Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

Exploring Data Literacy Levels in the Crowd – the Case of COVID-19

Shaoyang Fan, Lei Han, Gianluca Demartini, Shazia Sadiq

School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia
shaoyang.Fan@uq.net.au, L.han@uq.edu.au, g.demartini@uq.edu.au, shazia@itee.uq.edu.au

Abstract

During global health crises, the use of data becomes critical to control the spread of infections, to inform the general public and to foster safe behaviors. The ability of people to read and understand data (i.e., data literacy) has the potential to affect human behaviors. In this paper, we study non-expert human subjects' ability to make accurate interpretations of complex pandemic data visualizations designed for general public consumption. We present them with popular plots and graphs that have been shown by traditional and social media, and ask them to answer questions to assess their data literacy at three levels: extracting information, finding relationships among data, and expanding or predicting information. Our results show the presence of variance in interpretations and reveal insights into how messages communicated through data may be perceived differently by different people. We also highlight the importance of designing communication strategies that ensure the spread of the right message through data.

Introduction

During global health crises, critical decision-makers like epidemiologists, government decision-makers, and scientists have a heightened reliance on timely and accurate data to act as effectively as possible. At the same time, public-facing data also becomes critical in such situations when even a single individual behavior can have a positive or negative impact on the spread and severity of a global pandemic (Nicolaides et al. 2020; Peiris, Guan, and Yuen 2004). During the COVID-19 pandemic, various measures such as infection rate and R0 trends and messages like "flatten the curve" have been presented across traditional and social media on a daily basis (Repišti et al. 2020; Cinelli et al. 2020; Hussain 2020; Lee et al. 2021). Thus, understanding how data visualizations and analytical results are consumed by the general public becomes a health and public safety imperative.

Our aim is to investigate how the public is looking at data visualizations and messages during a global health crisis. Specifically, we explore the accuracy of the interpretations they make and how an individual's context impacts the way data is consumed and perceived by them. In this paper, we present the results of a study conducted by means of crowdsourcing to explore the ability of people to read, understand and interpret pandemic related visualizations. A number of factors may influence such a study, including but not limited to the representation type of the data story (e.g., the type of chart or the interactivity level) and the complexity of the question being asked. The multi-faceted nature of the setting requires a systematic approach to the study design.

Accordingly, we run a crowdsourcing experiment to ask workers to answer questions based on different types of data and visualisations presented to them. Following the methodology of three-level comprehension of statistical information (Friel, Curcio, and Bright 2001), for each type of visualization we investigate the capability of crowd workers to: (i) read and find specific information (Level 1); (ii) extract relationships between different data points (Level 2); and (iii) to analyze information and predict ongoing trends (Level 3). Moreover, we additionally look at those misunderstandings (i.e., wrong answers to our questions) that the public has made with high confidence, as such phenomenon implies the general public not aware of mistakes having been made. This is a case of unsuccessful communication, which, during pandemics, may lead to critical consequences.

Related Work

Data literacy was defined by Lee, Kim, and Kwon (2016) as the capability to interpret and to extract information given the visually represented data. Boy et al. (2014) studied users' data visualization literacy through Amazon MTurk and showed that the public has can understand line charts while reading bar charts or scatter plots are more difficult for them. Lee, Kim, and Kwon (2016) proposed a data visualization literacy test following psychological and educational measurement. Their method has shown to be effective in evaluating users' ability to learn from unfamiliar data visualizations. Börner et al. (2016) found that people's data visualization literacy varies from one person another significantly, and most participants cannot understand popular visualization types appearing in newspapers, textbooks and magazines. Related to this line of research, we look at how people can make sense of various popular COVID-19 related visualizations designed for general public consumption.

In early years, Wood (1968) summarized three facets in mathematical information comprehension: translation, interpretation and extrapolation. Later, Friel, Curcio, and Bright (2001) developed these concepts and defined three comprehension levels for statistical graphs: (i) describing graph content; (ii) sorting factors from graph; and (iii) identifying implied consequences that may not be displayed in a graph. Based on these graph comprehension levels, Galesic and Garcia-Retamero (2011) constructed a graphical literacy scale to assess the skills in understanding various charts. Similarly, Wakeling et al. (2015) developed three levels of questions to evaluate the difficulty of comprehending business data visualizations. Different from these works that only studied primitive graphs, our work aims at understanding reader' data literacy abilities for complex visualizations.

Study Design

To investigate if and how the public can consume data stories and messages related to global health crises such as COVID-19, we have approached the problem through a data literacy lens using popular COVID-19 data stories, and by observing how people can understand the information conveyed by the most frequently used data visualizations. To this end, we focus on the following research questions (RQs):

- (RQ#1) How well can people understand health-related data visualizations using their data literacy skills?
- (RQ#2) To what extent and how the accuracy of the public understanding COVID-19 related data stories is correlated with their perceived confidence?

The Selection of Data Visualizations

In our experiment, we select seven data visualizations that are frequently used by trusted sources conveying COVID-19 related information (Lee, Kim, and Kwon 2016; Lee et al. 2021), including mainstream news agencies[1,2], academia[3,4], and most visited and governmental websites[5,6,7]. The seven visualizations cover the most frequent visualization types, as discussed by Lee, Kim, and Kwon (2016); Lee et al. (2021). For example, visualization D is Johns Hopkins University's COVID-19 dashboard, which has received up to 3 to 5 billion visits per day (Recker 2021). We keep the originality of these visualizations and have not altered any information in them, such as the titles, legends, and data.

Figure 1 and 2 show the 7 selected data visualizations. Visualization A shows the total number of confirmed cases in the US over time. Visualization B shows positive cases scenarios across different levels of physical distancing compliance. Visualization C shows the number of confirmed cases in different countries over time. Visualization D shows an interactive dashboard with a map, cases, deaths, and growth curves. Visualization E is a sortable table showing the number of confirmed, recovered, death and cases per one million inhabitants in different countries. Visualization F compares mortality and infection rates between COVID-19 and other pandemics. Visualization G shows the cumulative number of deaths since the 10th death case in each country.

Experimental Setup

To study how people can understand these data visualizations, we deploy our experiment over crowdsourcing platforms by asking crowd worker questions that can be answered from the information communicated by the visualizations. Figure 1 shows the task interface (UI). In each page, we show one visualization on the left of the task UI, and three questions on the right. The three questions are sorted by different levels of difficulty in graph literacy (see Tab. 1). For each question, we also provide workers with a slider-bar to indicate their perceived confidence in answering the question. Each worker is expected to read all the 7 visualizations (in random order), and thus, each participant has to answer 21 questions. We embed Javascript in the tasks to log their low-level interaction with our task UI, including mouse clicks and any input from the keyboard. All workers have read our information consent document and accepted such data being collected while working on the task. As a quality control method, we check (i) if they spend certain amount of time on each visualization, and (ii) whether they write or select any information that does not exist in the visualization (e.g., country name in Visualization C, see Fig. 2ii). Therefore, we have recruited a total of 300 workers who passed our quality check through MTurk[8] and Toloka[9]. Among them, 100 workers are from US, India, and Russia, respectively, because US and India are the two main sources of workers on MTurk (Ross et al. 2010) and Toloka is a Russian crowdsourcing platform. Due to our choice of using popular visualizations, participants may have been exposed to some of the visualizations prior to the study. Recruiting participants via crowdsourcing from three different countries reduces the risk of prior exposition bias, as crowd workers from Russia and India may be less likely to have seen popular western visualizations before the study.

Three Levels of Question Difficulty

To ask questions for each visualization, we consider 3 different levels, following the levels of graph literacy defined by Friel, Curcio, and Bright (2001).

- Level 1 (*information finding*): Extracting obvious information from the graph as answers to the question (Q1);
- Level 2 (*recognizing relationships*): Determining the relationship in the data shown in the graph (Q2); and
- Level 3 (*making inferences*): Analyzing, expanding, and predicting the information presented with the graph, as well as explaining how to obtain the answers (Q3).

Table 1 lists all the questions in our experiment (also see Fig. 1 for Vis. A questions). All researchers reviewed all

[1] https://www.abc.net.au/news/2020-03-25/coronavirus-covid-19-modelling-stay-home-chart/12084144

[2] https://www.bbc.com/news/world-51235105

[3] https://www.nature.com/articles/d41586-020-00758-2

[4] https://coronavirus.jhu.edu/map.html

[5] https://www.google.com/covid19-map/

[6] https://www.visualcapitalist.com/infection-trajectory-flattening-the-covid19-curve/

[7] https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html

[8] https://www.mturk.com/

[9] https://toloka.yandex.com/

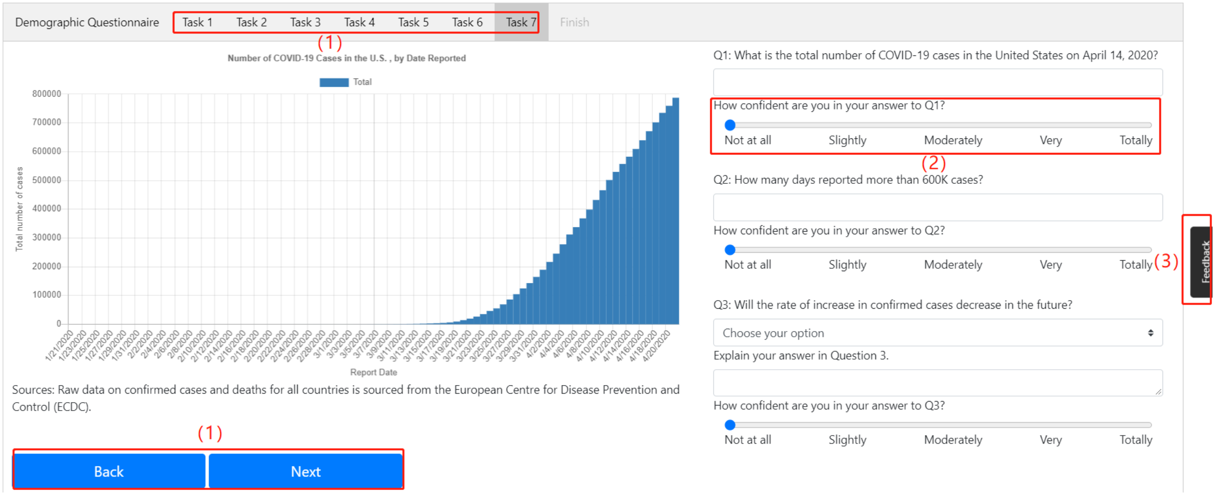


Figure 1: Task page involving *Visualization A*: (1) tabs and buttons allow worker to revise answers; (2) a 5-point Likert scale is used to report the self-perceived confidence in answering questions; (3) feedback system.

	Level of Difficulty		
	Level-1 (Q1)	*Level-2* (Q2)	*Level-3* (Q3)
B	At which point in time do the physical distancing measures begin?	At 100 days, what is the expected difference of active cases between 80% and 90% compliance?	The number of active cases can be kept under 6000 after100 days if 1 in 7 people comply with physical distancing?
C	Which country has the most total confirmed cases?	As of Apr 1, 2020, how many countries have more confirmed cases than China?	If the current trend continue, which country will have the most cases?
D	When did the number of confirmed cases begin to exceed 10K in Spain?	How many days did Spain have more than 8K daily cases in March and April?	On March 3, did South Korea or China have more new daily cases?
E	Which country has seen the most cases relative to population?	Which country has seen the most deaths relative to the confirmed cases?	Which two countries are expected to have the most recovery in the coming days?
F	Which disease has the highest fatality rate?	What three diseases have the highest R0?	Is the R0 for COVID-19 expected to rise as time goes by?
G	Which country has less than 10 000 deaths?	Which country has a similar spread pattern with Italy?	Is national lock-down working?

Table 1: Questions for each visualization. The questions for Visualization A is shown in Figure 1.

questions to make sure each question clearly reflects the relevant visualization without using complex or ambiguous vocabulary, and strictly follows the criteria of the three question difficulty levels. Level-3 (Q3) are multiple-choice questions, and thus we also ask workers to justify their answers.

Results

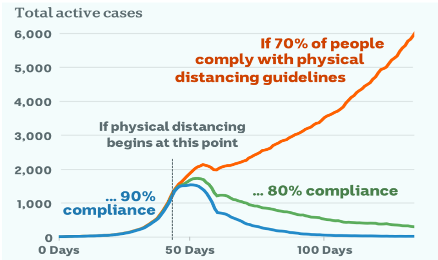
Task Performance (RQ#1)

We conducted the experiment between 9 May 2020 and 14 May 2020 over the MTurk platform and between 18 June 2020 and 21 June 2020 over the Toloka platform. In total, 300 workers have completed 6300 questions, in which we collected 28 072 behavior log entries. As workers self-reported, 68.33% of the participants are male while 30.67% are female. The average age of all participants is 32.26 years, and 74.67% of them possess university degrees. On average, each worker spent 23.8 minutes ($SD = 13.0$) to answer the questions for all seven visualizations.

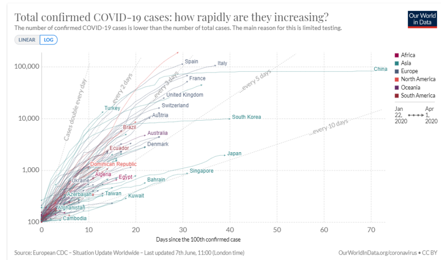
Table 2 shows the answer accuracy across different visualizations and levels of question difficulty. Overall, workers achieve the highest accuracy in Visualization C (47.8%), while their performance in Visualization D and E is the worst. Workers' performance in Level-1 questions is generally better than the other two for each visualization (except for Visualization D). Performance on Level-3 questions, however, may not be necessarily worst than that on Level-2 questions. In Visualization C and D, for example, their accuracy on Q2 is the worst among the three questions. To answer Q2, they need to find the relationship between multiple data points in the presented visualization. Graphs with too much information may increase the complexity of the visualization, which results in a higher cognitive load.

To determine the relationship between workers' background variables and performance, we used rank-based non-parametric tests due to non-normally distributed data. The spearman's rank-order correlation tests are run for continuous variables such as age, while the Kruskal-Wallis H test is employed for nominal or ordinal variables such as gender, education, income level, and country. Test results showed no statistically significant differences in accuracy across demographic variables, with all $p > 0.05$.

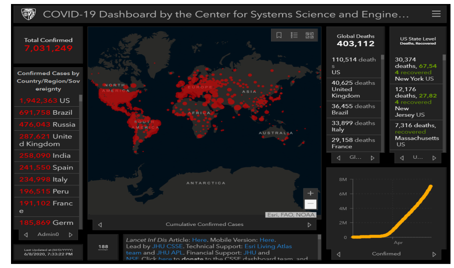
Figure 3 shows the time spent by participants on each visualization. We can see that *visualization D*, *visualization C*



i *Visualization B*

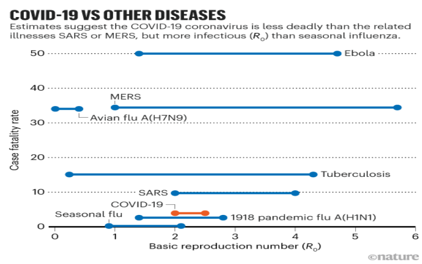


ii *Visualization C*

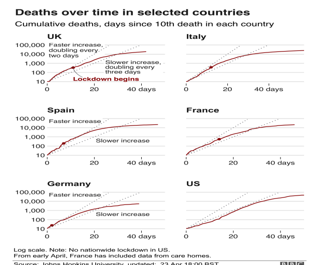


iii *Visualization D*

iv *Visualization E*



v *Visualization F*



vi *Visualization G*

Figure 2: Data visualizations used in the study.

	A	*B*	*C*	*D*	*E*	*F*	*G*
Level-1 (Q1)	191 (63.7%)	151 (50.3%)	235 (78.3%)	87 (29%)	106 (35.3%)	132 (44%)	195 (65%)
Level-2 (Q2)	136 (45.3%)	79 (26.3%)	75 (25%)	40 (13.3%)	36 (12%)	114 (38%)	86 (28.7%)
Level-3 (Q3)	44 (14.7%)	56 (18.7%)	120 (40%)	97 (32.3%)	22 (7.3%)	18 (6%)	20 (6.7%)
Total	371 (41.2%)	286 (31.8%)	430 (47.8%)	224 (24.9%)	164 (18.2%)	264 (29.3%)	301 (33.4%)

Table 2: Number (and percentage) of correct answers with a breakdown of visualizations and question difficulties.

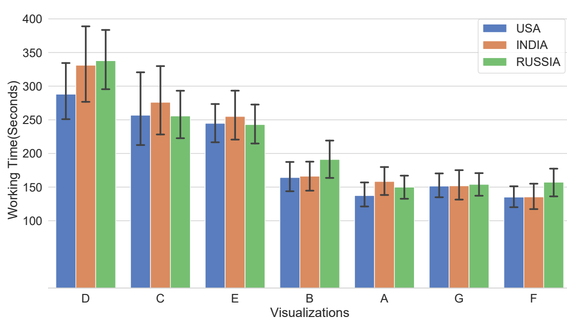


Figure 3: Average task time by visualization.

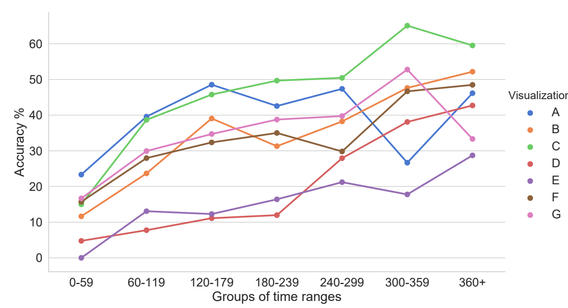


Figure 4: Accuracy rate over time spent by workers.

and *visualization E* require longer time to work on (318.747, 262.82, and 247.58 seconds, respectively) than the other four visualizations (less than 180 seconds). In fact, the former group (*visualization C*, *visualization D* and *visualization E*) consists of interactive visualizations while the other four are static. The performance of workers on *visualization C* is the best, while they have the lowest accuracy on *visualization D* and *visualization E* (see Table. 2). These results show that interactive visualization requires longer time to explore but do not necessarily incur into higher comprehension difficulties (e.g., *visualization C*).

To investigate the relationship between time spent and answer quality, we divide the time spent into 60 seconds intervals. Figure 4 shows answer quality over time spent on each visualization. We can see that in the first 180 seconds, the accuracy is increasing together with time spent. However, when more than 180 seconds are spent on the visualizations, the accuracy does not necessarily increase.

1	2	3	4	5
17.85%	16.77%	23.53%	30.77%	54.88%

Table 3: Average accuracy over **Confidence level**.

Group	High Confidence (confidence score $\geqslant 4$)	Low Confidence (confidence score $\leqslant 2$)
Correct	1455	189
Wrong	2054	919
Total	3509	1108

Table 4: Number of correct and wrong answers over confidence levels (i.e., high and low confidence groups).

	Level-1 (Q1)	Level-2 (Q2)	Level-3 (Q3)	Total
E	105 (27.42%)	142 (37.08%)	136 (35.51%)	383
G	47 (14.24%)	114 (34.55%)	169 (51.21%)	330
F	96 (30.19%)	87 (27.36%)	135 (42.45%)	318
D	90 (34.22%)	87 (33.08%)	86 (32.7%)	263
B	60 (23.53%)	72 (28.24%)	123 (48.24%)	255
A	63 (21.88%)	73 (25.35%)	152 (52.78%)	288
C	29 (13.36%)	102 (47%)	86 (39.63%)	217
	490 (23.86%)	677 (32.96%)	887 (43.18%)	2054

Table 5: Number (and percentage) of incorrect answers with high confidence (i.e., confidence score $\geqslant 4$) for each visualization (sorted by the total number of errors in descending order). The numbers in parentheses indicate the percentage to total number of errors for each visualization.

Worker Confidence (RQ#2)

Table 3 shows the mean accuracy over the five confidence levels. It is clear that the accuracy for highest confidence is higher than that for other confidence levels. Table 4 presents the number of correct and wrong answers that have been made over high and low confidence levels. We observe that 3509 responses are provided with a high confidence level (55.7% of the total 6300 answers), among which 2054 answers (58.5% of 3509) are incorrect. Although workers with high confidence may lead to more accurate answers compared to those with low self-confidence, there are also a large number of high confidence answers that are actually wrong.

Table 5 shows the breakdown of the incorrect answers with high confidence by visualizations and question levels. Overall, misunderstanding becomes more frequent when the question level is higher; 23.86%, 32.96% and 43.18% of high-confidence errors happen while answering Q1, Q2 and Q3, respectively. By manually analyzing the justifications provided in Q3, we find that part of the high-confidence errors are caused by workers who answer the questions based on their own knowledge or beliefs rather than on the information presented in the chart. Furthermore, Visualization C receives more high-confidence errors when answering Q2 (47%), compared to other visualizations. Note that, workers need to compare multiple data points to answer Q2, and the acquisition of a single data point requires interactive operations (e.g., hover over country names). Therefore, they may miss some data which may lead to wrong conclusions without awareness. In addition, there are a large proportion of high-confidence errors in Q1 for Visualization D (34.22%), as it contains the most comprehensive information in the dashboard, which may increase cognitive complexity.

Discussion and Conclusions

In this paper, we study how the general public consumes data related to public health crises. Through a crowdsourcing experiment by asking workers a set of questions that require comprehension of popular COVID-19 data visualizations, we show that there exists variability in understanding these visualizations. We find that misunderstandings and incorrect interpretations of the data are common and likely to happen in some cases, which may lead to a dangerous outcome as the safety of general public behaviors could be affected by the way they perceive the data presented to them. In our experiment, we do not observe strong correlation between data literacy levels and worker demographics. Thus, it becomes important to design such data visualizations to maximize their communication effectiveness for everyone.

Generally speaking, interactive visualizations often take longer time and more effort to explore. Thus, data publishers should be mindful about this, if the aim is to communicate clear messages efficiently to a broad audience. Abundance of information may even elicit the least correct interpretation. We find that spending too much time on the visualization does not necessarily lead to a better understanding (see Fig. 4). This can be explained by the fact that the advantage of spending time to extract richer information may be offset by increasing the cognitive load and mental effort (Huang, Eades, and Hong 2009). Our results show that one popular COVID-19 dashboard from JHU (i.e., Visualization D) may be complicated for the public to understand (e.g., when answering Q1, see Tab. 2). This is consistent with Recker (2021) suggesting that current COVID-19 dashboard designs need improvements. When policymakers employ data visualizations to convey important information to the public, these visualizations should focus on critical information and remove or downplay unimportant data facets.

Moreover, a number of high-confidence errors indicate making misinterpretations without awareness. Through the manual inspection of workers' justifications, we found that part of the high-confidence errors was caused by answering questions based on their own experiences or beliefs rather than on the data presented to them in the visualization. In our future work, we will delve into how to reduce misunderstandings from crisis data visualizations.

Acknowledgments

This work is partially supported by the ARC Discovery Project (Grant No. DP190102141) and the by the ARC Training Centre for Information Resilience (Grant No. IC200100022). We also thank Toloka for providing access to their platform.

References

Börner, K.; Maltese, A.; Balliet, R. N.; and Heimlich, J. 2016. Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization*, 15(3): 198–213.

Boy, J.; Rensink, R. A.; Bertini, E.; and Fekete, J.-D. 2014. A principled way of assessing visualization literacy. *IEEE transactions on visualization and computer graphics*, 20(12): 1963–1972.

Cinelli, M.; Quattrociocchi, W.; Galeazzi, A.; Valensise, C. M.; Brugnoli, E.; Schmidt, A. L.; Zola, P.; Zollo, F.; and Scala, A. 2020. The COVID-19 social media infodemic. *Scientific Reports*, 10(1): 1–10.

Friel, S. N.; Curcio, F. R.; and Bright, G. W. 2001. Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in mathematics Education*, 124–158.

Galesic, M.; and Garcia-Retamero, R. 2011. Graph literacy: A cross-cultural comparison. *Medical Decision Making*, 31(3): 444–457.

Huang, W.; Eades, P.; and Hong, S.-H. 2009. Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization*, 8(3): 139–152.

Hussain, W. 2020. Role of social media in COVID-19 pandemic. *The International Journal of Frontier Sciences*, 4(2): 59–60.

Lee, C.; Yang, T.; Inchoco, G. D.; Jones, G. M.; and Satyanarayan, A. 2021. Viral Visualizations: How Coronavirus Skeptics Use Orthodox Data Practices to Promote Unorthodox Science Online. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–18.

Lee, S.; Kim, S.-H.; and Kwon, B. C. 2016. Vlat: Development of a visualization literacy assessment test. *IEEE transactions on visualization and computer graphics*, 23(1): 551–560.

Nicolaides, C.; Avraam, D.; Cueto-Felgueroso, L.; González, M. C.; and Juanes, R. 2020. Hand-hygiene mitigation strategies against global disease spreading through the air transportation network. *Risk Analysis*, 40(4): 723–740.

Peiris, J. S.; Guan, Y.; and Yuen, K. Y. 2004. Severe acute respiratory syndrome. *Nature medicine*, 10(12): S88–S97.

Recker, J. 2021. Improving the State-Tracking Ability of Corona Dashboards. *the European Journal of Information Systems*.

Repišti, S.; Jovanović, N.; Kuzman, M. R.; Medved, S.; Jerotić, S.; Ribić, E.; Majstorović, T.; Simoska, S. M.; Novotni, L.; Milutinović, M.; et al. 2020. How to measure the impact of the COVID-19 pandemic on quality of life: COV19-QoL–the development, reliability and validity of a new scale. *Global Psychiatry*.

Ross, J.; Irani, L.; Silberman, M. S.; Zaldivar, A.; and Tomlinson, B. 2010. Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, 2863–2872. New York, NY, USA: ACM.

Wakeling, S.; Clough, P.; Wyper, J.; and Balmain, A. 2015. Graph literacy and business intelligence: Investigating user understanding of dashboard data visualizations. *Business Intelligence Journal*, 20(4): 8–19.

Wood, R. 1968. Objectives in the teaching of mathematics. *Educational Research*, 10(2): 83–98.