2009). The impact of a standardized patient program on student learning of communication skills. American Journal of Pharmaceutical Education, 73(1), 4. https://doi.org/10.5688/aj730104

Richins, M. (1997). The consumption emotion set. Journal of Consumer Research, 24, 127-46.

Roach, E., & Okrainec, A. (2021). Telesimulation for remote simulation and assessment. Journal of Surgical Oncology, 124(2), 193-199. https://doi.org/10.1002/jso.26505

Roach, E., & Okrainec, A. (2021). Telesimulation for remote simulation and assessment. *Journal of Surgical Oncology*, *124*(2), 193-199. https://doi.org/10.1002/jso.26505

Robinson, B. K., & Dearmon, V. (2013). Evidence-based nursing education: Effective use of instructional design and simulated learning environments to enhance knowledge transfer in undergraduate nursing students. Journal of Professional Nursing, 29(4), 203-209. https://doi.org/10.1016/j.profnurs.2012.04.022

Romme, G. L. (2002). Microworlds for management education and learning. Tilburg, Netherlands: Tilburg University.

Roseman, I. J. (1984). Cognitive determinants of emotion: A structural theory. Review of personality & social psychology, 59(5), 899-915. https://doi.org/10.1037/0022-3514.59.5.899

Roseman, I. J. (1996). Appraisal theory. Appraisal processes in emotion: Theory, methods, research, 3(19), 241-278. https://doi.org/10.1080/026999396380240

Roseman, I. J., Spindel, M. S., & Jose, P. E. (1990). Appraisals of emotion-eliciting events: Testing a theory of discrete emotions. Journal of Personality and Social Psychology, 59(5), 899-915. https://doi.org/10.1037/0022-3514.59.5.899

Rubin, D. C., & Talarico, J. M. (2009). A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. Memory, 17(8), 802-808. https://doi.org/10.1080/09658210903130764

Rubio, S., Diaz, E., Martin, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. Applied Psychology, 53(1), 61-86. https://doi.org/10.1111/j.1464-0597.2004.00161.x

Rudd, M. D. (2012). Brief cognitive behavioral therapy (BCBT) for suicidality in military populations. Military Psychology, 24(6), 592-603. https://doi.org/10.1080/08995605.2012.736325

Rudestam, K. E., & Newton, R. R. (2007). The method chapter: Describing your research plan. *Surviving your dissertation: A comprehensive guide to content and process*, *8*(2), 87-117. https://doi.org/10.2307/3172836

Rudolph, J. W., Simon, R., & Raemer, D. B. (2007). Which reality matters? Questions on the path to high engagement in healthcare simulation. Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare, 2(3), 161-163. https://doi.org/10.1097/sih.0b013e31813d1035

Rudolph, J. W., Simon, R., Raemer, D. B., & Eppich, W. J. (2008). Debriefing as formative assessment: Closing performance gaps in medical education. Academic Emergency Medicine, 15(11), 1010-1016. https://doi.org/10.1111/j.1553-2712.2008.00248.x

Rumrill, P. D., Fitzgerald, S. M., & Merchant, W. R. (2010). Using scoping literature reviews as a means of understanding and interpreting existing literature. Work, 35(3), 399-404. https://doi.org/10.3233/wor-2010-0998

Russell, J. A. (2003). Core affect and the psychological construction of emotion. Psychological Review, 110(1), 145-172. https://doi.org/10.1037/0033-295x.110.1.145

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. Journal of Research in Personality, 11(3), 273-294. https://doi.org/10.1016/0092-6566(77)90037-x

Ruth, J. A., Brunel, F. F., & Otnes, C. C. (2002). Linking thoughts to feelings: Investigating cognitive appraisals and consumption emotions in a mixed-emotions context. Journal of the Academy of Marketing Science, 30(1), 44-58. https://doi.org/10.1177/03079459994317

Ruth, J. A., Brunel, F. F., & Otnes, C. C. (2002). Linking thoughts to feelings: Investigating cognitive appraisals and consumption emotions in a mixed-emotions context. *Journal of the Academy of Marketing Science*, *30*(1), 44-58. https://doi.org/10.1177/03079459994317

Rutten, N., Van Joolingen, W. R., & Van der Veen, J. T. (2012). The learning effects of computer simulations in science education. Computers & Education, 58(1), 136-153. https://doi.org/10.1016/j.compedu.2011.07.017

Said, S., Gozdzik, M., Roche, T. R., Braun, J., Rössler, J., Kaserer, A., Spahn, D. R., Nöthiger, C. B., & Tscholl, D. W. (2020). Validation of the raw national aeronautics and space administration task load index (NASA-TLX) questionnaire to assess perceived workload in patient monitoring tasks: Pooled analysis study using mixed models. *Journal of Medical Internet Research*, *22*(9), e19472. https://doi.org/10.2196/19472

Sander, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. Neural Networks, 18(4), 317-352. https://doi.org/10.1016/j.neunet.2005.03.001

Sanseau, E., Lavoie, M., Tay, K., Good, G., Tsao, S., Burns, R., Thomas, A., Heckle, T., Wilson, M., Kou, M., & Auerbach, M. (2021). TeleSimBox: A perceived effective alternative for experiential learning for medical student education with

social distancing requirements. AEM Education and Training, 5(2). https://doi.org/10.1002/aet2.10590

Sawyer, T., Eppich, W., Brett-Fleegler, M., Grant, V., & Cheng, A. (2016). More than one way to debrief. Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare, 11(3), 209-217. https://doi.org/10.1097/sih.0000000000000148

Scherer, K. R. (1988). Criteria for emotion-antecedent appraisal: A review. Cognitive Perspectives on Emotion and Motivation, 89-126. https://doi.org/10.1007/978-94-009-2792-6_4

Scherer, Y. K., Bruce, S. A., & Runkawatt, V. (2007). A comparison of clinical simulation and case study presentation on nurse practitioner students' knowledge and confidence in managing a cardiac event. International Journal of Nursing Education Scholarship, 4(1). https://doi.org/10.2202/1548-923x.1502

Schiavenato, M. (2009). Reevaluating simulation in nursing education: Beyond the human patient simulator. Journal of Nursing Education, 48(7), 388-394. https://doi.org/10.3928/01484834-20090615-06

Schlairet, M. C., Schlairet, T. J., Sauls, D. H., & Bellflowers, L. (2015). Cognitive load, emotion, and performance in high-fidelity simulation among beginning nursing students: A pilot study. Journal of Nursing Education, 54(3). https://doi.org/10.3928/01484834-20150218-10

Schnall, S. (2010). Affect, mood and emotions. International Encyclopedia of Education, 544-548. https://doi.org/10.1016/b978-0-08-044894-7.00597-2

Schön, D. A. (1984). The reflective practitioner: How professionals think in action. Routledge.

Schubert, A., & Schubert, G. (2020). Internationality at university level. *Scientometrics*, *123*(3), 1341-1364. https://doi.org/10.1007/s11192-020-03443-3

Schuff, D., Corral, K., & Turetken, O. (2011). Comparing the understandability of alternative data warehouse schemas: An empirical study. Decision Support Systems, 52(1), 9-20. https://doi.org/10.1016/j.dss.2011.04.003

Schupp, H. T., Stockburger, J., Codispoti, M., Junghofer, M., Weike, A. I., & Hamm, A. O. (2007). Selective visual attention to emotion. Journal of Neuroscience, 27(5), 1082-1089. https://doi.org/10.1523/jneurosci.3223-06.2007

Scopus. (2002). *Research Intelligence. Scopus Content Coverage Guide*. Elsevier | An Information Analytics Business | Empowering Knowledge. https://www.elsevier.com/?a=69451

Sears, A., & Jacko, J. A. (2012). Emotion in Human–Computer Interaction Scott Brave and Clifford Nass. In The Human-Computer Interaction Handbook (pp. 113-128). CRC Press.

Sekhon, J. S., & Titiunik, R. (2012). When natural experiments are neither natural nor experiments. *American Political Science Review*, *106*(1), 35-57. https://doi.org/10.1017/s0003055411000542

Shen, L., Wang, M., & Shen, R. (2009). Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment. Journal of Educational Technology & Society, 12(2), 176-189. https://doi.org/10.1007/978-3-540-85170-7_16

Skavronskaya, L., Moyle, B., & Scott, N. (2020). The experience of novelty and the novelty of experience. Frontiers in Psychology, 11. https://doi.org/10.3389/fpsyg.2020.00322

So, H. Y., Chen, P. P., Wong, G. K., & Chan, T. T. (2019). Simulation in medical education. Journal of the Royal College of Physicians of Edinburgh, 49(1), 52-57. https://doi.org/10.4997/jrcpe.2019.112

Sørensen, J. L., Østergaard, D., LeBlanc, V., Ottesen, B., Konge, L., Dieckmann, P., & Van der Vleuten, C. (2017). Design of simulation-based medical education and advantages and disadvantages of in situ simulation versus off-site simulation. BMC Medical Education, 17(1). https://doi.org/10.1186/s12909-016-0838-3

Sui, J., & Humphreys, G. W. (2015). The integrative self: How self-reference integrates perception and memory. Trends in Cognitive Sciences, 19(12), 719-728. https://doi.org/10.1016/j.tics.2015.08.015

Sundarasen, S., Chinna, K., Kamaludin, K., Nurunnabi, M., Baloch, G. M., Khoshaim, H. B., Hossain, S. F., & Sukayt, A. (2020). Psychological impact of COVID-19 and lockdown among University students in Malaysia: Implications and policy recommendations. International Journal of Environmental Research and Public Health, 17(17), 6206. https://doi.org/10.3390/ijerph17176206

Tao, J., & Tan, T. (2005). Affective computing: A review. In International Conference on Affective computing and intelligent interaction (pp. 981-995). Springer, Berlin, Heidelberg.

Taylor, H., & Reyes, H. (2012). Self-efficacy and resilience in baccalaureate nursing students. International Journal of Nursing Education Scholarship, 9(1), 1-13. https://doi.org/10.1515/1548-923x.2218

Taylor, R., Jung, J., Loewen, P., Spencer, C., Dossa, A., & De Lemos, J. (2013). Online versus live delivery of education to pharmacists in a large multicentre health region: A non-inferiority assessment of learning outcomes. *The Canadian Journal of Hospital Pharmacy*, *66*(4). https://doi.org/10.4212/cjhp.v66i4.1271

Thompson, E., Millman, Z. B., Okuzawa, N., Mittal, V., DeVylder, J., Skadberg, T., Buchanan, R. W., Reeves, G. M., & Schiffman, J. (2015). undefined. Journal of Nervous & Mental Disease, 203(5), 342-351. https://doi.org/10.1097/nmd.0000000000000287

Triberti, S., Chirico, A., La Rocca, G., & Riva, G. (2017). Developing emotional design: Emotions as cognitive processes and their role in the design of interactive technologies. Frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.01773

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-scr): Checklist and explanation. Annals of Internal Medicine, 169(7), 467-473. https://doi.org/10.7326/m18-0850

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Publishing Company.

Turner, J. R., & Baker, R. (2020). Collaborative research: Techniques for conducting collaborative research from the science of team science (SciTS). *Advances in Developing Human Resources*, *22*(1), 72-86. https://doi.org/10.1177/1523422319886300

Turrise, S. L., Thompson, C. E., & Hepler, M. (2020). Virtual simulation: Comparing critical thinking and satisfaction in RN-BSN students. *Clinical Simulation in Nursing*, *46*, 57-61. https://doi.org/10.1016/j.ecns.2020.03.004

Tyng, C. M., Amin, H. U., Saad, M. N., & Malik, A. S. (2017). The influences of emotion on learning and memory. Frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.01454

Valiente, C., Swanson, J., DeLay, D., Fraser, A. M., & Parker, J. H. (2020). Emotion-related socialization in the classroom: Considering the roles of teachers, peers, and the classroom context. Developmental Psychology, 56(3), 578-594. https://doi.org/10.1037/dev0000863

Van Eck, N. J., & Waltman, L. (2007). VOS: A new method for visualizing similarities between objects. *Studies in Classification, Data Analysis, and Knowledge Organization*, 299-306. https://doi.org/10.1007/978-3-540-70981-7_34

Van Eck, N. J., & Waltman, L. (2014). Visualizing bibliometric networks. *Measuring Scholarly Impact*, 285-320. https://doi.org/10.1007/978-3-319-10377-8_13

Van Gorp, L., Boroş, S., Bracke, P., & Stevens, P. A. (2017). Emotional support on re-entry into the home country: Does it matter for repatriates' adjustment who the providers are? International Journal of Intercultural Relations, 58, 54-68. https://doi.org/10.1016/j.ijintrel.2017.04.007

Veldhuijzen, G., Klemt-Kropp, M., Terhaar sive Droste, J. S., Van Balkom, B., Van Esch, A. A., & Drenth, J. P. (2021). Computer-based patient education is non-inferior to nurse counselling prior to colonoscopy: A multicenter randomized controlled trial. *Endoscopy*, *53*(03), 254-263. https://doi.org/10.1055/a-1225-8708

Virtanen, K., Mansikka, H., Kontio, H., & Harris, D. (2021). Weight watchers: NASA-TLX weights revisited. *Theoretical Issues in Ergonomics Science*, 1-24. https://doi.org/10.1080/1463922x.2021.2000667

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and Noninferiority testing. *Journal of General Internal Medicine*, *26*(2), 192-196. https://doi.org/10.1007/s11606-010-1513-8

Walters, S. (2020). The Experience of Emotion. Psychology-1st Canadian Edition, 56, 100751. https://doi.org/10.1016/j.newideapsych.2019.100751

Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., Li, Y., Meng, X., & Xu, B. (2021). A deep learning algorithm using CT images to screen for corona virus disease (COVID-19). European Radiology. https://doi.org/10.1007/s00330-021-07715-1

Watson, D., & Tellegen, A. (1999). Issues in dimensional structure of affect--effects of descriptors, measurement error, and response formats: Comment on Russell and Carroll (1999). Psychological Bulletin, 125(5), 601-610. https://doi.org/10.1037/0033-2909.125.5.601

Watson, L., & Spence, M. T. (2007). Causes and consequences of emotions on consumer behaviour. European Journal of Marketing, 41(5/6), 487-511. https://doi.org/10.1108/03090560710737570

Weeks, D. L., & Molsberry, D. M. (2008). Pediatric advanced life support re-training by videoconferencing compared to face-to-face instruction: A planned non-inferiority trial. *Resuscitation*, *79*(1), 109-117. https://doi.org/10.1016/j.resuscitation.2008.05.008

Wenk, M., Waurick, R., Schotes, D., Wenk, M., Gerdes, C., Van Aken, H. K., & Pöpping, D. M. (2008). Simulation-based medical education is no better than

problem-based discussions and induces misjudgment in self-assessment. Advances in Health Sciences Education, 14(2), 159-171. https://doi.org/10.1007/s10459-008-9098-2

White, H., & Sabarwal, S. (2014). Quasi-experimental design and methods. *Methodological briefs: impact evaluation*, *8*, 1-16.

Widmark, A., Gunnlaugsson, A., Beckman, L., Thellenberg-Karlsson, C., Hoyer, M., Lagerlund, M., Kindblom, J., Ginman, C., Johansson, B., Björnlinger, K., Seke, M., Agrup, M., Fransson, P., Tavelin, B., Norman, D., Zackrisson, B., Anderson, H., Kjellén, E., Franzén, L., … Nilsson, P. (2019). Ultra-hypofractionated versus conventionally fractionated radiotherapy for prostate cancer: 5-year outcomes of the HYPO-RT-PC randomised, non-inferiority, phase 3 trial. *The Lancet*, *394*(10196), 385-395. https://doi.org/10.1016/s0140-6736(19)31131-6

Wiener, F., & Vinaver, J. (1977). Computer simulation of medical reasoning in the assessment of kidney disease. *International Journal of Bio-Medical Computing*, *8*(3), 203-215. https://doi.org/10.1016/0020-7101(77)90004-6

Williams, B., Abel, C., Khasawneh, E., Ross, L., & Levett-Jones, T. (2016). undefined. Advances in Medical Education and Practice, 181. https://doi.org/10.2147/amep.s98462

Wolff, A. C., Pesut, B., & Regan, S. (2010). New graduate nurse practice readiness: Perspectives on the context shaping our understanding and expectations. Nurse Education Today, 30(2), 187-191. https://doi.org/10.1016/j.nedt.2009.07.011

Wosnitza, M., & Volet, S. (2005). Origin, direction and impact of emotions in social online learning. Learning and Instruction, 15(5), 449-464. https://doi.org/10.1016/j.learninstruc.2005.07.009

Wu, C., Huang, Y., & Hwang, J. (2015). Review of affective computing in education/learning: Trends and challenges. British Journal of Educational Technology, 47(6), 1304-1323. https://doi.org/10.1111/bjet.12324

Yang, W., Liu, J., Shan, Z., Tian, H., Zhou, Z., Ji, Q., Weng, J., Jia, W., Lu, J., Liu, J., Xu, Y., Yang, Z., & Chen, W. (2014). Acarbose compared with metformin as initial therapy in patients with newly diagnosed type 2 diabetes: An open-label, non-inferiority randomised trial. *The Lancet Diabetes & Endocrinology*, *2*(1), 46-55. https://doi.org/10.1016/s2213-8587(13)70021-4

You, J. W., & Kang, M. (2014). The role of academic emotions in the relationship between perceived academic control and self-regulated learning in online learning. Computers & Education, 77, 125-133. https://doi.org/10.1016/j.compedu.2014.04.018

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2014). State of science: Mental workload in ergonomics. Ergonomics, 58(1), 1-17. https://doi.org/10.1080/00140139.2014.956151

Young, J. Q., Thakker, K., John, M., Friedman, K., Sugarman, R., Van Merriënboer, J. J., Sewell, J. L., & O'Sullivan, P. S. (2021). Exploring the relationship between emotion and cognitive load types during patient handovers. *Advances in Health Sciences Education*, *26*(5), 1463-1489. https://doi.org/10.1007/s10459-021-10053-y

Ypinazar, V., & Margolis, S. (2006). Clinical simulators: Applications and implications for rural medical education. Rural and Remote Health. https://doi.org/10.22605/rrh527

Yurko, Y. Y., Scerbo, M. W., Prabhu, A. S., Acker, C. E., & Stefanidis, D. (2010). Higher mental workload is associated with poorer Laparoscopic performance as measured by the NASA-TLX tool. Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare, 5(5), 267-271. https://doi.org/10.1097/sih.0b013e3181e3f329

Ziv, A., Wolpe, P. R., Small, S. D., & Glick, S. (2003). Simulation-based medical education. *Academic Medicine*, *78*(8), 783-788. https://doi.org/10.1097/00001888-200308000-00006

Zou, H., Li, N., & Cao, L. (2017). Emotional response–based approach for assessing the sense of presence of subjects in virtual building evacuation studies. Journal of Computing in Civil Engineering, 31(5), 04017028. https://doi.org/10.1061/(asce)cp.1943-5487.0000679

Zupic, I., & Čater, T. (2013). Bibliometric methods in management and organization: A review. *Organizational research methods*, *18*(3), 429-472. https://doi.org/10.2139/ssrn.2267251

Appendix 1: Research Instrument

Please type the Simulation Session Code given by your instructor:

What is your gender ?

- ☐ Male
- ☐ Female
- ☐ Other (specify)

What age category do you fall into ?

- ○ Less than 21
- ○ 22 – 25
- ○ 26 – 31
- ○ 32 – 37
- ○ 37 – 40
- ○ Over 40

Which race/ethnicity best describes you?

- ☐ American Indian or Alaskan Native
- ☐ Asian/Pacific Islander
- ☐ Black or African American
- ☐ Hispanic/Latin
- ☐ White/caucasian
- ☐ Decline to state
- ☐ Other not specified

Please select your class standing (if student) or your current professional situation :

- ○ Undergrad / Pre-Med
- ○ Graduate Student (Master/MSN/DNP/PhD/MD student)
- ○ Resident / Fellow
- ○ Nurse, MD or other healthcare professional (not currently in school/residency)

Overall, how many semesters of experience with traditional simulation (non virtual) do you have ? (either as an observer or as participant)

Less than 1 semester	1 - 2 semesters	3 - 4 semesters	5 - 6 semesters	Over 6 semesters
○	○	○	○	○

Estimate here how aroused you actually feel. By “arousal” we meant how “worked-up” you feel.

You might experience high arousal in one of a variety of ways, for example as excitement or anxiety or anger. Low arousal might also be experienced by you in one of a number of different ways, for example as relaxation or boredom or calmness

Low Arousal	○ ○ ○ ○ ○ ○	High Arousal

Please select a point between each pair of emotions that best describes how you feel now.
The selected point should be closer to the emotion you felt

Bored	○ ○ ○ ○ ○	Alert
Lethargic	○ ○ ○ ○ ○	Excited
Depressed	○ ○ ○ ○ ○	Elated
Tense	○ ○ ○ ○ ○	Calm
Sad	○ ○ ○ ○ ○	Happy
Upset	○ ○ ○ ○ ○	Contended
Stressed	○ ○ ○ ○ ○	Serene
Nervous	○ ○ ○ ○ ○	Relaxed

Please select your role during the Virtual Simulation Session:

Leader	Team Member
○	◉

Please answer the following questions regarding your Perception of Fidelity (Realism) of the virtual simulation

	Strongly disagree	Disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree	NA (Not Applicable)
The scenario resembled a real-life situation	○	○	○	○	○	○	○
Real life factors, situation, and variables were built into the simulation scenario	○	○	○	○	○	○	○

If your level of presence in the real world is "100" and your level of presence is "1"' if you lack presence, rate your level of presence in this virtual world (presence is a "feeling of being there"). Enter a number 1-100.

0 10 20 30 40 50 60 70 80 90 100

Rate your level of presence

Estimate here how aroused you actually feel. By “arousal” we meant how “worked-up” you feel.

You might experience high arousal in one of a variety of ways, for example as excitement or anxiety or anger. Low arousal might also be experienced by you in one of a number of different ways, for example as relaxation or boredom or calmness

Low Arousal	○ ○ ○ ○ ○ ○	High Arousal

Please select a point between each pair of emotions that best describes how you felt during the simulation (not how you felt either before or after the simulation). The selected point should be closer to the emotion you felt .

Bored	○ ○ ○ ○ ○	Alert
Lethargic	○ ○ ○ ○ ○	Excited
Depressed	○ ○ ○ ○ ○	Elated
Tense	○ ○ ○ ○ ○	Calm
Sad	○ ○ ○ ○ ○	Happy
Upset	○ ○ ○ ○ ○	Contended
Stressed	○ ○ ○ ○ ○	Serene
Nervous	○ ○ ○ ○ ○	Relaxed

Please select the best option regarding your experience in the simulation session you just participated. The selected point should be closer to your response to the question:

How enjoyable was your overall experience ?	not at all enjoyable ○	2 3 4 5 6 ○ ○ ○ ○ ○	extremely enjoyable ○
How pleasant or unpleasant was your overall experience ?	not at all pleasant ○	2 3 4 5 6 ○ ○ ○ ○ ○	extremely pleasant ○
To what extent was the overall experience consistent to what you wanted to achieve during this session?	very much inconsistent ○	2 3 4 5 6 ○ ○ ○ ○ ○	very much consistent ○
In comparison to what you desired, how consistent was your overall experience ?	very much inconsistent ○	2 3 4 5 6 ○ ○ ○ ○ ○	very much consistent ○
To what extent do you see the overall experience in this session important or not important in achieving your learning goals or needs?	not at all important ○	2 3 4 5 6 ○ ○ ○ ○ ○	extremely important ○
How well did you understand what was happening around you?			a great deal

	no understanding	2	3	4	5	6	of understanding
How sure were you about what was happening?	not sure at all	2	3	4	5	6	extremely sure
To what extent did the overall experience was something you expected or unexpected?	unexpected	2	3	4	5	6	expected
How usual or unusual was your overall experience ?	unusual	2	3	4	5	6	usual
To what extent was the overall experience you had in the simulation session familiar or unfamiliar ?	unfamiliar	2	3	4	5	6	familiar

Please select the best option for each question regarding your experience in the simulation session:

	Not at all					Extremely
How well could you predict what was going to happen ?						
To what extent did you feel that circumstances beyond anyone's control determined the outcomes/results of this simulation?						
How responsible did you think someone or something other than yourself was for the outcomes/results of the simulation session?						
When you were in this session, how powerful (dominant) did you feel?						
When you were in this simulation session, how helpless did you feel?						
	Not at all					Extremely
To what extent did you feel that you could influence what was happening in the simulation session?						
How important was this simulation session to you?						
To what extent did you feel cheated or wronged in this simulation session ?						
How fair did you think what happened to you in this simulation session was?						

Considering the simulation session you just participated, please respond to the following questions:

Mental Demand- How mentally demanding was the simulation?

Very Low ○ ○ ○ ○ ○ ○ ○ Very High

Physical Demand - How physically demanding was the task?

Very Low ○ ○ ○ ○ ○ ○ ○ Very High

Temporal Demand - How hurried or rushed was the pace of the simulation?

Very Low ○ ○ ○ ○ ○ ○ ○ Very High

Performance - How successful were you in accomplishing what you were asked to do?

Good ○ ○ ○ ○ ○ ○ ○ Poor

Effort - How hard did you have to work to accomplish your level of performance?

Very Low ○ ○ ○ ○ ○ ○ ○ Very High

Frustration - How insecure, discouraged, irritated, stressed, and annoyed were you?

Very Low ○ ○ ○ ○ ○ ○ ○ Very High

Regarding the case/setting/condition to be depicted in the current simulation session, how many times have you seen it in clinical settings (not counting simulation sessions) ?

Never Seen ○ Seen once ○ Seen 2-4 times ○ Seen > 4 times ○

→

Considering the debriefing session you just participated, please respond to the following questions:

Estimate here how aroused you actually feel. By “arousal” we meant how “worked-up” you feel.

You might experience high arousal in one of a variety of ways, for example as excitement or anxiety or anger. Low arousal might also be experienced by you in one of a number of different ways, for example as relaxation or boredom or calmness

Low Arousal	○ ○ ○ ○ ○ ○	High Arousal

Please select a point between each pair of emotions that best describes how you felt during the debriefing (not how you felt either before or after the debriefing). The selected point should be closer to the emotion you felt .

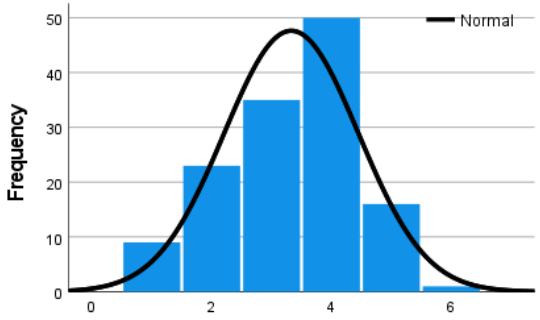
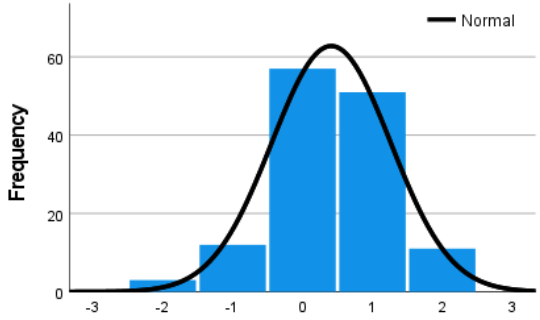
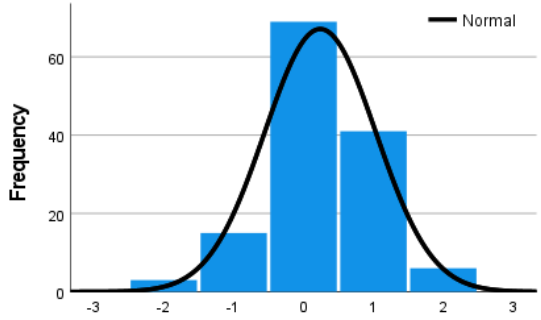
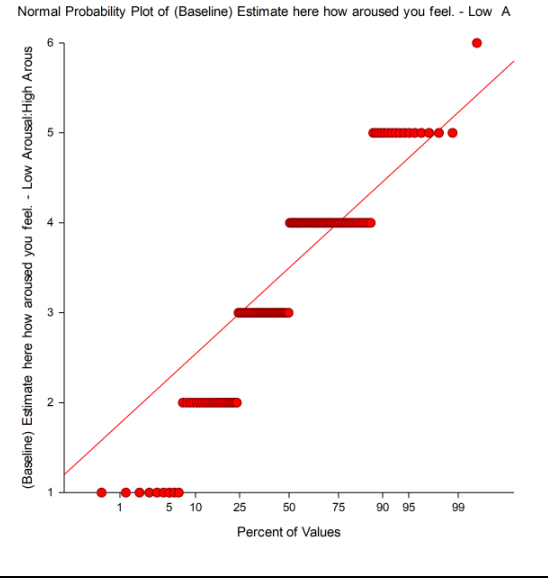
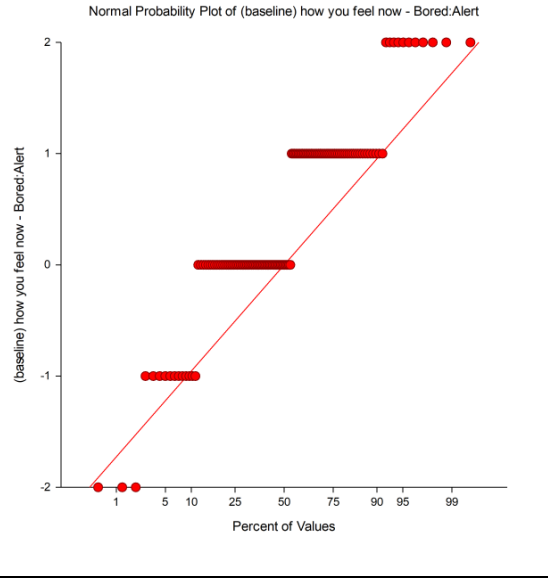
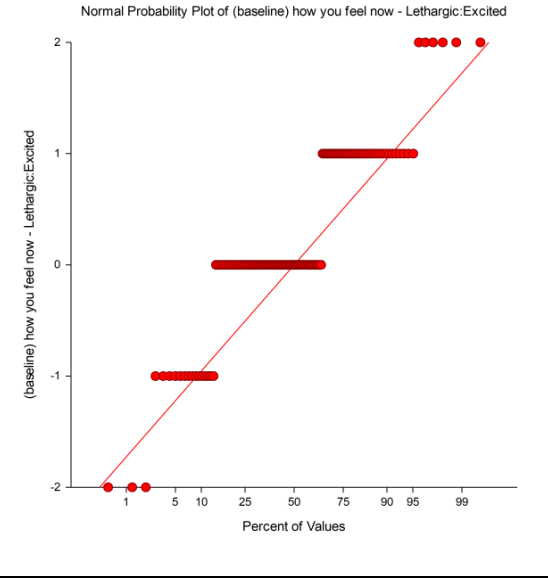
Bored	○ ○ ○ ○ ○	Alert
Lethargic	○ ○ ○ ○ ○	Excited
Depressed	○ ○ ○ ○ ○	Elated
Tense	○ ○ ○ ○ ○	Calm
Sad	○ ○ ○ ○ ○	Happy
Upset	○ ○ ○ ○ ○	Contended
Stressed	○ ○ ○ ○ ○	Serene
Nervous	○ ○ ○ ○ ○	Relaxed

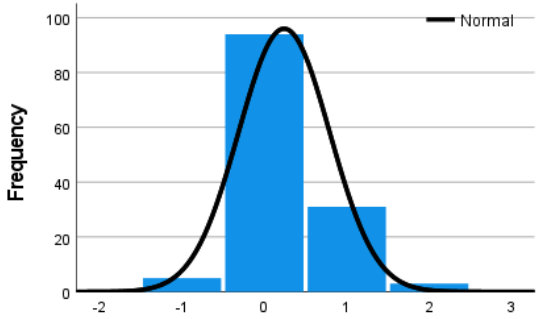
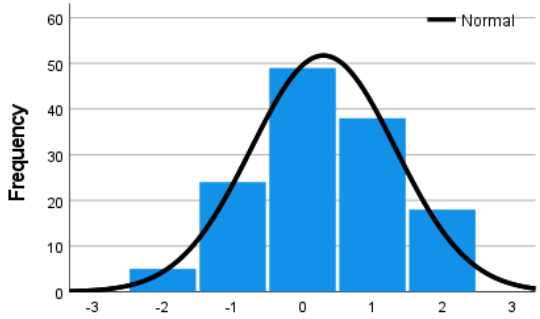
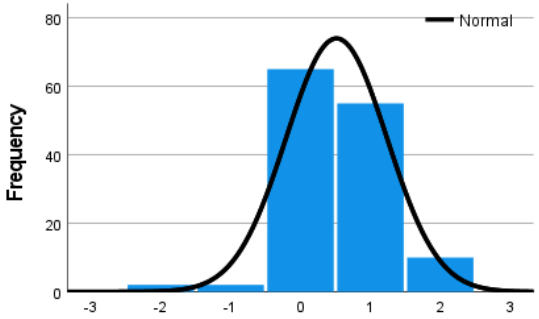
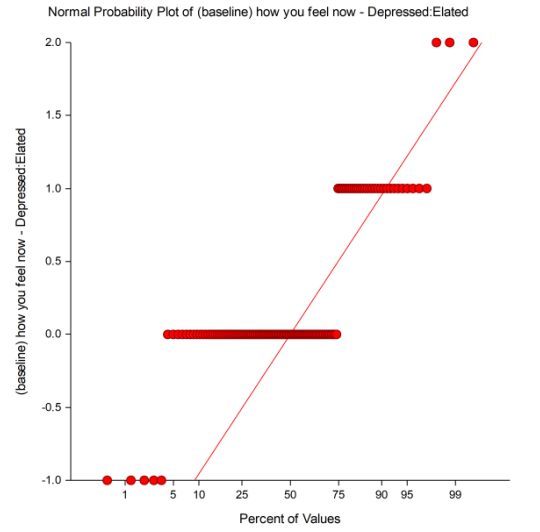
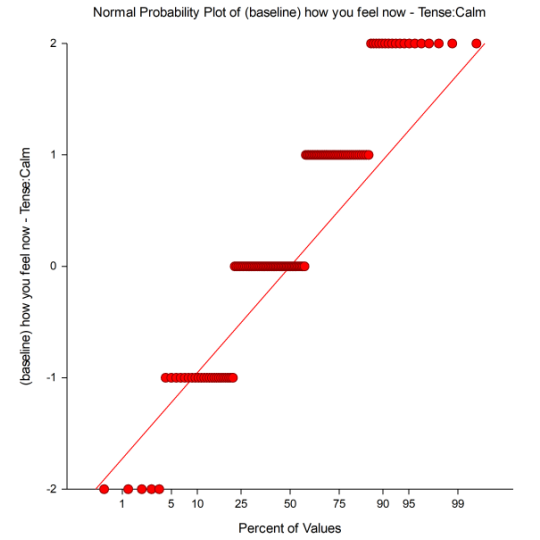
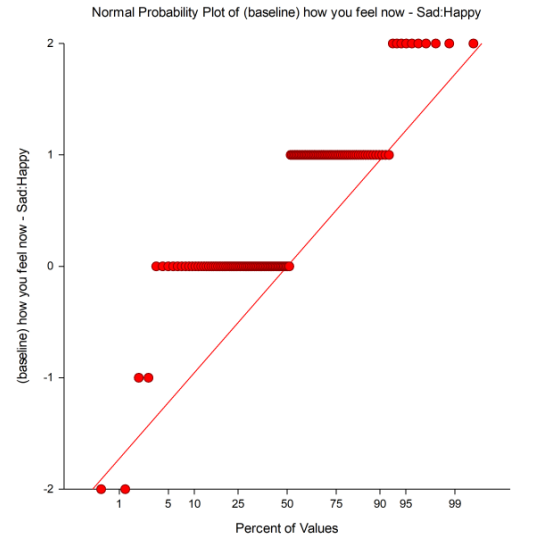
Please answer the following questions regarding Feedback/Guided Reflection

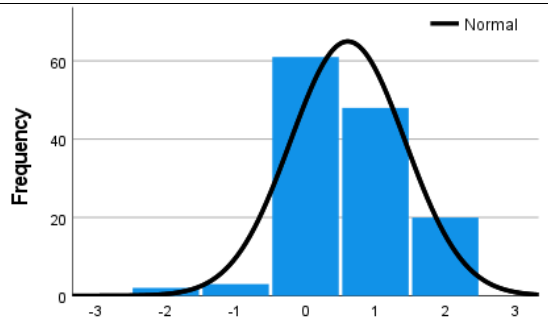
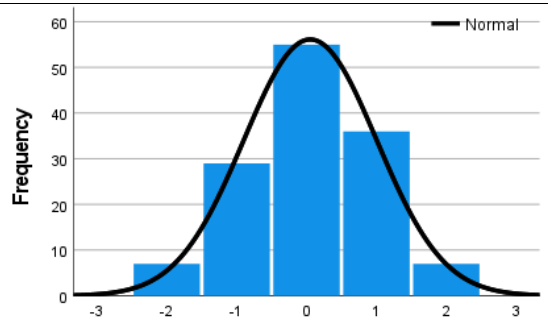
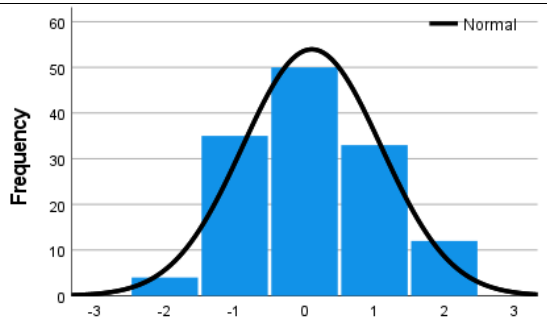
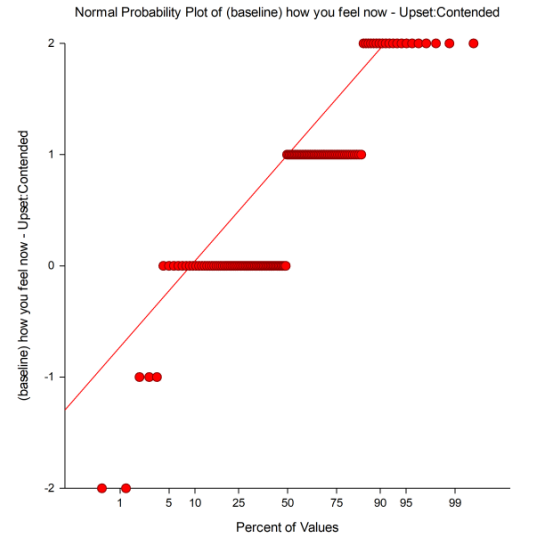
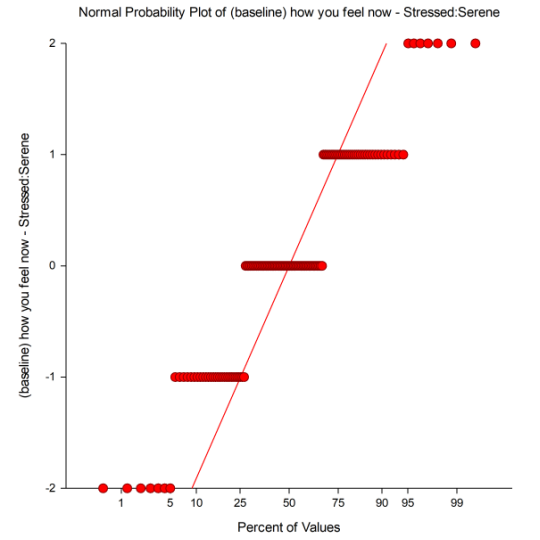
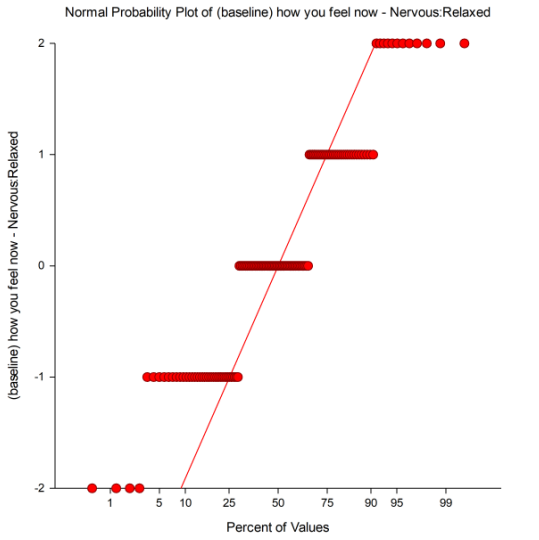
	Strongly disagree	Disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree	NA (Not Applicable)
Feedback provided has constructive	○	○	○	○	○	○	○
Feedback was provided in a timely manner	○	○	○	○	○	○	○
The simulation allowed me to analyze my own behavior and actions	○	○	○	○	○	○	○
There was an opportunity after the simulation to obtain guidance/feedback from the teacher in order to build knowledge to another level	○	○	○	○	○	○	○

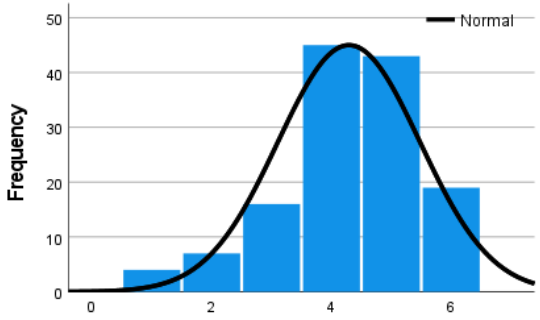
Appendix 2: Frequency distribution of study variables

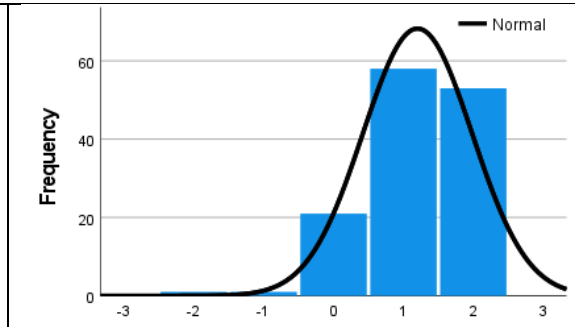
Emotion State Variables (Baseline, Post-Simulation, and Post-Debriefing)

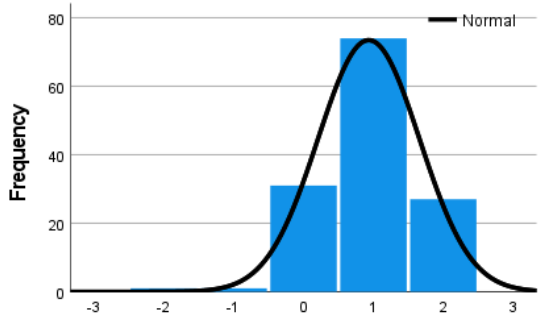
Q2.7_1	Q2.8_1	Q2.8_2
		
		

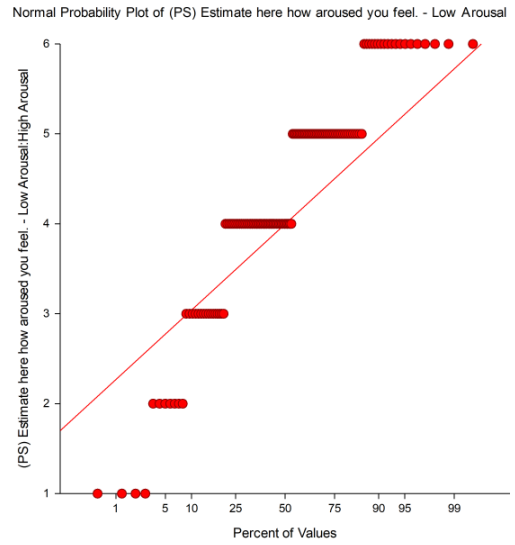
Q2.8_3	Q2.8_4	Q2.8_5
		
		
Q2.8_6	Q2.8_7	Q2.8_8

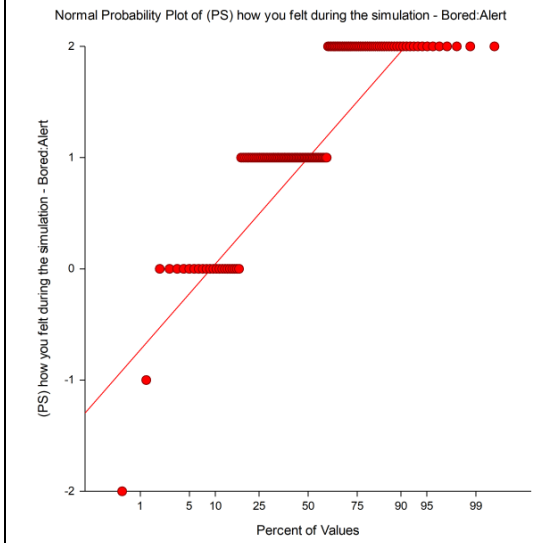
		
		
Q4.5_1	**Q4.6_1**	**Q4.6_2**

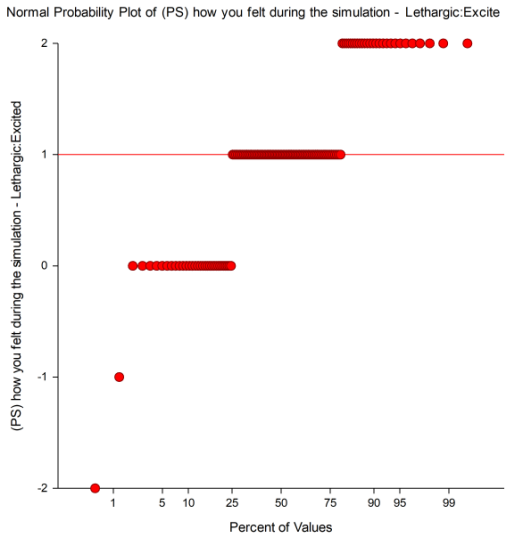




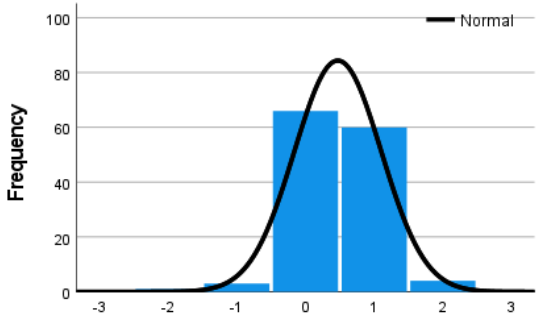


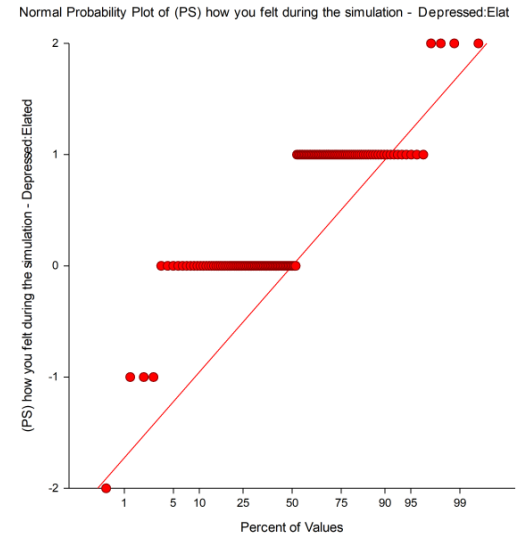




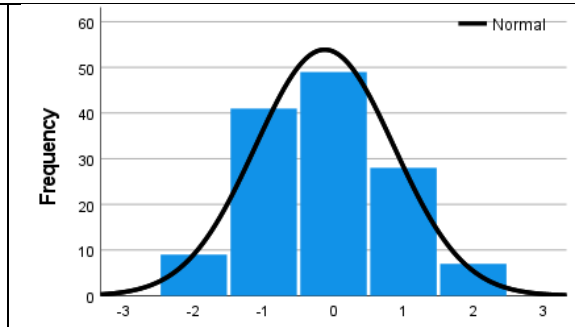


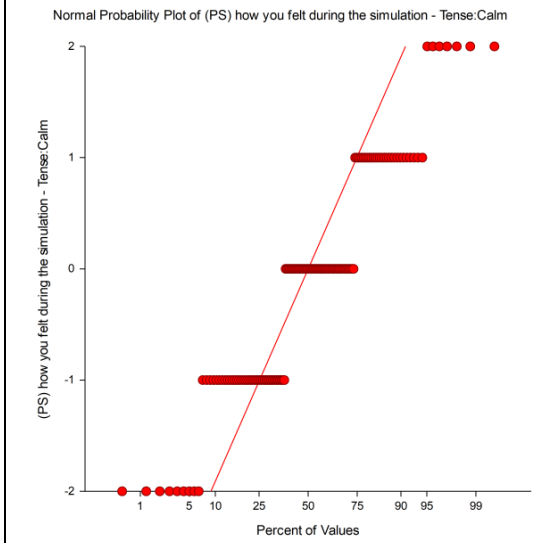
Q4.6_3	Q4.6_4	Q4.6_5



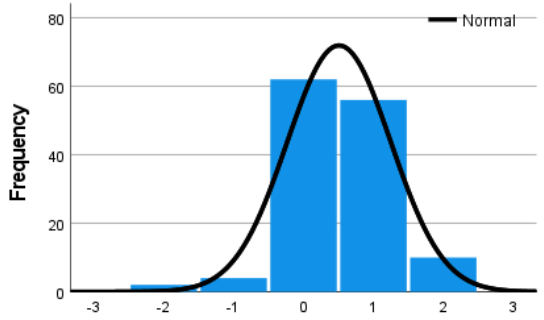


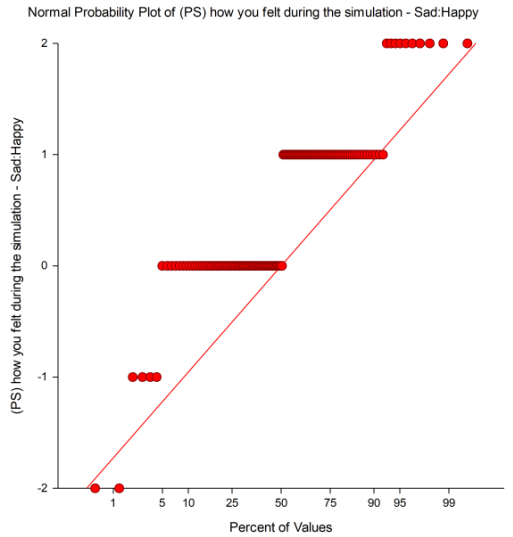
Q4.6_6



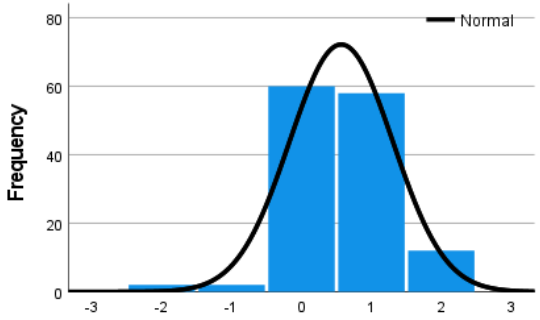


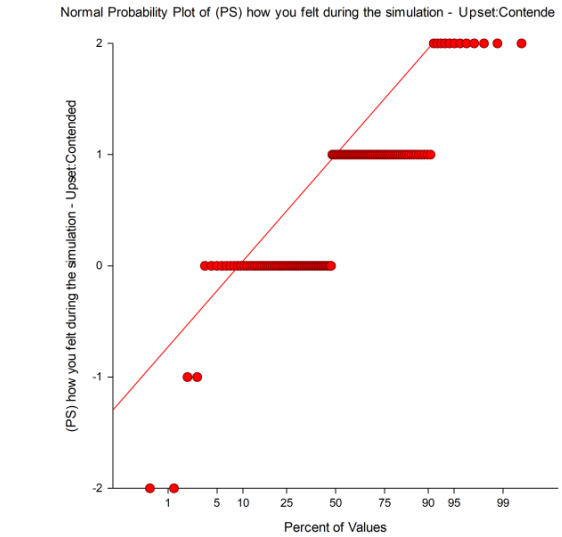
Q4.6_7



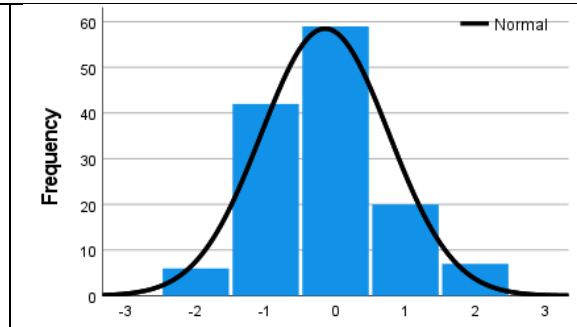


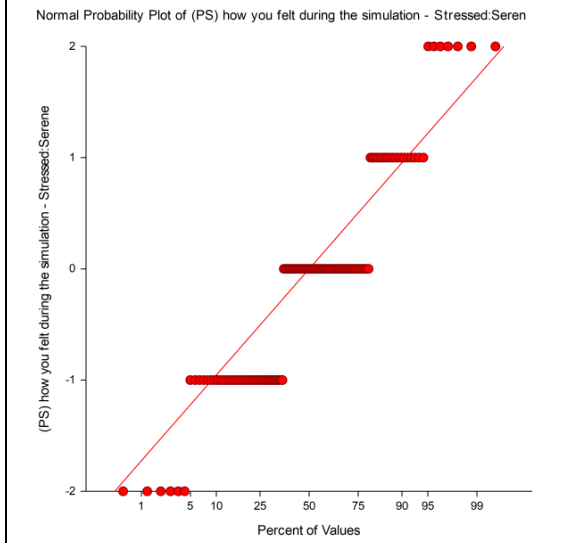
Q4.6_8



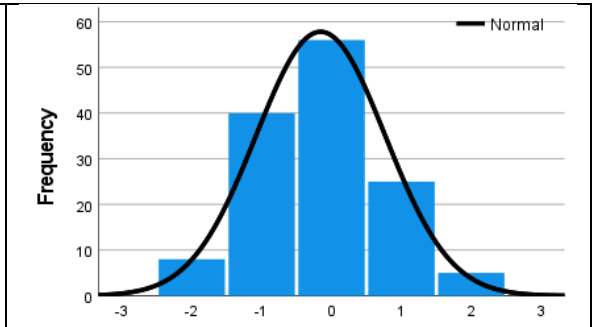


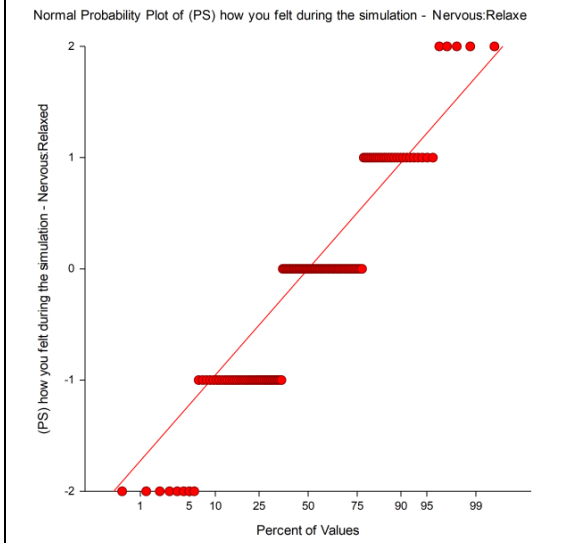
Q6.2_1



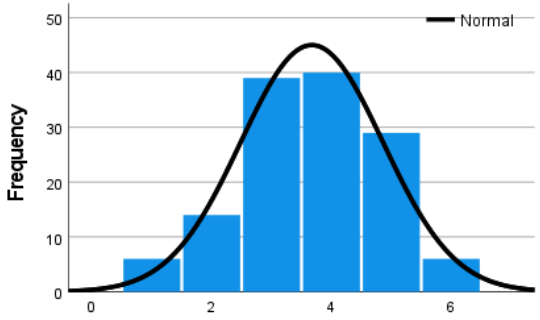
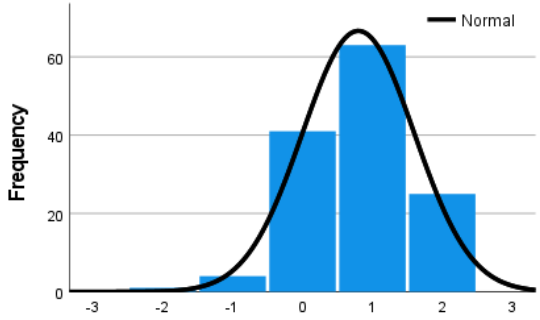
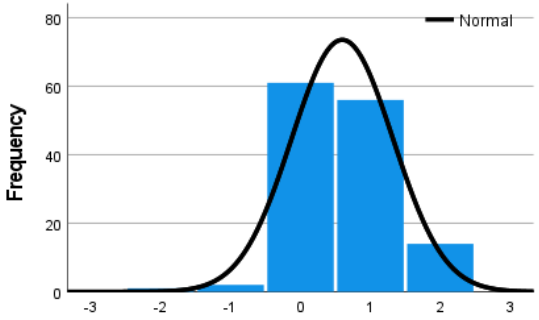
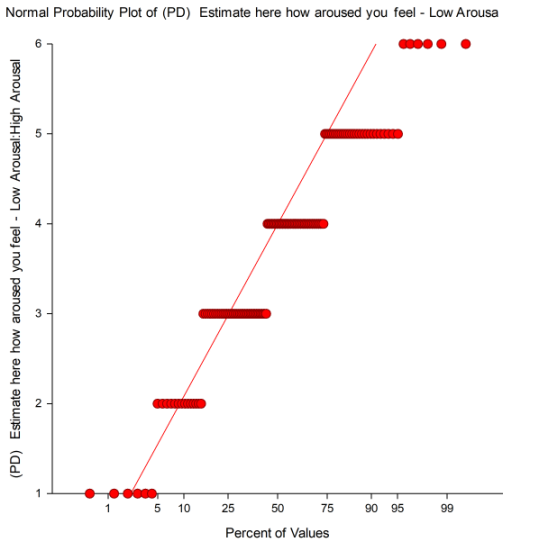
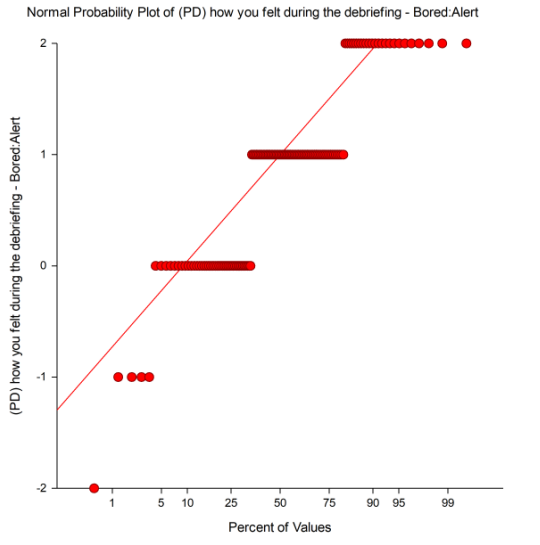
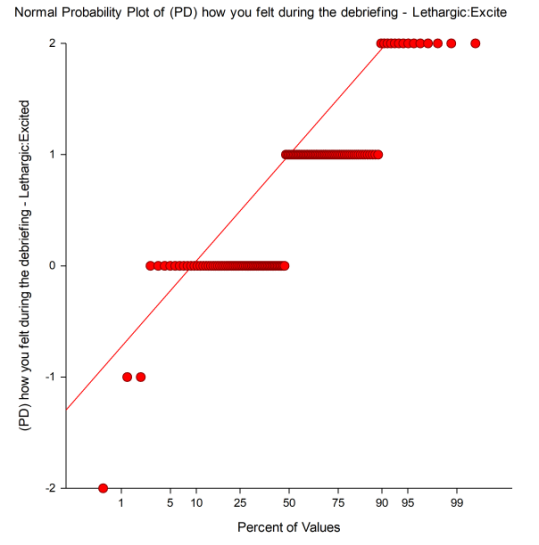


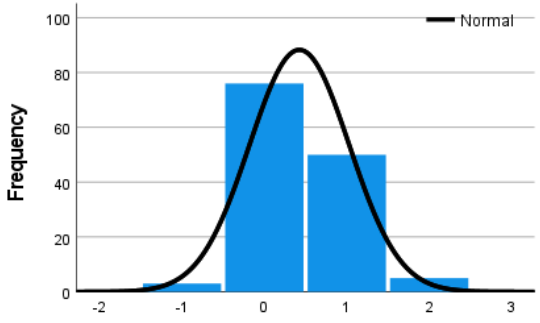
Q6.3_1

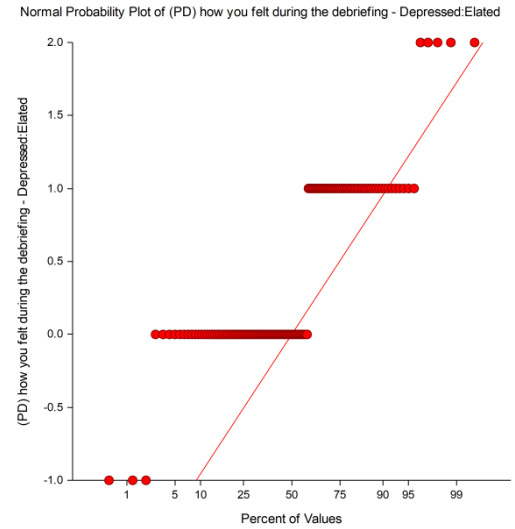




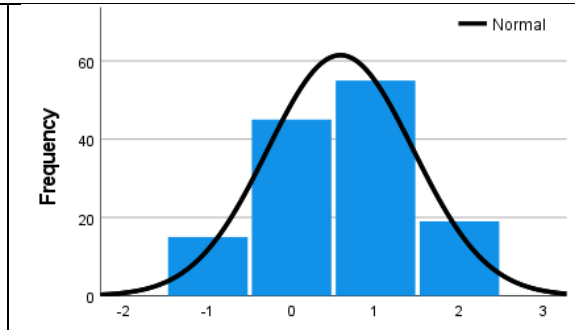
Q6.3_2

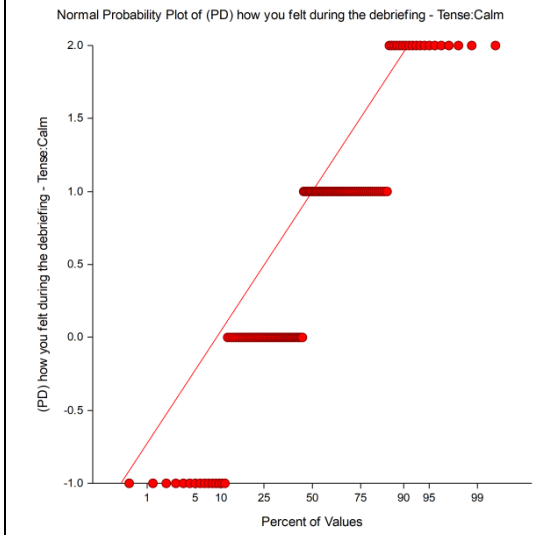
		
		
Q6.3_3	**Q6.3_4**	**Q6.3_5**



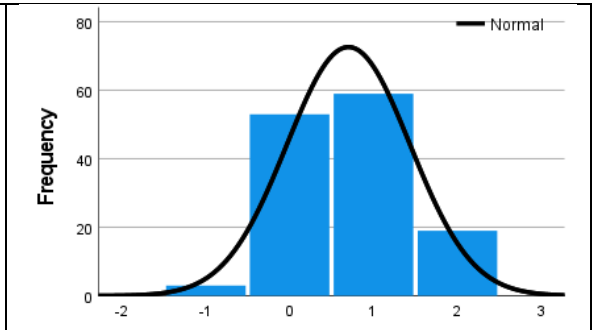


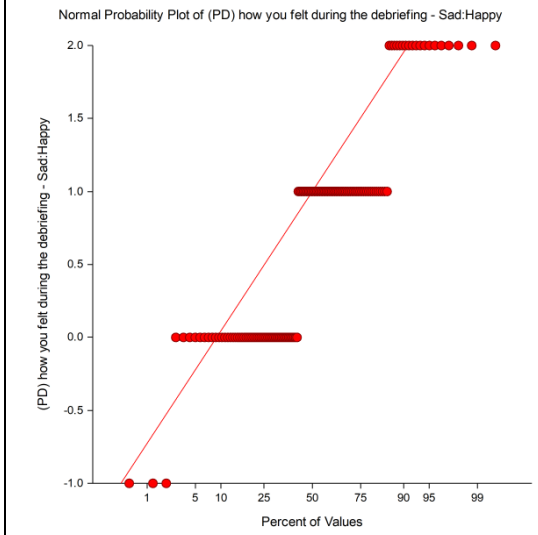
Q6.3_6



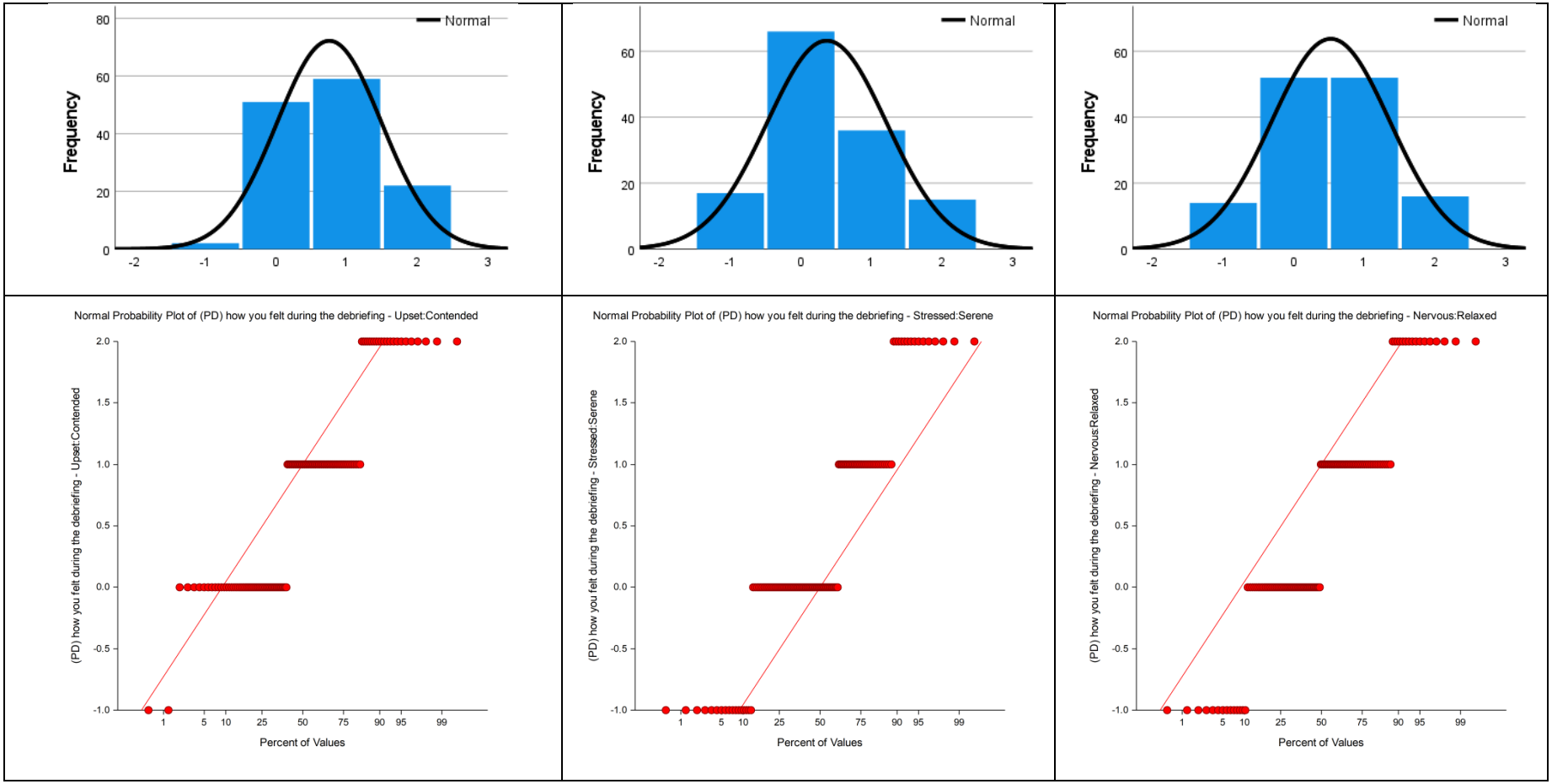


Q6.3_7

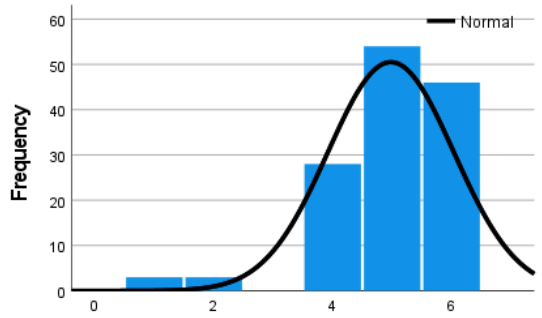
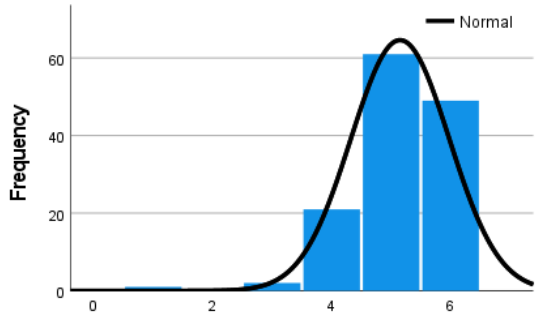
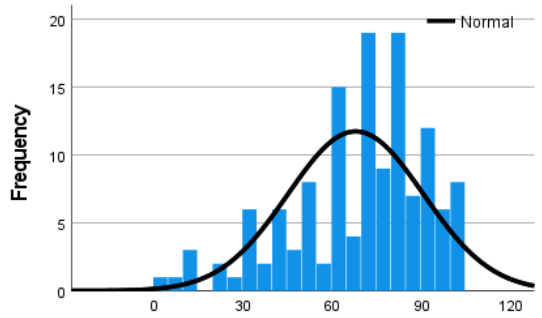
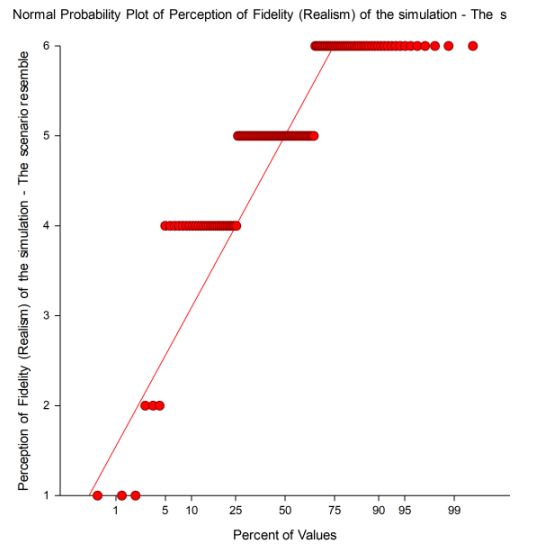
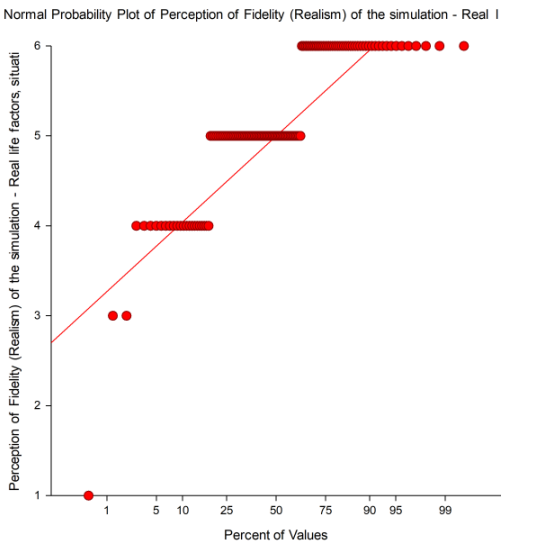
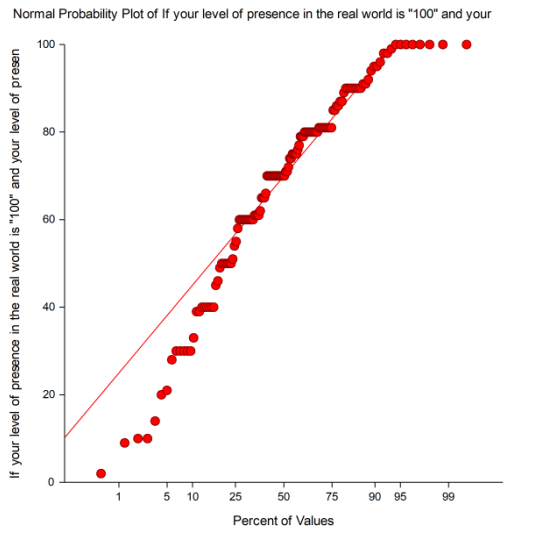


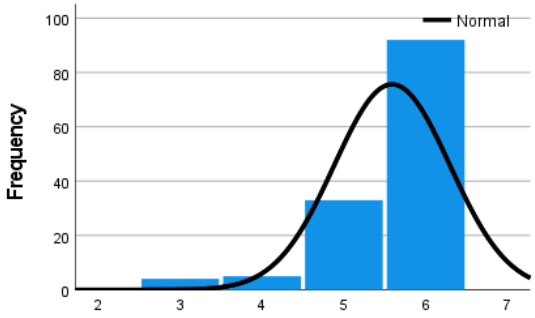


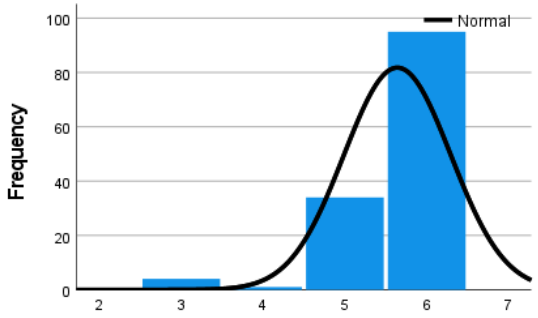
Q6.3_8

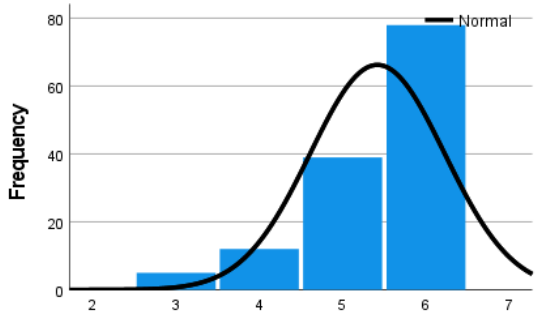

Normal
Frequency
Normal Probability Plot of (PD) how you felt during the debriefing - Upset:Contended
(PD) how you felt during the debriefing - Upset:Contended
Normal Probability Plot of (PD) how you felt during the debriefing - Stressed:Serene
(PD) how you felt during the debriefing - Stressed:Serene
Normal Probability Plot of (PD) how you felt during the debriefing - Nervous:Relaxed
(PD) how you felt during the debriefing - Nervous:Relaxed
Percent of Values

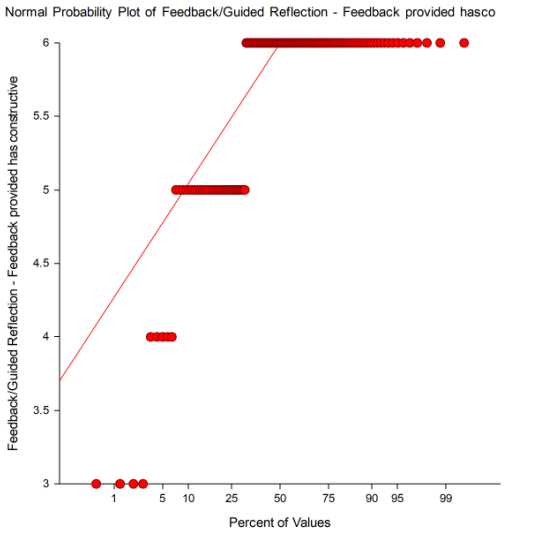
Feedback and Perception of Presence Variables (Post-Simulation, and Post-Debriefing)

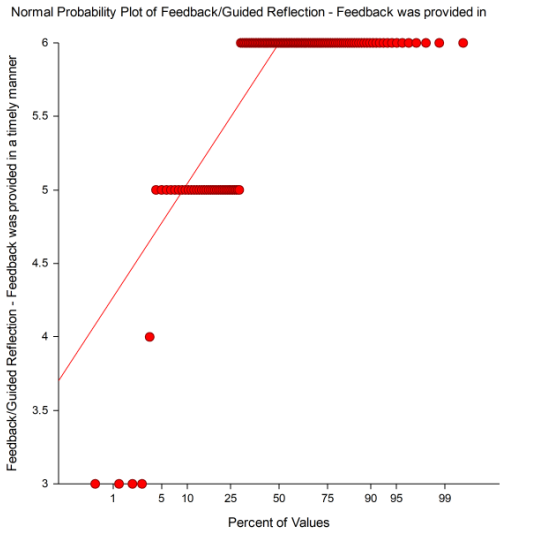
Q4.3_1	Q4.3_2	Q4.4_1
		
		
Q6.4_1	Q6.4_2	Q6.4_3

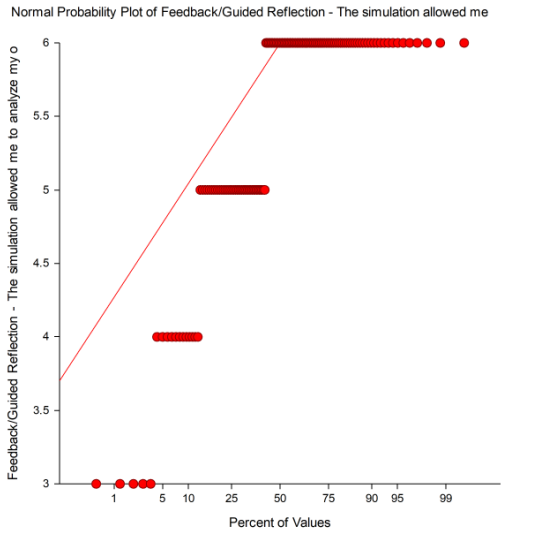

Normal
Frequency
100
80
60
40
20
0
2
3
4
5
6
7


Normal
Frequency
100
80
60
40
20
0
2
3
4
5
6
7

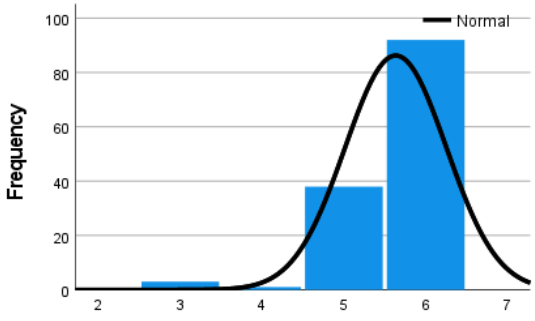

Normal
Frequency
80
60
40
20
0
2
3
4
5
6
7

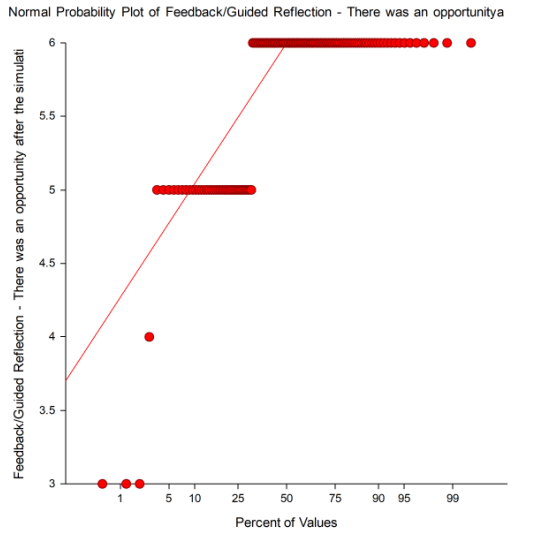

Normal Probability Plot of Feedback/Guided Reflection - Feedback provided hasco
Feedback/Guided Reflection - Feedback provided has constructive
6
5.5
5
4.5
4
3.5
3
1
5
10
25
50
75
90
95
99
Percent of Values


Normal Probability Plot of Feedback/Guided Reflection - Feedback was provided in
Feedback/Guided Reflection - Feedback was provided in a timely manner
6
5.5
5
4.5
4
3.5
3
1
5
10
25
50
75
90
95
99
Percent of Values

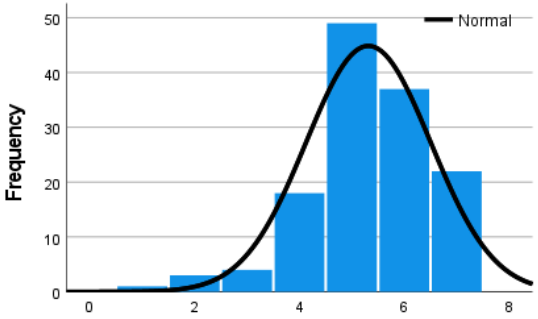
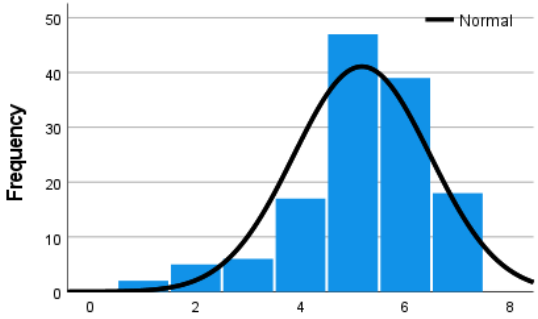
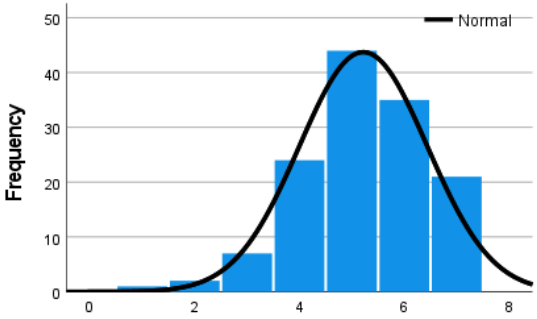
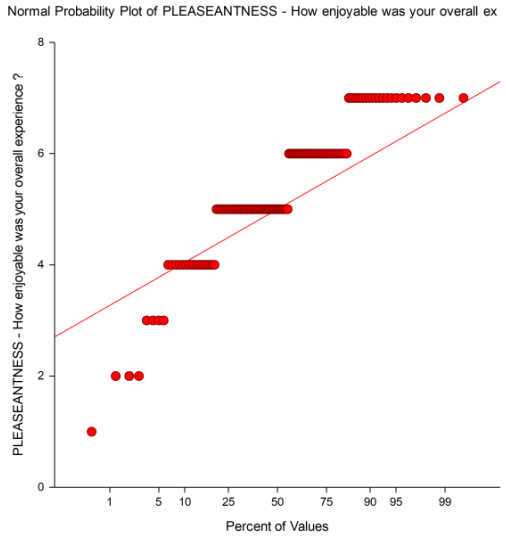
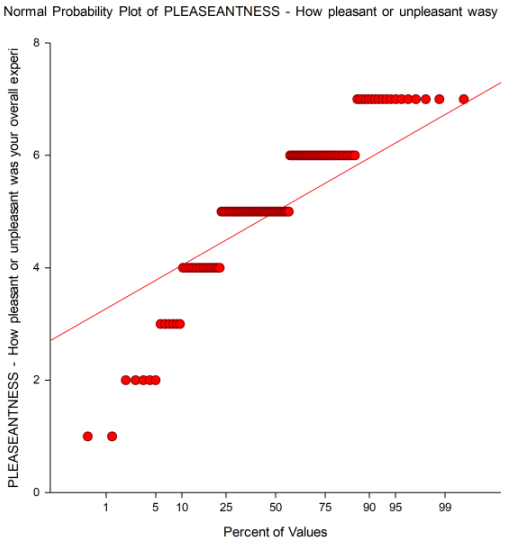
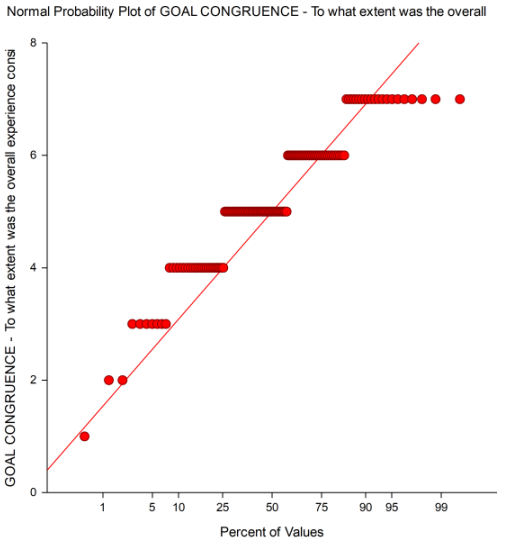

Normal Probability Plot of Feedback/Guided Reflection - The simulation allowed me
Feedback/Guided Reflection - The simulation allowed me to analyze my o
6
5.5
5
4.5
4
3.5
3
1
5
10
25
50
75
90
95
99
Percent of Values

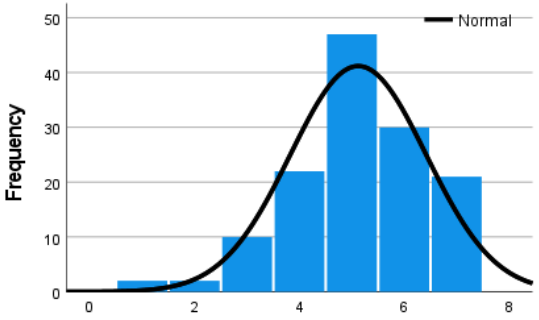
Q6.4_4	

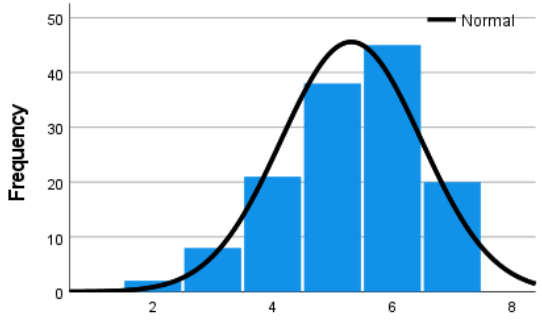

Normal
Frequency
100
80
60
40
20
0
2
3
4
5
6
7

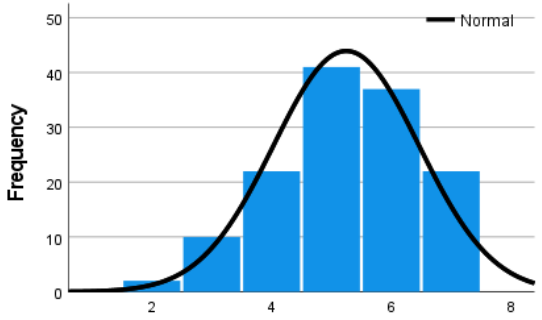

Normal Probability Plot of Feedback/Guided Reflection - There was an opportunitya
Feedback/Guided Reflection - There was an opportunity after the simulati
6
5.5
5
4.5
4
3.5
3
1
5
10
25
50
75
90
95
99
Percent of Values

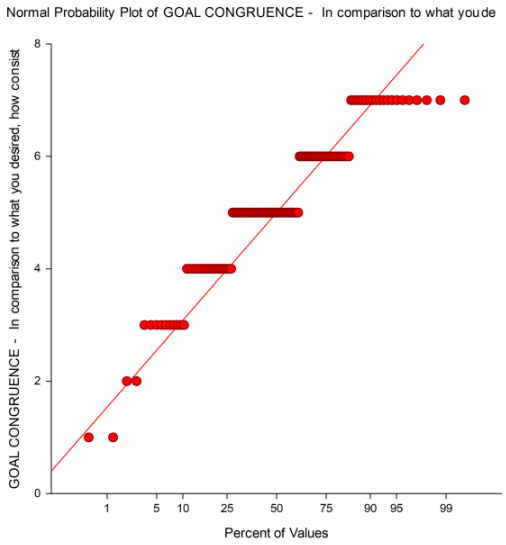
Emotional Appraisal Process (Post-Simulation)

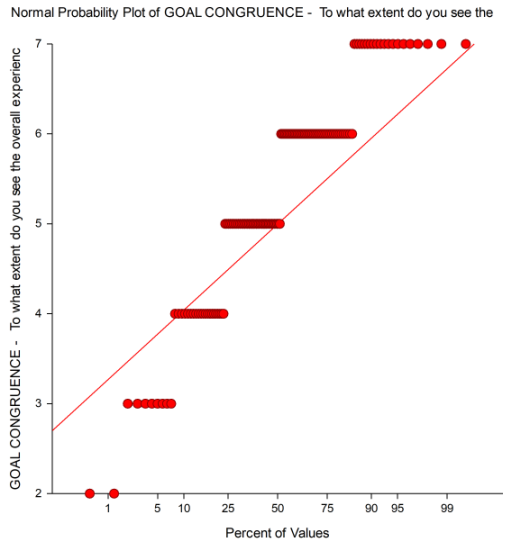
Q4.7_1	Q4.7_2	Q4.7_3
		
		
Q4.7_4	Q4.7_5	Q4.7_6

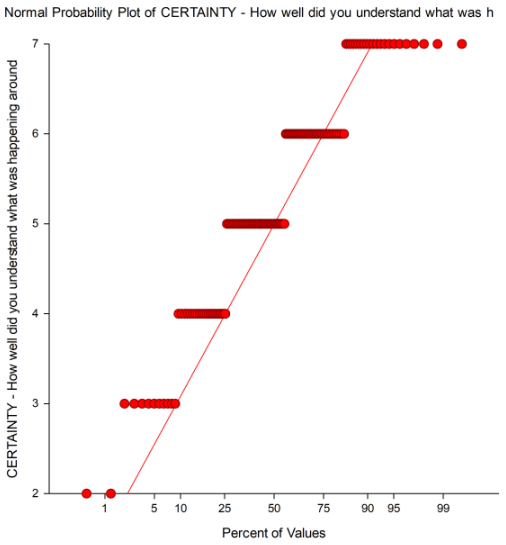




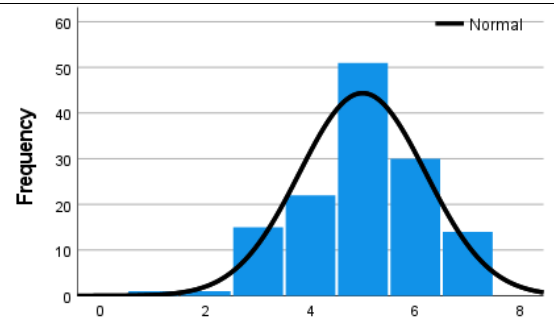
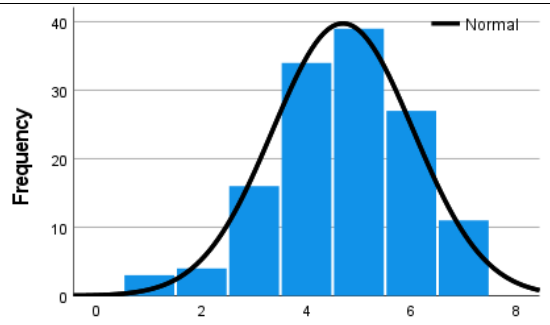
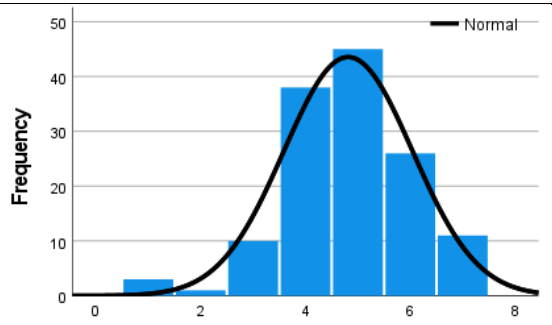
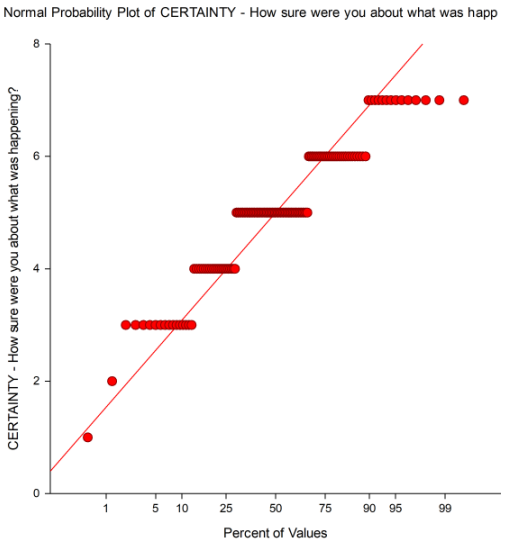
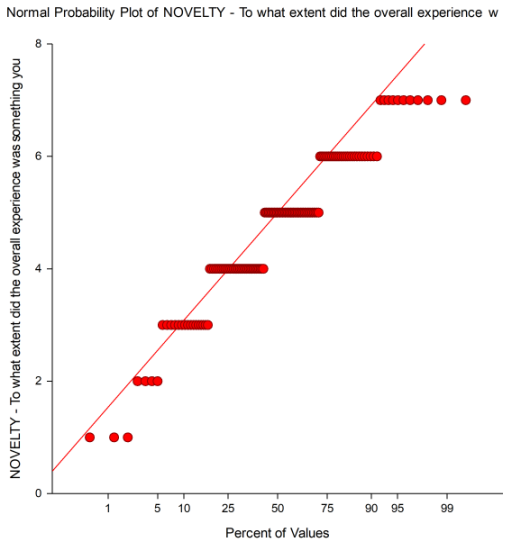
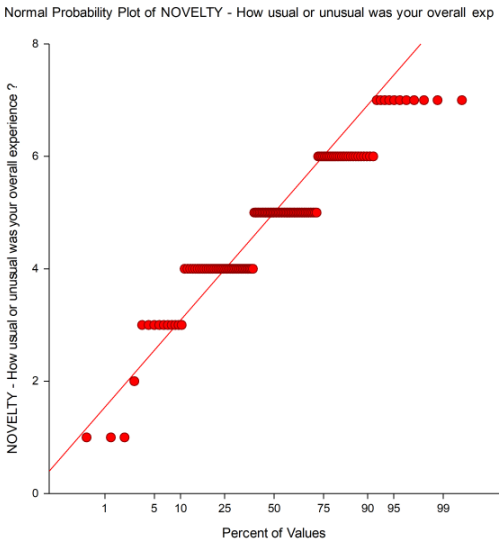


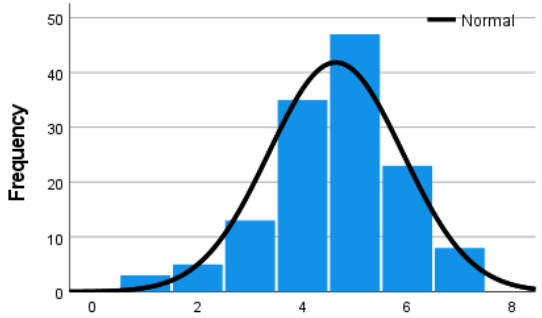


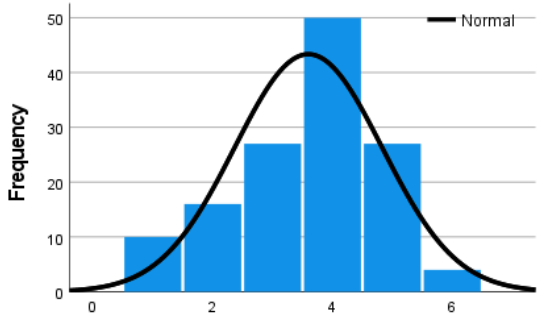


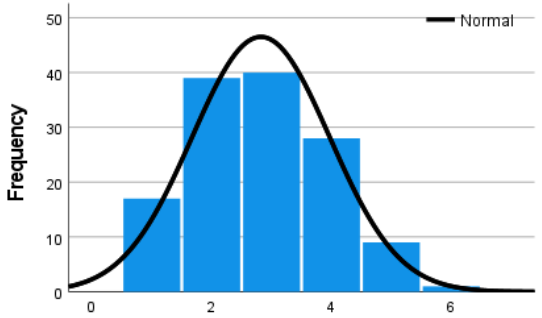


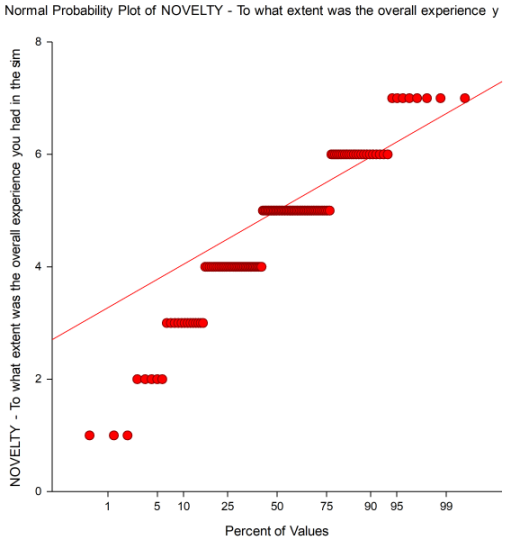
Q4.7_7	Q4.7_8	Q4.7_9

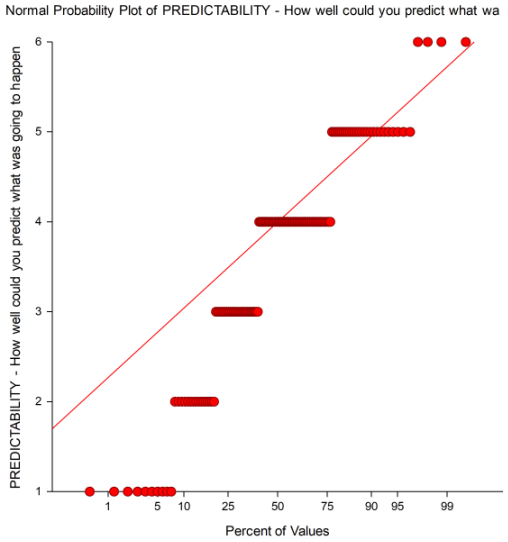
		
		
Q4.7_10	Q4.8_1	Q4.8_2

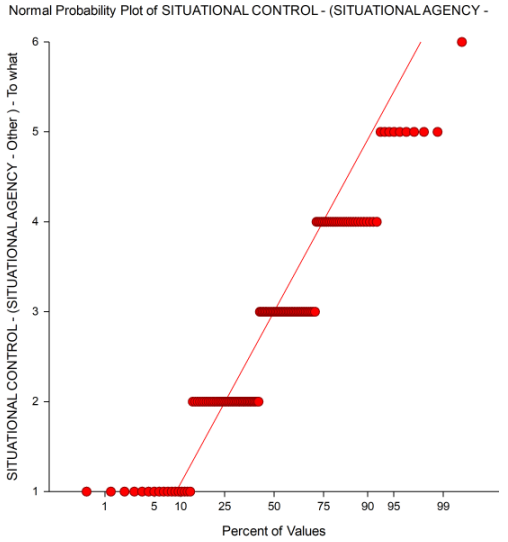




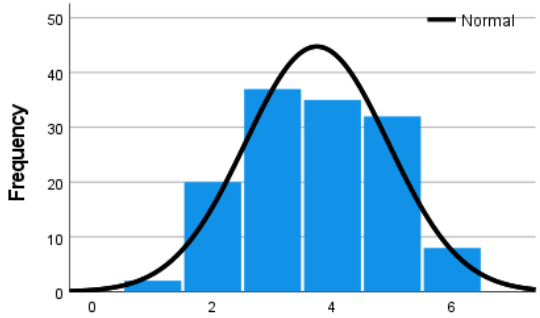
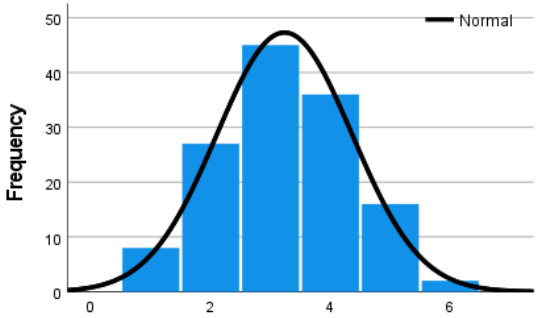
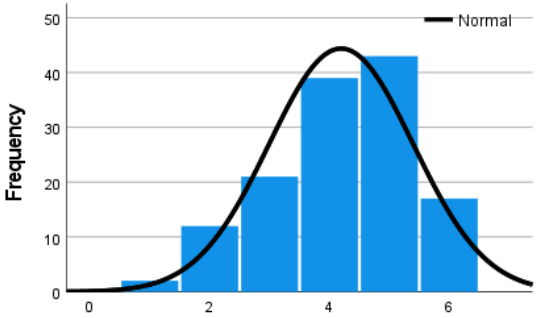
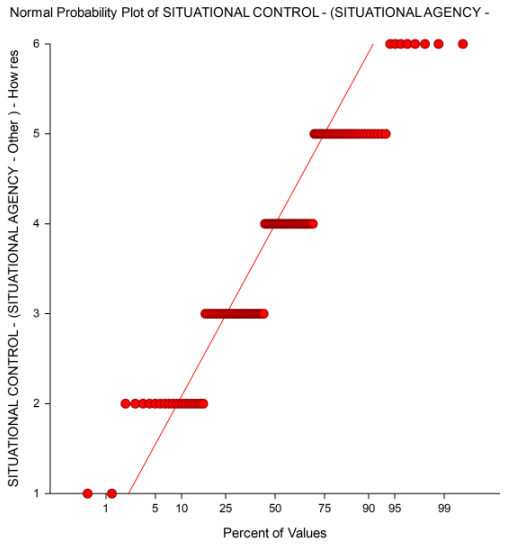
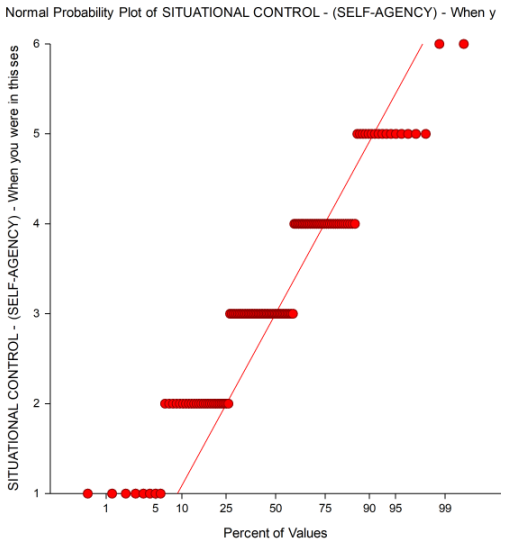
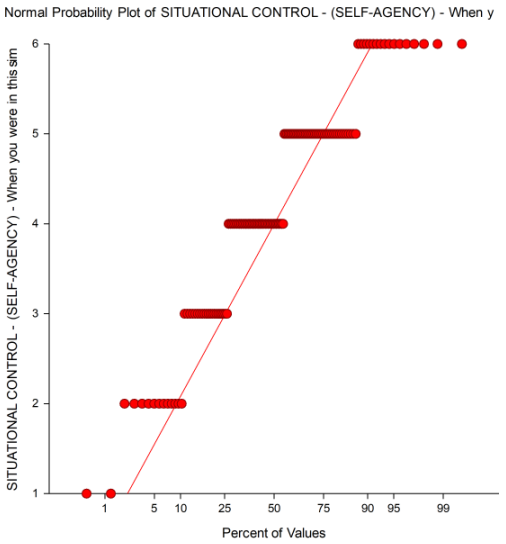


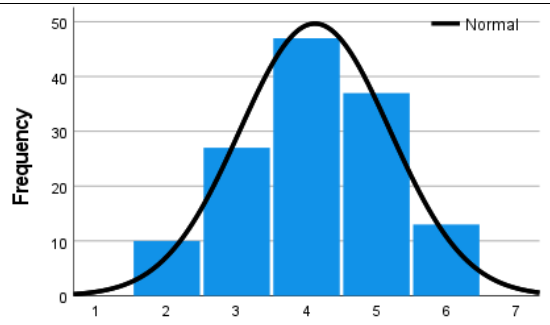
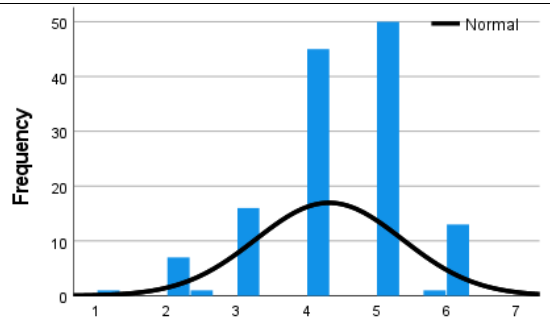
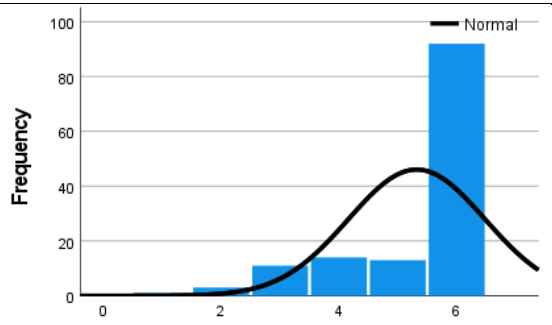
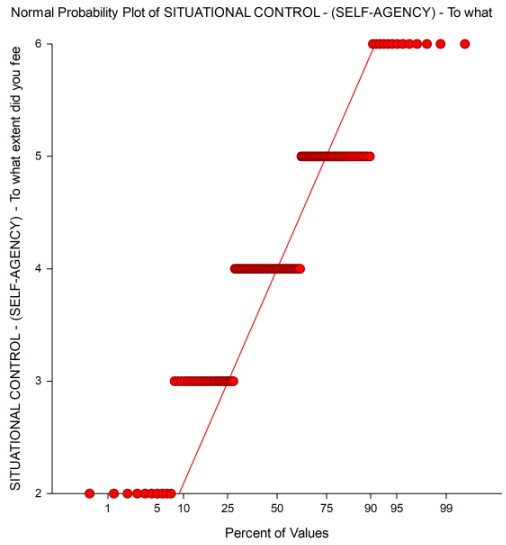
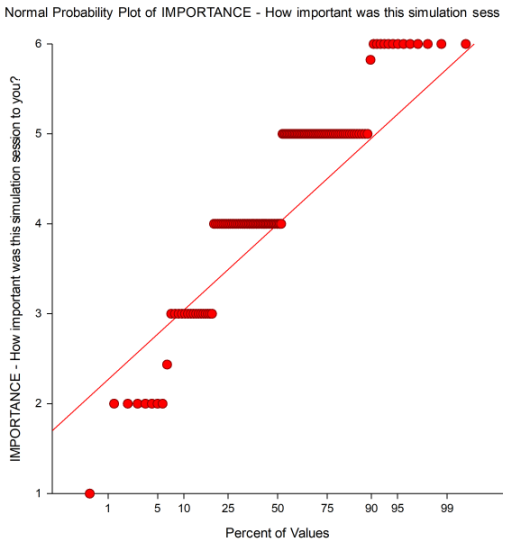
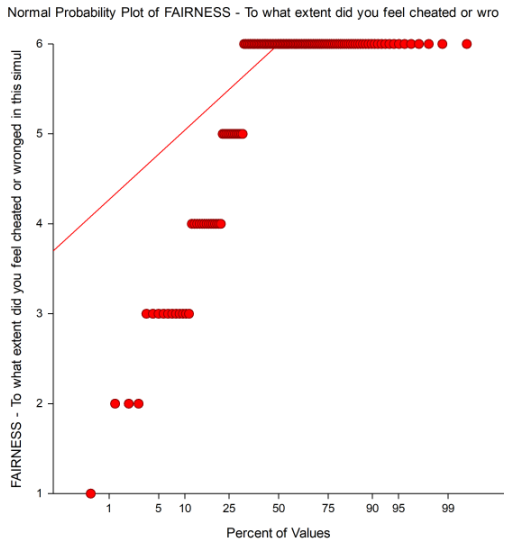


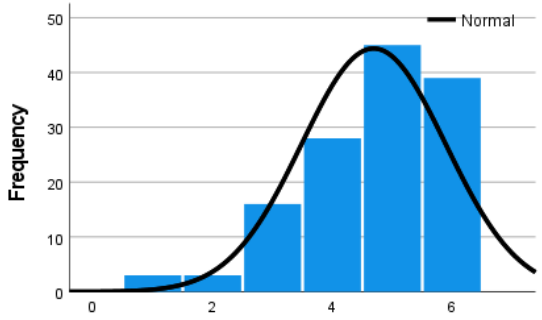


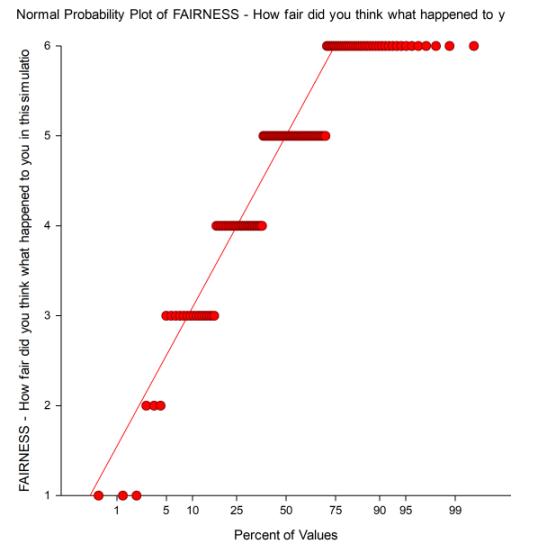


Q4.8_3	Q4.8_4	Q4.8_5_I

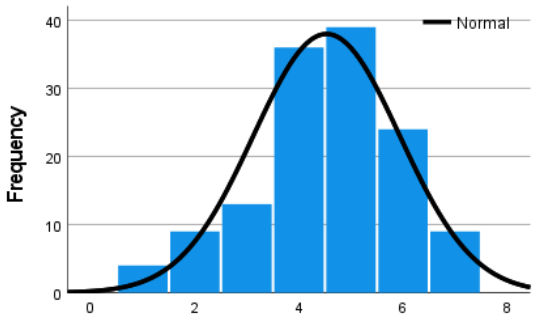
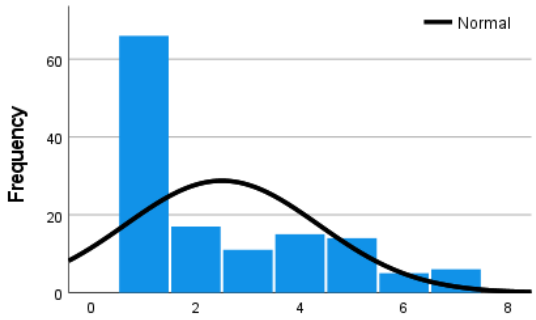
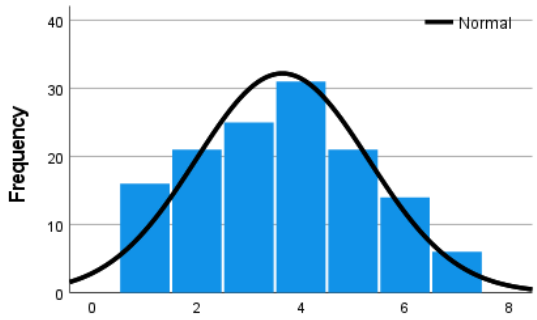
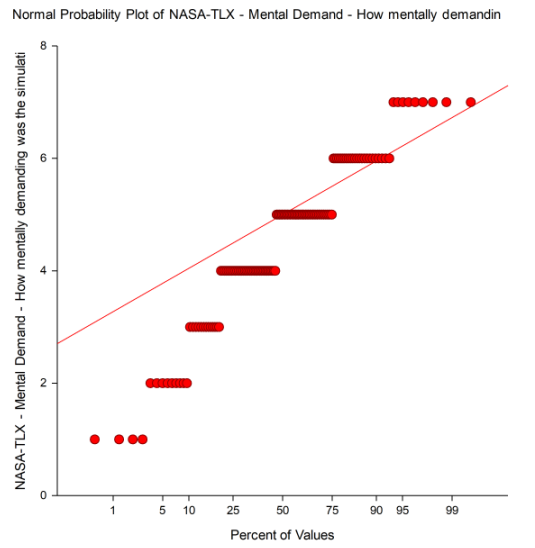
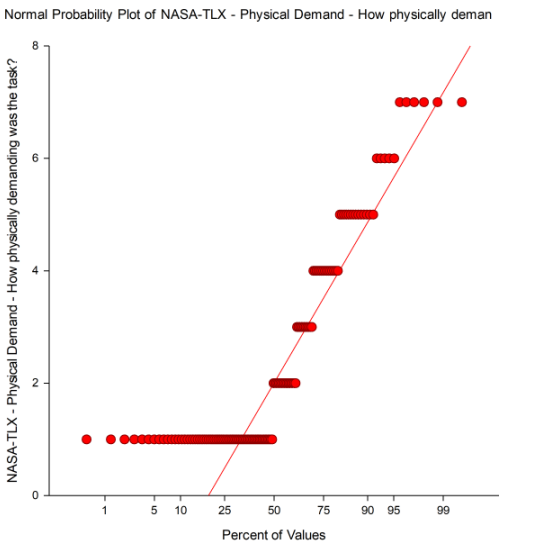
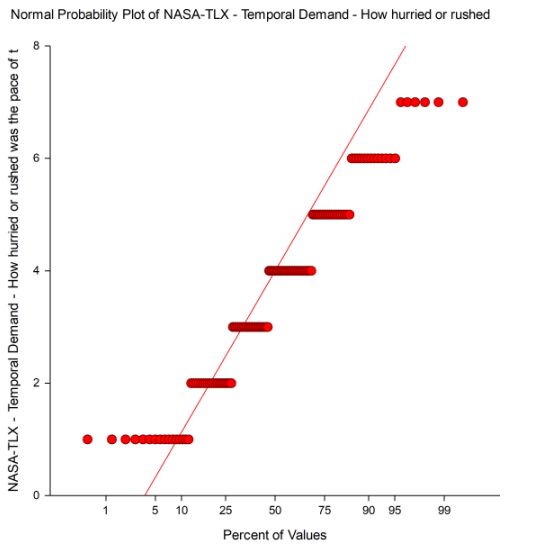
		
		
Q4.8_6	Q4.8_7	Q4.8_8

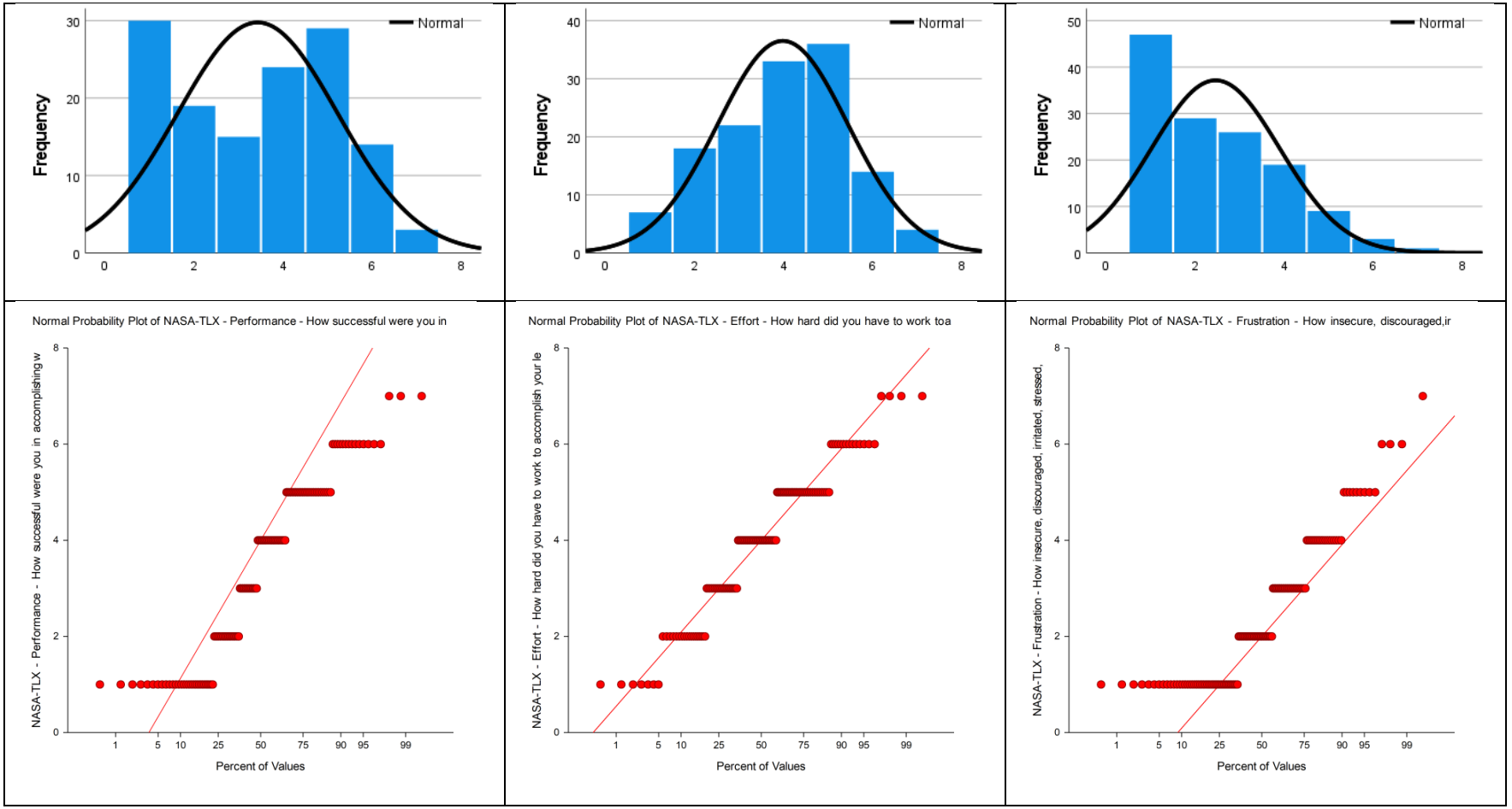
 Frequency Normal 50 40 30 20 10 0 1 2 3 4 5 6 7	 Frequency Normal 50 40 30 20 10 0 1 2 3 4 5 6 7	 Frequency Normal 100 80 60 40 20 0 0 2 4 6
 Normal Probability Plot of SITUATIONAL CONTROL - (SELF-AGENCY) - To what SITUATIONAL CONTROL - (SELF-AGENCY) - To what extent did you fee 6 5 4 3 2 1 5 10 25 50 75 90 95 99 Percent of Values	 Normal Probability Plot of IMPORTANCE - How important was this simulation sess IMPORTANCE - How important was this simulation session to you? 6 5 4 3 2 1 1 5 10 25 50 75 90 95 99 Percent of Values	 Normal Probability Plot of FAIRNESS - To what extent did you feel cheated or wro FAIRNESS - To what extent did you feel cheated or wronged in this simul 6 5 4 3 2 1 1 5 10 25 50 75 90 95 99 Percent of Values
Q4.8_9		


Normal
Frequency
50
40
30
20
10
0
0
2
4
6


Normal Probability Plot of FAIRNESS - How fair did you think what happened to y
FAIRNESS - How fair did you think what happened to you in this simulatio
6
5
4
3
2
1
1
5
10
25
50
75
90
95
99
Percent of Values

NASA-TLX (Post-Simulation)

Q4.9_1	Q4.9_2	Q4.9_3
		
		
Q4.9_4	Q4.9_5	Q4.9_6


Normal
Frequency
Normal Probability Plot of NASA-TLX - Performance - How successful were you in
NASA-TLX - Performance - How successful were you in accomplishing w
Percent of Values
Normal Probability Plot of NASA-TLX - Effort - How hard did you have to work toa
NASA-TLX - Effort - How hard did you have to work to accomplish your le
Percent of Values
Normal Probability Plot of NASA-TLX - Frustration - How insecure, discouraged,ir
NASA-TLX - Frustration - How insecure, discouraged, irritated, stressed,
Percent of Values

Appendix 3: Frequency distribution and Skewness and Kurtosis analysis of study variables

Emotion State Variables (Baseline, Post-Simulation, and Post-Debriefing)

Variable	Description	N	Mean	Std. Deviation	Skewness	Std. Error of Skewness	Kurtosis	Std. Error of Kurtosis
Q2.7_1	(Baseline) Estimate here how aroused you feel. - Low Arousal:High Arousal	134	3.33	1.122	-0.323	0.209	-0.479	0.416
Q2.8_1	(baseline) how you feel now - Bored:Alert	134	0.41	0.851	-0.309	0.209	0.325	0.416
Q2.8_2	(baseline) how you feel now - Lethargic:Excited	134	0.24	0.796	-0.185	0.209	0.498	0.416
Q2.8_3	(baseline) how you feel now - Depressed:Elated	134	0.24	0.551	0.88	0.209	1.318	0.416
Q2.8_4	(baseline) how you feel now - Tense:Calm	134	0.3	1.033	-0.088	0.209	-0.53	0.416
Q2.8_5	(baseline) how you feel now - Sad:Happy	134	0.51	0.723	-0.174	0.209	1.163	0.416
Q2.8_6	(baseline) how you feel now - Upset:Contended	134	0.6	0.823	-0.053	0.209	0.314	0.416
Q2.8_7	(baseline) how you feel now - Stressed:Serene	134	0.05	0.952	-0.105	0.209	-0.301	0.416
Q2.8_8	(baseline) how you feel now - Nervous:Relaxed	134	0.1	0.991	0.164	0.209	-0.565	0.416
Q4.5_1	(PS) Estimate here how aroused you feel. - Low Arousal:High Arousal	134	4.29	1.188	-0.693	0.209	0.408	0.416
Q4.6_1	(PS) how you felt during the simulation - Bored:Alert	134	1.2	0.783	-0.849	0.209	0.988	0.416
Q4.6_2	(PS) how you felt during the simulation - Lethargic:Excited	134	0.93	0.728	-0.491	0.209	1.043	0.416
Q4.6_3	(PS) how you felt during the simulation - Depressed:Elated	134	0.47	0.634	-0.247	0.209	0.892	0.416
Q4.6_4	(PS) how you felt during the simulation - Tense:Calm	134	-0.13	0.992	0.165	0.209	-0.444	0.416
Q4.6_5	(PS) how you felt during the simulation - Sad:Happy	134	0.51	0.743	-0.249	0.209	0.969	0.416
Q4.6_6	(PS) how you felt during the simulation - Upset:Contended	134	0.57	0.74	-0.235	0.209	1.062	0.416
Q4.6_7	(PS) how you felt during the simulation - Stressed:Serene	134	-0.15	0.914	0.362	0.209	0.055	0.416
Q4.6_8	(PS) how you felt during the simulation - Nervous:Relaxed	134	-0.16	0.925	0.145	0.209	-0.211	0.416
Q6.2_1	(PD) Estimate here how aroused you feel - Low Arousal:High Arousal	134	3.67	1.188	-0.212	0.209	-0.339	0.416
Q6.3_1	(PD) how you felt during the debriefing - Bored:Alert	134	0.8	0.802	-0.328	0.209	0.201	0.416

Q6.3_2	(PD) how you felt during the debriefing - Lethargic:Excited	134	0.6	0.727	0.073	0.209	0.371	0.416
Q6.3_3	(PD) how you felt during the debriefing - Depressed:Elated	134	0.43	0.605	0.498	0.209	-0.08	0.416
Q6.3_4	(PD) how you felt during the debriefing - Tense:Calm	134	0.58	0.87	-0.117	0.209	-0.627	0.416
Q6.3_5	(PD) how you felt during the debriefing - Sad:Happy	134	0.7	0.736	0.193	0.209	-0.591	0.416
Q6.3_6	(PD) how you felt during the debriefing - Upset:Contended	134	0.75	0.74	0.204	0.209	-0.746	0.416
Q6.3_7	(PD) how you felt during the debriefing - Stressed:Serene	134	0.37	0.846	0.355	0.209	-0.404	0.416
Q6.3_8	(PD) how you felt during the debriefing - Nervous:Relaxed	134	0.52	0.838	0.006	0.209	-0.55	0.416

Feedback and Perception OF Presence Variables (Post-Simulation, and Post-Debriefing)

	Description	N	Mean	Std. Deviation	Skewness	Std. Error of Skewness	Kurtosis	Std. Error of Kurtosis
Q4.3_1	Perception of Fidelity (Realism) of the simulation - The scenario resembled a real-life situation	134	4.98	1.058	-1.578	0.209	3.597	0.416
Q4.3_2	Perception of Fidelity (Realism) of the simulation - Real life factors, situation, and variables were built into the simulation scenario	134	5.149	0.8272	-1.257	0.209	3.673	0.416
Q4.4_1	Rate your level of presence	134	67.7239	22.7743	-0.779	0.209	0.128	0.416
Q6.4_1	Feedback/Guided Reflection - Feedback provided has constructive	134	5.59	0.707	-1.945	0.209	3.818	0.416
Q6.4_2	Feedback/Guided Reflection - Feedback was provided in a timely manner	134	5.64	0.654	-2.264	0.209	5.961	0.416
Q6.4_3	Feedback/Guided Reflection - The simulation allowed me to analyze my own behavior and actions	134	5.42	0.807	-1.342	0.209	1.19	0.416
Q6.4_4	Feedback/Guided Reflection - There was an opportunity after the simulation to obtain guidance/feedback from the teacher in order to build knowledge to another level	134	5.63	0.62	-2.065	0.209	5.431	0.416

Emotional Appraisal Process (Post-Simulation)

Variable	Description	N	Mean	Std. Deviation	Skewness	Std. Error of Skewness	Kurtosis	Std. Error of Kurtosis
Q4.7_1	PLEASEANTNESS - How enjoyable was your overall experience ?	134	5.31	1.192	-0.738	0.209	1.077	0.416
Q4.7_2	PLEASEANTNESS - How pleasant or unpleasant was your overall experience ?	134	5.17	1.301	-0.907	0.209	1.043	0.416
Q4.7_3	GOAL CONGRUENCE - To what extent was the overall experience consistent to what you wanted to achieve during this session?	134	5.22	1.222	-0.524	0.209	0.375	0.416
Q4.7_4	GOAL CONGRUENCE - In comparison to what you desired, how consistent was your overall experience ?	134	5.12	1.298	-0.581	0.209	0.484	0.416
Q4.7_5	GOAL CONGRUENCE - To what extent do you see the overall experience in this session important or not important in achieving your learning goals or needs?	134	5.31	1.172	-0.521	0.209	-0.135	0.416
Q4.7_6	CERTAINTY - How well did you understand what was happening around you?	134	5.25	1.217	-0.383	0.209	-0.377	0.416
Q4.7_7	CERTAINTY - How sure were you about what was happening?	134	4.99	1.205	-0.378	0.209	0.108	0.416
Q4.7_8	NOVELTY - To what extent did the overall experience was something you expected or unexpected?	134	4.69	1.345	-0.384	0.209	0.057	0.416
Q4.7_9	NOVELTY - How usual or unusual was your overall experience ?	134	4.81	1.227	-0.454	0.209	0.831	0.416
Q4.7_10	NOVELTY - To what extent was the overall experience you had in the simulation session familiar or unfamiliar ?	134	4.63	1.278	-0.492	0.209	0.456	0.416
Q4.8_1	PREDICTABILITY - How well could you predict what was going to happen?	134	3.6	1.233	-0.459	0.209	-0.342	0.416
Q4.8_2	SITUATIONAL CONTROL - (SITUATIONAL AGENCY - Other) - To what extent did you feel that circumstances beyond	134	2.82	1.149	0.237	0.209	-0.528	0.416

	anyone's control determined the outcomes/results of this simulation?							
Q4.8_3	SITUATIONAL CONTROL - (SITUATIONAL AGENCY - Other) - How responsible did you think someone or something other than yourself was for the outcomes/results of the simulation session?	134	3.74	1.195	-0.017	0.209	-0.766	0.416
Q4.8_4	SITUATIONAL CONTROL - (SELF-AGENCY) - When you were in this session, how powerful (dominant) did you feel?	134	3.23	1.13	0.04	0.209	-0.429	0.416
Q4.8_5_I	SITUATIONAL CONTROL - (SELF-AGENCY) - When you were in this simulation session, how helpless did you feel? (Inverted)	134	4.19	1.204	-0.46	0.209	-0.348	0.416
Q4.8_6	SITUATIONAL CONTROL - (SELF-AGENCY) - To what extent did you feel that you could influence what was happening in the simulation session?	134	4.12	1.076	-0.131	0.209	-0.555	0.416
Q4.8_7	IMPORTANCE - How important was this simulation session to you?	134	4.32	1.051	-0.589	0.209	0.2	0.416
Q4.8_8	FAIRNESS - To what extent did you feel cheated or wronged in this simulation session ? (inverted)	134	5.32	1.161	-1.62	0.209	1.675	0.416
Q4.8_9	FAIRNESS - How fair did you think what happened to you in this simulation session was?	134	4.69	1.204	-0.894	0.209	0.533	0.416

NASA-TLX (Post-Simulation)

Variable	Description	N	Mean	Std. Deviation	Skewness	Std. Error of Skewness	Kurtosis	Std. Error of Kurtosis
Q4.9_1	NASA-TLX - Mental Demand - How mentally demanding was the simulation?	134	4.53	1.407	-0.432	0.209	-0.035	0.416
Q4.9_2	NASA-TLX - Physical Demand - How physically demanding was the task?	134	2.5	1.859	0.959	0.209	-0.321	0.416
Q4.9_3	NASA-TLX - Temporal Demand - How hurried or rushed was the pace of the simulation?	134	3.64	1.661	0.117	0.209	-0.789	0.416
Q4.9_4	NASA-TLX - Performance - How successful were you in accomplishing what you were asked to do? (I)	134	3.43	1.795	0.013	0.209	-1.25	0.416
Q4.9_5	NASA-TLX - Effort - How hard did you have to work to accomplish your level of performance?	134	3.98	1.464	-0.194	0.209	-0.579	0.416
Q4.9_6	NASA-TLX - Frustration - How insecure, discouraged, irritated, stressed, and annoyed were you?	134	2.46	1.439	0.776	0.209	-0.13	0.416

Appendix 4: Variables Transformation

```
COMPUTE Cheerfullness_baseline=(Q2.8_6_T * 0.442)+ (Q2.8_3_T*0.366) +
(Q2.8_5_T * 0.386).
EXECUTE.
DATASET ACTIVATE DataSet9.
DATASET CLOSE DataSet8.
COMPUTE Cheerfullness_baseline=(Q2.8_6_T * 0.442)+ (Q2.8_3_T*0.366) +
(Q2.8_5_T * 0.386).
EXECUTE.
COMPUTE  Cheerfullness_ps=(Q4.6_6_T  *  0.442)+  (Q4.6_3_T*0.366)  +
(Q4.6_5_T * 0.386).
EXECUTE.
COMPUTE  Cheerfullness_pd=(Q6.3_6_T  *  0.442)+  (Q6.3_3_T*0.366)  +
(Q6.3_5_T * 0.386).
EXECUTE.
COMPUTE Invigoration_baseline=(Q2.8_1_T*0.585) + (Q2.8_2_T* 0.509).
EXECUTE.
COMPUTE Invigoration_ps=(Q4.6_1_T *0.585) + (Q4.6_2_T* 0.509).
EXECUTE.
COMPUTE Invigoration_pd=(Q6.3_1_T *0.585) + (Q6.3_2_T* 0.509).
EXECUTE.
COMPUTE  tranquility_baseline=(Q2.8_4_T*0.369)  +  (Q2.8_7_T*0.360)  +
(Q2.8_8_T*0.383).
EXECUTE.
COMPUTE   tranquility_ps=(Q4.6_4_T*0.369)   +   (Q4.6_7_T*0.360)   +
(Q4.6_8_T*0.383).
EXECUTE.
COMPUTE   tranquility_pd=(Q6.3_4_T*0.369)   +   (Q6.3_7_T*0.360)   +
(Q6.3_8_T*0.383).
EXECUTE.
COMPUTE Pleasantness=(Q4.7_1 * 0.545) + (Q4.7_2 * 0.51).
EXECUTE.
COMPUTE Goal_Congruence=(0.365*Q4.7_3) + (0.355*Q4.7_4)+(Q4.7_5*0.367).
EXECUTE.
COMPUTE Certainty=(Q4.7_6 * 0.496) + (Q4.7_7* 0.571).
EXECUTE.
COMPUTE Novelty=(Q4.7_8_I*0.426)+ (Q4.7_9_I*0.356) + (Q4.7_10_I*0.388).
EXECUTE.
COMPUTE Predictatibility=Q4.8_1.
EXECUTE.
COMPUTE Situational_Control=(Q4.8_4*0.467) + (Q4.8_6*0.689).
EXECUTE.
COMPUTE Importance=Q4.8_7.
EXECUTE.
COMPUTE Fidelity=(Q4.3_1*0.563)+ (Q4.3_2 * 0.511).
EXECUTE.
COMPUTE Feedback=(Q6.4_1 * 0.376) + (Q6.4_2 * 0.242) + (Q6.4_3 * 0.267)
+ (Q6.4_4*0.207).
EXECUTE.
COMPUTE Nasa_RawTLX=Q4.9_1+Q4.9_2+Q4.9_3+Q4.9_4+Q4.9_5+Q4.9_6.
EXECUTE.
COMPUTE Nasa_RawTLX_NoPhysical=Q4.9_1+Q4.9_3+Q4.9_4+Q4.9_5+Q4.9_6.
EXECUTE.
COMPUTE Nasa_Task_Demand=Q4.9_1+Q4.9_2+Q4.9_3.
EXECUTE.
COMPUTE Nasa_Task_Demand_noPhysical=Q4.9_1+Q4.9_3.
EXECUTE.
COMPUTE Nasa_Task_Subject_Demand=Q4.9_4+Q4.9_5+Q4.9_6.
EXECUTE.
COMPUTE FAS_Simulation=Q4.5_1-Q2.7_1.
```

```
EXECUTE.
COMPUTE FAS_Debriefing=Q6.2_1-Q4.5_1.
EXECUTE.
```