University of South Wales
Prifysgol De Cymru

Metrics for Evaluating Cyber Security Data Visualizations in Virtual Reality

Daniel Harris, Marius Miknis, Connor Smith, Ian Wilson

 # July 2021 #

kess2 dwyrain east
Ysgoloriaethau Sgiliau Economi Gwybodaeth
Knowledge Economy Skills Scholarships

UNDEB EWROPEAIDD
EUROPEAN UNION
Llywodraeth Cymru
Welsh Government
Cronfa Gymdeithasol Ewrop
European Social Fund

Abstract

Write last

1 Introduction

Evaluation in the areas of cyber security, data visualisation, and virtual reality have historically been difficult and combining these areas has compounded the issue. Evaluation issues in these areas are presented in this section.

1.1 Evaluation Difficulties

There is a lack of evaluation for these types of visualisations which has led to unproven claims of effectiveness. Researchers acknowledge the lack of testing [1]–[3], and plan to rectify this in the future by conducting user-based evaluations. However, none of these researchers are yet to do so. The reason these inspiring researchers have not evaluated their systems is due to a lack of guidance on how they should be performed [4]–[7]. This paper proposes a solution to this lack of guidance by producing a list of evaluation metrics to inform the development of more effective and more regular evaluations in the future.

1.2 Performance vs. Preference

Papers that do include evaluations use subjective (non-observational) participant feedback to prove effectiveness rather than using objective (observational) analysis [8]–[10]. An example of this could be asking a participant which system they thought was quicker rather than timing them. This style of subjective evaluation is discouraged by other researchers who explain that participants can prefer systems that actually perform worse, due to aesthetics, novelty, and familiarity [11]–[14]. When using non-observational metrics to evaluate a system you are not investigating the effectiveness of a system but instead investigating a participant's opinions of its effectiveness. Non-observational metrics can give valuable insights into user satisfaction [15], however they should not be used alone to measure the effectiveness of a system when observational metrics are available.

Understanding the availability of observational and non-observational metrics is difficult due to the infancy of and lack of literature concerned with this combined field. This paper resolves this issue by surveying and dividing existing evaluation metrics into observational and non-observational categories that will better inform future research and produce more accurate assessments of system effectiveness.

Drew et al – Need to read to double check *NOW THAT* I have access

2 Related Work

Virtual reality visualisations cannot be evaluated with traditional metrics. Cyber security data visualisation evaluations have expectations that are not applicable when viewing data in virtual reality # ref – repurpose cyber refs from para below? #, most notably:

1. The data will be viewed on 2-dimentional displays.
2. The data will be viewed from a single perspective.
3. The data will be displayed with a consistent background.
4. Interaction methods will involve a mouse, trackpad, keyboard, or touch screen.
5. Users will have experience with input methods and output formats.

Virtual reality visualisations cannot be evaluated with traditional metrics and therefor research needs to be conducted into metrics specific to this combined topic area. Gathering and analysing evaluation metrics is common practice in cyber security, data visualisation, and virtual reality [9], [16] # add more refs, see Notion - Related Work #. However, there has yet to be a synthesis of metrics for works overlapping in these topics as evidenced in section 1.1.

why is the below important?

Past works that survey evaluation metrics have also identified the importance of separating metrics into observational and non-observational categories [9], [16] # <- Maybe add more if I can #. However, none of these papers simultaneously surveys metrics from all three topics of cyber security, data visualisation, and virtual reality.

2.1 Other efforts

Other researchers have identified the lack of guidance in this area and produce solutions in the form of defining tasks that should be conducted during an evaluation [14], [17]–[20]. However, there is no explanation of how the tasks should be used to gauge effectiveness. This paper supplies this missing information by documenting the metrics that can be used when evaluating task effectiveness.

2.1.1 Non-user evaluations

~ An evaluation framework for network security visualizations

- list of metrics of how judge cyber security data viz *Programs*

3 Methods

User-based evaluation metrics were researched from the three overlapping topics of cyber security, data visualisation and virtual reality (Figure 1). Metrics applicable to the overlap of all three topic areas were documented. The methods used to find and document these metrics is discussed in this section.

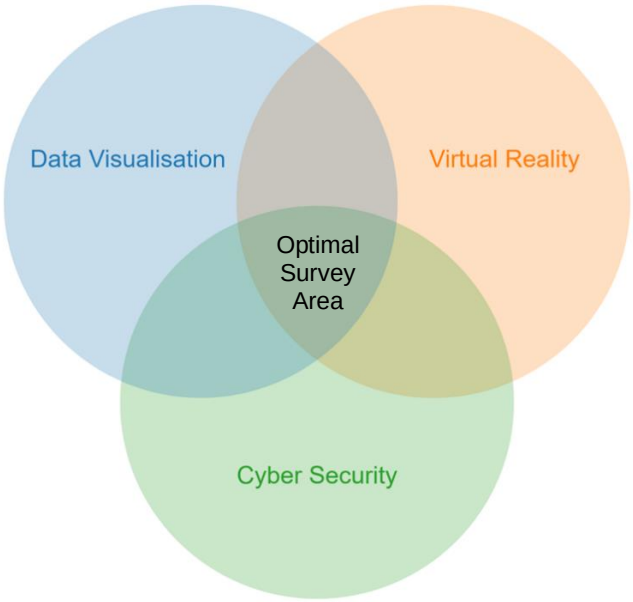

Figure 1 - Venn diagram of survey topic areas

3.1 Search method

Due to the small number of papers specifically concerned with all three topic areas this research also investigated any paper residing in one of the topic areas that involved user-based evaluation. Priority was given to papers that overlap two or more of the topic areas, however these were scarce due to the difficulties discussed in section 1.1. The search for new papers continued until no new meaningful metric was discovered after a considerable length of time. The term “Virtual Reality” in the search context also includes augmented reality, extended reality, and digital stereoscopy. The full list of topic areas surveyed for user-based evaluation metrics is shown in Figure 2 and is grouped into areas of survey priority (1 = highest, 3 = lowest).

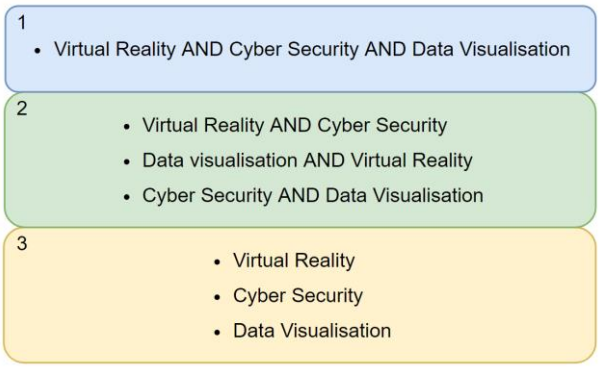

Figure 2 - Grouped list of surveyed topic areas

The search terms were used with a range of databases, most notably:

- IEEE
- Association for Computer Machinery
- Semantic Scholar
- Springer
- ScienceDirect

3.2 Metric categories

The discovered metrics were divided into the two categories of observational and non-observational. Observational metrics are gathered passively from participants solely through observing and/or recording their actions and should be gathered without a participant's active awareness. An example of an observational metric would be the length of time it takes a participant to perform a given task. The observational metrics were further subdivided into four sub-categories:

1. *Precision*. The accuracy of participants at performing tasks and discovering insights.
2. *Period*. The measurement of time taken to perform actions during the experiment.
3. *Progression.* Tracking the physical movement of a participant during the experiment.
4. *Portion.* Relating to counting the occurrences of actions or outcomes.

Non-observational metrics are gathered actively from participants by interacting with them for the purpose of data gathering. An example of a non-observational metric would be an interview where participants are consciously aware that they are providing feedback. The non-observational metrics were further subdivided into four sub-categories:

1. *Procurement*. Defines a subjective method of gathering information from participants without explicitly defining the question or discussion topic used. These should not be considered as standalone evaluation metrics but instead be used as a method to gather other non-observational metrics.
2. *Physiological*. Relating to a participant's physiological response to the experiment. Examples include nausea and fatigue.
3. *Psychological*. Relating to a participant's emotional or mental response to the experiment and attempts to gauge the subconscious impact of an experiment.
4. *Personal*. Feedback given by participants portraying their opinions of aspects of the experiment.

3.3 Distribution

Papers were surveyed as evenly as practicable between the three topic areas. As shown in Figure 3 there were less cyber security papers surveyed compared to the other areas. This

was due to a lack of relevant user-based evaluation papers in this area. Although the reason for this absence of papers cannot be unequivocally identified it is likely due to the added difficulty in obtaining access to SOCs, security analysts, and security datasets. #find reference for this#

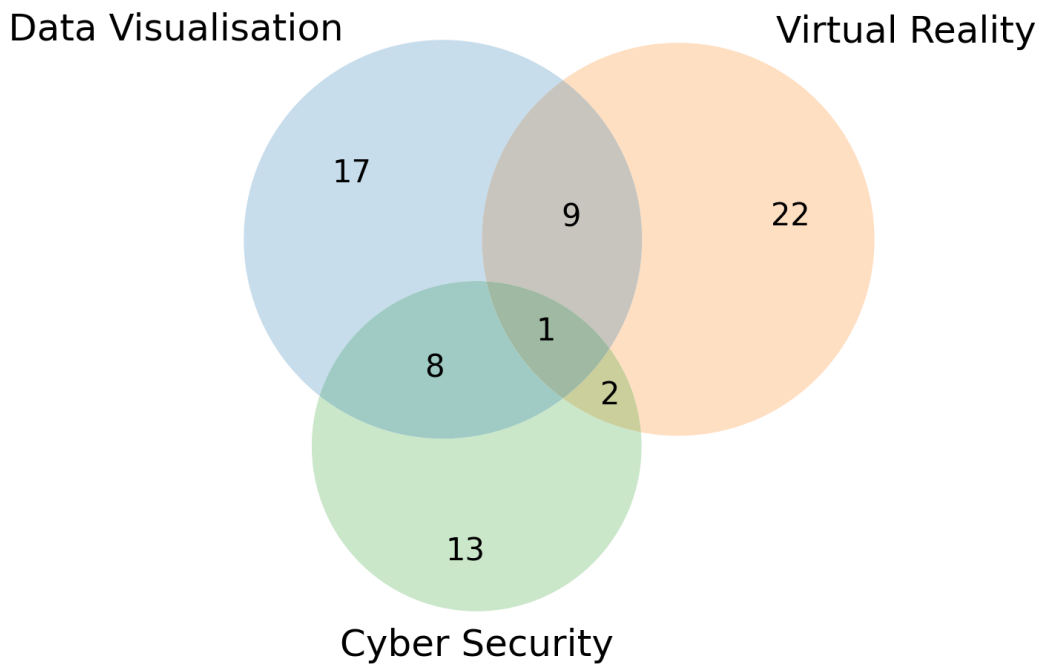

Figure 3 - Distribution of surveyed papers between the three topic areas

4 Results

Table 1 - Non-Observational Metric Data

Papers Surveyed	Non-Observational Metrics																																
	Questionnaire	Usability	Interview	Effectiveness/Usefulness	Perceived accuracy	Performance demand/Mental effort	Perceived Learning Effectiveness	Preferred	Immersion/Presence	Intention to use system	Satisfaction	Physical demand/Fatigue	Thinking aloud	Post task test	Frustration	Stimulation/Motivation	Nausea	User Experience Questionnaire	Perceived speed	Intuitiveness	Reality	Attractiveness	Novelty	Problem–solving capability	Differential emotion	Comfort	Dizziness	Cooperation	Score of pre task test	Temporal demand	Situation Awareness	Familiarity
Shared Surfaces and Spaces: Collaborative Data Visualisation in a Co-located Immersive Environment																																	
Design and Evaluation of Interactive Small Multiples Data Visualisation in Immersive Spaces	■																																
Immersion or diversion: Does virtual reality make data visualisation more effective?							■								■																		
Exploring Data in Virtual Reality: Comparisons with 2D Data Visualizations						■					■	■	■																				
How do new visual immersive systems influence gaming QoE? A use case of serious gaming with Oculus Rift	■	■							■																■								
A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods	■	■	■																														
Evaluation of a Virtual Reality Nasal Cavity Education Tool	■													■														■					
An Evaluation of the Effectiveness of Virtual Reality in Air Traffic Control	■						■																										
Evaluating Navigation Techniques for 3D Graph Visualizations in Virtual Reality						■	■																										
Operator impressions of 3d visualizations for cybersecurity analysts			■																														
STREAM: Exploring the Combination of Spatially-Aware Tablets with Augmented Reality Head-Mounted Displays for Immersive Analytics			■										■					■														
Evaluation of a Prototype Desktop Virtual Reality Model developed to Enhance Electrical Safety and Design in the Built Environment		■		■			■		■	■	■					■																
Investigating learners' attitudes toward virtual reality learning environments: Based on a constructivist approach									■	■						■								■								
Design and implementation of user experience model for augmented reality systems		■														■						■	■									
Popular Performance Metrics for Evaluation of Interaction in Virtual and Augmented Reality		■			■	■	■	■				■					■		■	■	■					■	■					
A survey of evaluation techniques used in augmented reality studies																																	
Usability Engineering for Augmented Reality: Employing User-based Studies to Inform Design																																	
A Survey of augmented reality		■																															
The relation between visualization size, grouping, and user performance																																	
Implementation and Evaluation of a High-presence Interior Layout Simulation System using Mixed Reality				■					■																								
The Effect of Exploration Mode and Frame of Reference in Immersive Analytics	■	■			■	■		■	■			■			■		■																
SAIBERSOC: Synthetic Attack Injection to Benchmark and Evaluate the Performance of Security Operation Centers	■																																
Time is money! Evaluating Augmented Reality Instructions for Time-Critical Assembly Tasks	■			■	■	■		■				■			■			■															
Flex-ER: A Platform to Evaluate Interaction Techniques for Immersive Visualizations			■	■						■																							
Melbourne 2100: Dystopian Virtual Reality to provoke civic engagement with climate change	■		■											■			■	■											■				
Visualization is Better! A Comparative Evaluation	■	■			■						■		■																				
Evaluation of Visualization Tools for Computer Network Defense Analysts: Display Design, Methods, and Results for a User Study	■	■	■	■	■		■	■		■			■							■		■											
Information visualization metrics and methods for cyber security evaluation	■	■	■					■																									
Capturing Human Cognition in Cyber-Security Simulations with NETS	■				■	■						■			■															■			
Cognitive Task Analysis Based Training for Cyber Situation Awareness	■	■		■										■																			
Advanced visualizations for network security				■									■	■																			
An Interdisciplinary Evaluation of Transactive Memory in Distributed Cyber Teams																					■										■	■	
Comparative Evaluation of Node-Link and Sankey Diagrams for the Cyber Security Domain			■					■			■																						
A Design Methodology and Implementation for Corporate Network Security Visualization: A Modular-Based Approach	■	■		■	■		■			■	■								■														

Table 2 - Observational Metric Data

Papers Surveyed	Observational Metrics																						
	Task accuracy/True Positives	Time taken to perform task	Missed events/False Negatives	Head movement	False positives	Words spoken	Reaction time	True Negatives	Interacted elements	Insights discovered	Body movement	Response time	Learning time	Depth of insights	Breadth of insights	Controller movement	Gaze change	Time looking at objects	Travel distance	Travel time	Steps to recovery	Distance of eye movement
Shared Surfaces and Spaces: Collaborative Data Visualisation in a Co-located Immersive Environment				■					■		■					■						
Design and Evaluation of Interactive Small Multiples Data Visualisation in Immersive Spaces	■	■															■	■	■	■		
Immersion or diversion: Does virtual reality make data visualisation more effective?	■	■																				
Exploring Data in Virtual Reality: Comparisons with 2D Data Visualizations	■									■				■	■							
How do new visual immersive systems influence gaming QoE? A use case of serious gaming with Oculus Rift																						
A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods		■	■																			
Evaluation of a Virtual Reality Nasal Cavity Education Tool													■									
An Evaluation of the Effectiveness of Virtual Reality in Air Traffic Control	■	■	■																			
Evaluating Navigation Techniques for 3D Graph Visualizations in Virtual Reality	■	■		■							■											
Operator impressions of 3d visualizations for cybersecurity analysts																						
STREAM: Exploring the Combination of Spatially-Aware Tablets with Augmented Reality Head-Mounted Displays for Immersive Analytics																						
Evaluation of a Prototype Desktop Virtual Reality Model developed to Enhance Electrical Safety and Design in the Built Environment																						
Investigating learners' attitudes toward virtual reality learning environments: Based on a constructivist approach																						
Design and implementation of user experience model for augmented reality systems																						
Popular Performance Metrics for Evaluation of Interaction in Virtual and Augmented Reality	■	■		■																		
A survey of evaluation techniques used in augmented reality studies	■		■	■																		
Usability Engineering for Augmented Reality: Employing User-based Studies to Inform Design	■	■																				
A Survey of augmented reality		■																				
The relation between visualization size, grouping, and user performance	■	■										■										
Implementation and Evaluation of a High-presence Interior Layout Simulation System using Mixed Reality																						
The Effect of Exploration Mode and Frame of Reference in Immersive Analytics	■	■																				
SAIBERSOC: Synthetic Attack Injection to Benchmark and Evaluate the Performance of Security Operation Centers	■		■																			
Time is money! Evaluating Augmented Reality Instructions for Time-Critical Assembly Tasks	■	■																				
Flex-ER: A Platform to Evaluate Interaction Techniques for Immersive Visualizations																						
Melbourne 2100: Dystopian Virtual Reality to provoke civic engagement with climate change	■																					
Visualization is Better! A Comparative Evaluation	■	■								■												
Evaluation of Visualization Tools for Computer Network Defense Analysts: Display Design, Methods, and Results for a User Study	■	■	■		■			■														
Information visualization metrics and methods for cyber security evaluation	■	■	■		■			■	■	■											■	■
Capturing Human Cognition in Cyber-Security Simulations with NETS	■					■	■															
Cognitive Task Analysis Based Training for Cyber Situation Awareness	■					■	■															
Advanced visualizations for network security	■	■			■				■													
An Interdisciplinary Evaluation of Transactive Memory in Distributed Cyber Teams	■	■	■		■	■	■	■														
Comparative Evaluation of Node-Link and Sankey Diagrams for the Cyber Security Domain	■	■																				
A Design Methodology and Implementation for Corporate Network Security Visualization: A Modular-Based Approach																						

A total of 54 metrics were discovered from 34 papers. Each metric is listed in order of frequency and shown in Table 1, Table 2, and defined in sections 4.1 and 4.2. There are 22 observational metrics and 32 non-observational metrics. Metric frequency tables are given in the appendix under Table 4 and Table 5.

In the metric descriptions a “task” is any activity or group of activities performed by participants during an experiment. The term task in this context can also be used to refer to the entire experiment. Examples of tasks have not been included as they will vairy depending on the goals of the system being evaluated. In the metric descriptions an “event” is something that happens (scripted or unscripted) during a task or group of tasks. An example of an event would be a security alert that appears two minutes after starting a task.

4.1 Observational Metrics

Precision

1. *Task accuracy/True positives*. The number of correctly performed actions or identified events during the task. Accuracy at which a participant performs the task [9], [16], [21]–[39].
2. *Missed events/False negatives*. The number of events missed during the task [16], [21], [25], [26], [30], [35], [40].
3. *False positives.* The number of incorrectly identified events [25], [26], [29], [30].
4. *True negatives.* The number of correctly ignored irrelevant events [25], [26], [30].
5. *Depth of insights*. The number of insights generated from a low percentage of data points [34].
6. *Breadth of insights*. The number of insights gathered from a high percentage of data points [34].

Period

1. *Time taken to perform task*. Length of time taken to perform the task [9], [22], [24]–[26], [29]–[33], [35]–[41].
2. *Reaction time.* Length of time taken to start responding to an event [27], [28], [30].
3. *Response time*. Length of time taken to finish responding to an event [38].
4. *Learning time*. Length of time taken to learn how to perform the task [42].
5. *Time looking at objects*. Length of time spent looking at different virtual elements [31].
6. *Travel time*. Amount of time spend moving compared to being stationary [31].

Progression

1. *Head movement*. Physical rotation and translation of the participant’s head [9], [16], [36].

2. *Body movement*. Physical locomotion of the participant [36], [43].
3. *Controller movement*. Physical movement of the controllers. Includes the movement of a participant's hands if using hand tracking [43].
4. *Travel distance*. Total distance travelled during a task. Can be physical or virtual movement [31].
5. Eye movement. Distance a participant's eyes move when performing a task [26].

Portion

1. *Insights discovered*. Number of insights discovered during the task [24], [26], [34].
2. *Interacted elements*. Number of times participants interact with virtual elements [26], [29], [43].
3. *Words spoken*. Number of words spoken between participants during a cooperative task [27], [28], [30].
4. *Steps to recovery*. Number of steps/procedures a participant takes to recover from a mistaken input or interaction [26].
5. *Gaze change*. Number of times or frequency which a participant changes the element they are looking at during the task [31].

4.2 Non-Observational Metrics

Procurement

1. *Questionnaire*. A set of pre-defined written questions given to participants to answer before, during and/or after undertaking the task. Questions can be qualitative or quantitative. Quantitative questions often use a Likert scale structure [21]–[28], [31], [35], [39], [40], [42], [44], [45].
2. *Interview*. An informal discussion (but can also be formalised) about the tasks performed. Discussion direction is often led by the participant's most memorable and noteworthy interactions, although the discussion can also be led by the interviewer [8], [23], [25], [26], [32], [40], [46], [47].
3. *Thinking aloud*. Participants vocalize their thoughts while performing the task. Their words are transcribed for future analysis [24], [25], [29], [34], [46].
4. *Post task test*. Participants take a test after performing the task. Test scores are used to gauge the impact of the task. This technique is commonly used to understand the memorability of data presented during the task [23], [28], [29], [42].
5. *User Experience Questionnaire.* Gathers a participant's psychological response of a system after performing the task [22], [23], [46].

6. *Differential emotion*. Gathers participant's emotional response to the task. Looks at a wide range of psychological responses including emotions such as interest, enjoyment, anger, and fear [44].
7. *Pre task test.* Participants take a test before performing the task. Test scores are used to understand a participant's pre-existing knowledge of the topic [23].

Physiological

1. *Physical demand/Fatigue*. Amount of physical exhaustion caused by the task [9], [22], [27], [34], [39].
2. *Nausea*. Extent of nauseousness induced by the task [9], [23], [39].
3. *Dizziness*. Extent of dizziness induced by the task [9].
4. *Reality*. Similarity the virtual environment has to real reality [9].
5. *Comfort*. The physical comfort of wearing and using the virtual reality equipment [9].

Psychological

1. *Performance demand/Mental effort*. Amount of mental exhaustion caused by the task [9], [22], [27], [34], [36], [39].
2. *Satisfaction*. Amount of enjoyment derived from the task [24], [32], [34], [45], [48].
3. *Immersion/Presence*. Depth of mental involvement while within the virtual environment [39], [44], [48]–[50].
4. *Frustration*. Amount of frustration caused by the task [22], [27], [33], [39].
5. *Stimulation/Motivation*. Level of enthusiasm, inspiration and engagement incited by the task [48], [49], [51].
6. *Intuitiveness*. Ease at which participants understood and felt comfortable with the task and/or interaction method [9], [25].
7. *Novelty*. The originality and unusualness of the task [51].
8. *Temporal demand*. The pressure the participant felt to perform the task at a specific pace or within a specific timeframe [27].
9. *Situation awareness*. The ease at which participants can gain situational awareness [30].

Personal

1. *Usability*. Perceived ease of use of the system [9], [24]–[26], [28], [39]–[41], [44], [45], [48], [51].
2. *Effectiveness/Usefulness*. Ease at which the task can be completed in relation to alternative methods of completing the task [22], [25], [28], [29], [45], [47], [48], [50].
3. *Perceived accuracy*. Participant's opinion on their performance and accuracy during the task [9], [22], [24], [25], [27], [39], [45].

4. *Preference*. A participant's personal preferred task. This metric should only be used when comparing between different approaches [9], [22], [25], [26], [32], [39].
5. *Intention to use*. Likeliness the participant believes they will continue using the system shown during the experiment in the future [25], [45], [47]–[49].
6. *Learnability*. Perceived ease at which the task was learnt and understood [9], [25], [33], [36], [45], [48].
7. *Perceived speed*. Participant's opinion on their speed during the task [9], [45].
8. *Cooperation*. Level of cooperation encouraged and facilitated by the task and environment [42].
9. *Attractiveness*. How visual appealing the participant thinks the system is [25], [51].
10. *Problem–solving capability*. Perceived aid the environment gave to solving the task [49].
11. *Familiarity*. The similarity the interface has with other systems used by the participant [30].

5 Discussion

Of the 34 papers surveyed 4 (11.76%) used only observational metrics, 9 (26.47%) used only non-observational metrics while the remaining 21 (61.76%) used a mixture of observation and non-observational metrics. Although most papers (75.53%) use at least one observational metric there is still a large amount (26.47%) that use only subjective metrics during evaluation and have no form of empirical validation. This gives further evidence to the claims made in section 1.2 that there is a lack of objective evaluation in this area.

Observational metrics were used a total of 74 times and non-observational metrics were used a total of 128 times. This equates to an average of 2.18 observational metrics and 3.76 non-observational metrics per paper. This shows a preference for non-observational metrics with the average paper using more non-observational metrics than observational metrics. Considering each topic area independently also shows a preference for non-observational metrics as shown in Table 3. There are three potential explanations for this preference:

1. There is large number of non-observational metrics compared to observational metrics meaning there are more non-observational metrics to *choose* from.
2. Observational metrics are typically more difficult to measure as apparatus needs to be used to record participant actions compared to asking participants their opinions.
3. There is a lack of understanding around the detrimental impact of using only non-observational metrics.

Table 3 - Average number of metrics used per paper in each topic area and in combination

	Observational	Non-Observational
Cyber Security	3.15	3.85
Data Visualisation	2.59	4.06
Virtual Reality	1.55	3.59
Combined	2.18	3.76

A table containing the surveyed metric data can be found on GitHub [52]. [1]

[1] https://github.com/danieljharris/Evaluation_Metrics

6 Conclusion and Future Work

Evaluations of work combining cyber security, data visualisation, and virtual reality are scarce due the infancy of the area and a lack of existing guidance (section 1.1). Evaluations that are conducted have an over reliance on subjective metrics, producing evaluations that analyse participant opinions of a system instead of objectively analysing it (section 1.2). This paper highlights these shortcomings and provides guidance on how to avoid them in future evaluations. Guidance is given in the form of listing, defining, and categorising metrics from user-based evaluations. Metric are categorised as “observational” or “non-observational” (section 3) to highlight the differences between reviewing a participant’s opinions of a system and objectively analysing a participant’s interactions with a system. All metrics are presented and defined in section 4.

Statistical analysis was performed to observe the differences between observational and non-observational metric use in the surveyed papers (section 5). The results show that 26.47% of papers do not use observational metrics when performing evaluations and rely solely on participant opinion to gauge the effectiveness of their systems, even though there is research warning against this practice (section 1.2). Potential reasons of this are a lack of guidance, an abundance of non-observational metrics, and added difficulty when using observational metrics.

By surveying, categorising, and presenting these metrics this paper provides guidance to produce better user-based studies and evaluations for future works that combine cyber security, data visualisation, and virtual reality.

Future work in this area could investigate and compare the depth of insights provided by each of metric during a user-based study to identify which metrics are better suited in specific scenarios. Additional work could also convert each metric into a criterion used for specific use cases. Examples of this work could be to set a criterion that stipulates the “Time taken to perform task” metric be less than 5 minutes per for all triage analysis tasks for a given system.

7 Acknowledgments

This work was supported by the KESS 2 programme. Knowledge Economy Skills Scholarships (KESS) is a pan-Wales higher-level skills initiative led by Bangor University on behalf of the HE sector in Wales. It is part funded by the Welsh Government’s European Social Fund (ESF) programme for East Wales.

8 References

[1] K. Kullman, L. Buchanan, A. Komlodi, D. Engel, and A. Komlodi, “Mental Model Mapping Method for Cybersecurity,” 2020, Accessed: Apr. 22, 2020. [Online]. Available: https://www.researchgate.net/publication/339943643_Mental_Model_Mapping_Method_for_Cybersecurity.

[2] A. Kabil, T. Duval, N. Cuppens, G. Le Comte, Y. Halgand, and C. Ponchel, “From Cyber Security Activities to Collaborative Virtual Environments Practices Through the 3D CyberCOP Platform,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11281 LNCS, pp. 272–287, doi: 10.1007/978-3-030-05171-6_14.

[3] S. Irshad, B. Seri Iskandar, M. Dayang, R. A. Rambli, M. Suziah, and B. Sulaiman, “An Interaction Design Model for Information Visualization in Immersive Augmented Reality platform,” 2019, doi: 10.1145/3365921.3365939.

[4] B. Lee, D. Brown, B. Lee, C. Hurter, S. Drucker, and T. Dwyer, “Data Visceralization: Enabling Deeper Understanding of Data Using Virtual Reality,” *IEEE Trans. Vis. Comput. Graph.*, pp. 1–1, Aug. 2020, doi: 10.1109/TVCG.2020.3030435.

[5] D. Staheli *et al.*, “Visualization evaluation for cyber security: trends and future directions,” in *ACM International Conference Proceeding Series*, Nov. 2014, vol. 10-Novembe, pp. 49–56, doi: 10.1145/2671491.2671492.

[6] A. Sethi, “EEVi : A Model Developed to Aid the Design and Evaluation Process of Cyber-Security Visualisation for Cyber-Security Analysts,” 2019.

[7] V. F. Mancuso, A. J. Strang, G. J. Funke, and V. S. Finomore, “Human factors of cyber attacks: A framework for human-centered research,” in *Proceedings of the Human Factors and Ergonomics Society*, Oct. 2014, vol. 2014-Janua, pp. 437–441, doi: 10.1177/1541931214581091.

[8] K. Kullman, N. Ben Asher, and C. Sample, “Operator impressions of 3d visualizations for cybersecurity analysts,” in *European Conference on Information Warfare and Security, ECCWS*, 2019, vol. 2019-July, pp. 257–266, [Online]. Available: https://www.researchgate.net/publication/334226184.

[9] A. Samini and K. L. Palmerius, “Popular performance metrics for evaluation of interaction in virtual and augmented reality,” in *Proceedings - 2017 International Conference on Cyberworlds, CW 2017 - in cooperation with: Eurographics*

Association International Federation for Information Processing ACM SIGGRAPH, Nov. 2017, vol. 2017-Janua, pp. 206–209, doi: 10.1109/CW.2017.25.

[10] A. D'Amico, L. Buchanan, D. Kirkpatrick, and P. Walczak, "Cyber operator perspectives on security visualization," in *Advances in Intelligent Systems and Computing*, 2016, vol. 501, pp. 69–81, doi: 10.1007/978-3-319-41932-9_7.

[11] A. D. Andre and C. D. Wickens, "When Users Want What's not Best for Them," *Ergon. Des. Q. Hum. Factors Appl.*, vol. 3, no. 4, pp. 10–14, Oct. 1995, doi: 10.1177/106480469500300403.

[12] R. W. Bailey, "Performance vs. Preference," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 37, no. 4, pp. 282–286, Oct. 1993, doi: 10.1177/154193129303700406.

[13] M. R. Drew, B. Falcone, and W. L. Baccus, "What does the system usability scale (SUS) measure?: Validation using think aloud verbalization and behavioral metrics," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Jul. 2018, vol. 10918 LNCS, pp. 356–366, doi: 10.1007/978-3-319-91797-9_25.

[14] S. Carpendale, "Evaluating information visualizations," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, vol. 4950 LNCS, pp. 19–45, doi: 10.1007/978-3-540-70956-5_2.

[15] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale," *Int. J. Hum. Comput. Interact.*, vol. 24, no. 6, pp. 574–594, Aug. 2008, doi: 10.1080/10447310802205776.

[16] A. Dünser, R. Grasset, and M. Billinghurst, "A Survey of evaluation techniques used in augmented studies," 2008, doi: 10.1145/1508044.1508049.

[17] A. Sethi, F. Paci, and G. Wills, "EEVi-framework for evaluating the effectiveness of visualization in cyber-security," in *2016 11th International Conference for Internet Technology and Secured Transactions, ICITST 2016*, Feb. 2017, pp. 340–345, doi: 10.1109/ICITST.2016.7856726.

[18] B. Shneiderman and C. Plaisant, "Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies," in *Proceedings of BELIV'06: BEyond time and errors - novel EvaLuation methods for Information Visualization. A workshop of the AVI 2006 International Working Conference*, 2006, p. 1, doi: 10.1145/1168149.1168158.

[19] D. J. Clark and B. Turnbull, “Experiment Design for Complex Immersive Visualisation,” Nov. 2020, doi: 10.1109/MilCIS49828.2020.9282380.

[20] N. Rakotondravony and H. P. Reiser, “Towards a Common Evaluation Framework for Cyber Security Visualizations,” 2017.

[21] M. Rosso, M. Campobasso, G. Gankhuyag, and L. Allodi, “SAIBERSOC: Synthetic Attack Injection to Benchmark and Evaluate the Performance of Security Operation Centers,” *ACM Int. Conf. Proceeding Ser.*, pp. 141–153, Oct. 2020, doi: 10.1145/3427228.3427233.

[22] J. Illing, P. Klinke, U. Grünefeld, M. Pfingsthorn, and W. Heuten, “Time is money! Evaluating Augmented Reality Instructions for Time-Critical Assembly Tasks,” in *ACM International Conference Proceeding Series*, Nov. 2020, pp. 277–287, doi: 10.1145/3428361.3428398.

[23] K. Ferris, G. G. Martinez, G. Wadley, and K. Williams, “Melbourne 2100: Dystopian Virtual Reality to provoke civic engagement with climate change,” in *ACM International Conference Proceeding Series*, Dec. 2020, pp. 392–402, doi: 10.1145/3441000.3441029.

[24] J. R. Goodall, “Visualization is better! A comparative evaluation,” in *6th International Workshop on Visualization for Cyber Security 2009, VizSec 2009 - Proceedings*, 2009, pp. 57–68, doi: 10.1109/VIZSEC.2009.5375543.

[25] C. J. Garneau and R. F. Erbacher, “Evaluation of Visualization Tools for Computer Network Defense Analysts: Display Design, Methods, and Results for a User Study.”

[26] J. T. Langton and A. Baker, “Information visualization metrics and methods for cyber security evaluation,” in *IEEE ISI 2013 - 2013 IEEE International Conference on Intelligence and Security Informatics: Big Data, Emergent Threats, and Decision-Making in Security Informatics*, 2013, pp. 292–294, doi: 10.1109/ISI.2013.6578846.

[27] N. A. Giacobe, M. D. McNeese, V. F. Mancuso, and D. Minotra, “Capturing human cognition in cyber-security simulations with NETS,” in *IEEE ISI 2013 - 2013 IEEE International Conference on Intelligence and Security Informatics: Big Data, Emergent Threats, and Decision-Making in Security Informatics*, 2013, pp. 284–288, doi: 10.1109/ISI.2013.6578844.

[28] Z. Huang, C. C. Shen, H. Doshi, N. Thomas, and H. Duong, “Cognitive task analysis based training for cyber situation awareness,” in *IFIP Advances in Information and Communication Technology*, 2015, vol. 453, pp. 27–40, doi: 10.1007/978-3-319-

18500-2_3.

[29] T. Nunnally, “ADVANCED VISUALIZATIONS FOR NETWORK SECURITY A Dissertation Presented to The Academic Faculty,” 2014.

[30] Mancuso Vincent Francis, “An Interdisciplinary Evaluation of Transactive Memory in Distributed Cyber Teams,” 2012. https://www.researchgate.net/publication/260856065_An_Interdisciplinary_Evaluation_of_Transactive_Memory_in_Distributed_Cyber_Teams (accessed Jun. 28, 2021).

[31] J. Liu, A. Prouzeau, B. Ens, and T. Dwyer, “Design and Evaluation of Interactive Small Multiples Data Visualisation in Immersive Spaces,” Jun. 2020, pp. 588–597, doi: 10.1109/vr46266.2020.00081.

[32] R. Blinder *et al.*, “Comparative Evaluation of Node-Link and Sankey Diagrams for the Cyber Security Domain,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Sep. 2019, vol. 11746 LNCS, pp. 497–518, doi: 10.1007/978-3-030-29381-9_31.

[33] B. J. H. Andersen, A. T. A. Davis, G. Weber, and B. C. Wunsche, “Immersion or diversion: Does virtual reality make data visualisation more effective?,” May 2019, doi: 10.23919/ELINFOCOM.2019.8706403.

[34] P. Millais, S. L. Jones, and R. Kelly, “Exploring Data in Virtual Reality: Comparisons with 2D Data Visualizations,” 2018, doi: 10.1145/3170427.3188537.

[35] Y. Lee, S. Marks, and A. M. Connor, “An Evaluation of the Effectiveness of Virtual Reality in Air Traffic Control,” in *Proceedings of the 2020 4th International Conference on Virtual and Augmented Reality Simulations*, Feb. 2020, pp. 7–17, doi: 10.1145/3385378.3385380.

[36] A. Drogemuller, A. Cunningham, J. Walsh, M. Cordeil, W. Ross, and B. Thomas, “Evaluating Navigation Techniques for 3D Graph Visualizations in Virtual Reality,” Nov. 2018, doi: 10.1109/BDVA.2018.8533895.

[37] J. L. Gabbard and J. I. Edward Swan, “Usability Engineering for Augmented Reality: Employing User-based Studies to Inform Design,” 2008.

[38] C. C. Gramazio, K. B. Schloss, and D. H. Laidlaw, “The relation between visualization size, grouping, and user performance,” *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1953–1962, Dec. 2014, doi: 10.1109/TVCG.2014.2346983.

[39] J. Wagner, W. Stuerzlinger, and L. Nedel, “The Effect of Exploration Mode and Frame

of Reference in Immersive Analytics," *IEEE Trans. Vis. Comput. Graph.*, 2021, doi: 10.1109/TVCG.2021.3060666.

[40] D. A. Bowman, J. L. Gabbard, and D. Hix, "A survey of usability evaluation in virtual environments: Classification and comparison of methods," *Presence: Teleoperators and Virtual Environments*, vol. 11, no. 4. MIT Press 238 Main St., Suite 500, Cambridge, MA 02142-1046 USA journals-info@mit.edu, pp. 404–424, Aug. 13, 2002, doi: 10.1162/105474602760204309.

[41] R. T. Azuma, "A survey of augmented reality," *Presence: Teleoperators and Virtual Environments*, vol. 6, no. 4. MIT Press Journals, pp. 355–385, Mar. 13, 1997, doi: 10.1162/pres.1997.6.4.355.

[42] S. Marks, D. White, and M. Mazdics, "Evaluation of a Virtual Reality Nasal Cavity Education Tool," in *Proceedings of 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering, TALE 2018*, Jan. 2019, pp. 193–198, doi: 10.1109/TALE.2018.8615344.

[43] B. Lee, X. Hu, M. Cordeil, A. Prouzeau, B. Jenny, and T. Dwyer, "Shared Surfaces and Spaces: Collaborative Data Visualisation in a Co-located Immersive Environment," 2020. Accessed: Nov. 09, 2020. [Online]. Available: https://github.com/benjaminchlee/FIESTA.

[44] I. Hupont, J. Gracia, L. Sanagustin, and M. A. Gracia, "How do new visual immersive systems influence gaming QoE? A use case of serious gaming with Oculus Rift," Jul. 2015, doi: 10.1109/QoMEX.2015.7148110.

[45] A. Luse, B. Mennecke, J. Triplett, N. Karstens, and D. Jacobson, "A Design Methodology and Implementation for Corporate Network Security Visualization: A Modular-Based Approach," *AIS Trans. Human-Computer Interact.*, vol. 3, no. 2, pp. 104–132, Jun. 2011, doi: 10.17705/1thci.00029.

[46] S. Hubenschmid SebastianHubenschmid *et al.*, "STREAM: Exploring the Combination of Spatially-Aware Tablets with Augmented Reality Head-Mounted Displays for Immersive Analytics," *CHI Conf. Hum. Factors Comput. Syst. (CHI '21)*, vol. 14, 2021, doi: 10.1145/3411764.3445298.

[47] M. J. Lobo, C. Hurter, and P. Irani, "Flex-ER: A Platform to Evaluate Interaction Techniques for Immersive Visualizations," *Proc. ACM Human-Computer Interact.*, vol. 4, no. ISS, Nov. 2020, doi: 10.1145/3427323.

[48] M. Barrett and J. Blackledge, "Evaluation of a Prototype Desktop Virtual Reality Model

Developed to Enhance Electrical Safety and Design in the Built Environment," *Environ. ISAST Trans. Comput. Intell. Syst.*, vol. 3, pp. 1–10, Jan. 2012, doi: 10.21427/D7862H.

[49] H. M. Huang, U. Rauch, and S. S. Liaw, "Investigating learners' attitudes toward virtual reality learning environments: Based on a constructivist approach," *Comput. Educ.*, vol. 55, no. 3, pp. 1171–1182, Nov. 2010, doi: 10.1016/j.compedu.2010.05.014.

[50] Y. Lu and T. Ishida, "Implementation and Evaluation of a High-presence Interior Layout Simulation System using Mixed Reality," 2020, doi: 10.22667/JISIS.2020.02.29.050.

[51] S. Irshad, D. R. A. Rambli, and S. Sulaiman, "Design and implementation of user experience model for augmented reality systems," in *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia*, Nov. 2020, pp. 48–57, doi: 10.1145/3428690.3429169.

[52] Harris Daniel, "Evaluation_Metrics," 2021. https://github.com/danieljharris/Evaluation_Metrics (accessed Jun. 30, 2021).

9 Appendix

Table 4 - Observational metric frequencies

Observational Metric	Frequency
Task accuracy/True positives	21
Time taken to perform task	17
Missed events/False negatives	7
Head movement	4
False positives	4
Insights discovered	3
Interacted elements	3
True negatives	3
Reaction time	3
Words spoken	3
Body movement	2
Response time	1
Learning time	1
Depth of insights	1
Breadth of insights	1
Controller movement	1
Gaze change	1
Time looking at objects	1
Travel distance	1

Travel time	1
Steps to recovery	1
Distance of eye movement	1

Table 5 - Non-observational metric frequencies

Non-Observational Metric	Frequency
Questionnaire	15
Usability	12
Interview	8
Effectiveness/Usefulness	8
Perceived accuracy	7
Performance demand/Mental effort	6
Perceived Learning Effectiveness	6
Preferred	6
Satisfaction	5
Physical demand/Fatigue	5
Immersion/Presence	5
Intention to use system	5
Thinking aloud	5
Post task test	4
Frustration	4
Stimulation/Motivation	3
User Experience Questionnaire	3
Nausea	3
Attractiveness	2
Perceived speed	2
Intuitiveness	2
Reality	2
Cooperation	1
Differential emotion	1
Problem–solving capability	1
Novelty	1
Comfort	1
Dizziness	1
Pre task test	1
Temporal demand	1
Situation Awareness	1
Familiarity	1