# DATA SCIENTISTS' EPISTEMIC THINKING FOR CREATING AND INTERPRETING VISUALIZATIONS AND THE IMPACT FOR STUDENTS' DATA VISUALIZATION LITERACY

Charlotte Bolch
School of Teaching and Learning
University of Florida
cbolch21@gmail.com

*Data visualizations have become an essential way to present information that facilitates communication of large datasets. The complex thinking processes (epistemic thinking) that data scientists engage in along the data visualization literacy continuum has not been fully explored. The common skills and strategies for interpreting and creating visualizations was investigated among of 16 researchers in Data Science using three rounds of Delphi panels. Skills and strategies were identified qualitatively using thematic analysis after Delphi panel 1 and then brought back to the researchers in Delphi panel 2 for them to rate the level of importance they attribute to those skills/strategies. Consensus was determined using a cutoff for the interquartile range for each skill/strategy and overall group ratings presented to researchers in Delphi panel 3 for them to adjust their ratings. Implications of the study regarding the thought processes of data visualization to inform curriculum development are provided.*

## BACKGROUND

Data is constantly being collected and stored to understand people's daily interactions such as their Internet web page history or their use of mobile devices (Takemura, 2018)**.** As a result of large and diverse data being collected, the use of visualization methods and techniques to make sense of these big datasets has become an essential way to represent and interpret information (Figueiras, 2013). Data visualization is a process for representing information to facilitate understanding, identify trends and patterns, and make inferences about data (Kapler & Wright, 2004). Azzam and Evergreen (2013) define data visualization as a process of creating a representation that consists of the following three criteria: (1) the data must be qualitative or quantitative, (2) the raw data is accurately represented and important information is not omitted, and (3) the data can be explored, examined, and communicated. Overall visualizations are intended to facilitate communication and exploration of statistical information that saturates daily life (Koparan & Güven, 2015; Philip, Olivares-Pasillas, & Rocha, 2016).

Adults need to understand how data and graphs are used to communicate information and influence people's decisions in order to become informed democratic citizens in society (Shaughnessy, 1992). This process of understanding how to evaluate, critique, and construct data visualizations is defined as data visualization literacy which is developed and constructed throughout an individuals' life. For this study, the definition of data visualization literacy was understood from the reconceptualized four resources model for literacy of visual and multi-model texts (Serafini, 2012). The four roles that the reader has when interacting with a multi-modal text are: reader as navigator, interpreter, designer, and interrogator. The four resources model allows data visualization literacy to be understood from the perspective of consuming and producing visualizations. A consumer of a visualization primarily assumes the roles of navigator and interpreter, while the roles of designer and interrogator are usually performed by the producer of a visualization.

The comprehension of visualizations requires complex thinking regarding being able to read the visualization and understand the data values being portrayed, make meaning from those data values and form an interpretation about the visualization. The thinking processes that data scientists (experts in Data Science) go through with a visualization requires them to consider information and claims from the visualization and involves processes of reasoning about the information and claims. These types of thinking processes can be understood through the framework of epistemic thinking which involves understanding how people think about certain information and knowledge (Barzilai & Zohar, 2016; Chinn, Rinehart, & Buckland, 2014). From the perspective of epistemic thinking, the complexity of comprehending a visualization through the process of reading, finding meaning, making an interpretation, and finally expanding and applying that knowledge can be identified. The

In: P Arnold (Ed.), *New Skills in the Changing World of Statistics Education*
*Proceedings of the Roundtable conference of the International Association for Statistical Education (IASE),*
July 2020, Online. Voorborg, The Netherlands. 

purpose of the study was to understand the experiences of data scientists regarding data visualization to define the high-end anchor of the learning progression of data visualization literacy. The study explored data scientists' common skills and strategies of data visualization and their epistemic thinking about interpreting and creating visualizations.

## METHODS

### *Research Questions*

The research questions for this study were (1) What common skills and strategies define *interpreting* visualizations based on the experiences from data scientists? (2) What common skills and strategies define *creating* visualizations based on the experiences from data scientists?

### *Research Design*

To answer the research questions, the Delphi method was used as an effective way to gather information about skills and strategies regarding data visualization from researchers in the field of Data Science. The Delphi method is defined as a group facilitation technique that uses multiple iterations or revisions of a survey administrated to a group of experts in the field until consensus is achieved (Hasson, Keeney, & McKenna, 2000; Manizade & Mason, 2011; Ritchie & Earnest, 1999; van Zolingen & Klaassen, 2003). The Delphi method "may be characterized as a method for structuring a group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem" (Linstone & Turoff, 1975, p. 3). The use of the Delphi method allowed for a rich understanding and to build a consensus of how data scientists interpret and create visualizations. With multiple rounds of questions, each panel was informed and refined by the analysis of responses from previous questions. In this study, a Delphi panel was defined as the group of participants that the researcher intentionally facilitated discussions with as the researcher presented participants with a synthesis of results from previous panels and the participants responded to those results. The study was composed of three rounds of Delphi panels using the survey software system, Qualtrics (Qualtrics, Provo, UT) to collect the responses from participants as well as present the synthesis of the results from the previous panel.

### *Sample of Participants*

The participants for this study were researchers from one university in the field of Data Science or researchers whose projects involved components of Data Science and they met the three inclusion criteria and did not meet any of the exclusion criteria. The inclusion criteria for participants were (1) affiliation with a Data Science Institute or Initiative or a department/school of statistics, biostatistics, computer science, mathematics, engineering, biomedical informatics, applied sciences, economics, biology, or agricultural sciences, (2) research focus involving working with high-dimensional data or big datasets, (3) research involving application of quantitative methods to understand data such as advanced statistical techniques, machine learning techniques, or text/language processing. The exclusion criteria were (1) no mention in research biography about modeling, analysis, or processing of data or (2) job position title is Lecturer or job position is focused only on teaching.

The sampling method for this study was based on purposeful sampling and did not allow the researcher to make conclusions that can be generalized to the larger population of data scientists. Each potential participant was sent an email invitation stating the purpose of the study, what their involvement would be, time commitment for the study, and asking them to click on the link to the survey if they are interested in participating in the study. Sixteen participants gave informed consent and IRB approval was obtained for the study from the Institutional Review Board at the university of the researcher.

A total of 16 participants completed Delphi panel 1 from a wide range of departments/schools such as the Department of Biology, School of Forest Resources and Conservation, Department of Computer and Information Science and Engineering, and School of Special Education, School Psychology, and Early Childhood Studies. The job titles for the participants were assistant professor (N = 7), associate professor (N = 5), professor (N = 3), and scientist (N = 1). Twelve of the sixteen participants describe themselves as a data scientist. However, the four participants that did not identify as a data scientist provided the following descriptions: engineer/medical physicist, evolutionary and computational biologist, modeler, and remote sensing/GIS analyst.

*Delphi Method*

The questions for Delphi panel 1 focused on understanding the research experiences of the participants and gathering their comments and thoughts on the skills/strategies they want people to use when interpreting and creating a data visualization. A definition of a data visualization from the literature was provided to them in order to make sure that all participants had a similar reference point of a data visualization when answering the questions.

In the second Delphi panel (N = 13), the results of the qualitative analysis from Delphi panel 1 were provided to the participants which was a list of emerging skills/strategies for interpreting and creating visualizations. There were two parts for Delphi panel 2: (1) interpreting an interactive visualization and (2) thinking about creating a visualization. The participants were asked to interpret an interactive visualization about flu trends from 2012 until the present and identify the skills/strategies they used from the list. Then, for each skill/strategy the participants were asked to rate the level of importance using a Likert scale (1=Not important to 5 = Extremely important) to start the process of building consensus about the general skills/strategies for interpreting a visualization. The second part of Delphi panel 2 included a similar format to part 1 but focused on the skills/strategies for creating a visualization. The participants were asked to think back and recall a recent visualization that they had created and describe that visualization in a few sentences. Then participants were asked to select all the skills/strategies from the list that they use when creating a visualization and rate those skills/strategies according to the level of importance when creating a visualization using the same Likert scale.

For the final and third Delphi panel (N = 13), the responses from the Delphi panel 2 in regards to everyone's ratings of the skills/strategies for interpreting and creating a data visualization were analyzed and boxplot visualizations were created to graphically display the overall ratings from all participants. The purpose of Delphi panel 3 was to see if any of the participants would like to adjust any of their ratings from Delphi panel 2 after seeing the boxplot visualizations. The third panel is important for the Delphi method because the goal is to reach a consensus among the diverse panel of experts in the study about their opinions regarding the skills/strategies for interpreting and creating a data visualization (Ritchie & Earnest, 1999).

*Data Analysis*

The open-ended response data that was collected in Delphi panels 1 and 2 were analyzed using thematic analysis by Braun & Clarke (2006). Also, the literacy framework of the four resources model was used to identify the roles that participants engage in when interpreting and creating visualizations. For Delphi panels 2 and 3, the rating scores for each skill/strategy were analyzed using descriptive statistics (N, median, 1st quartile, 3rd quartile, etc.) to determine whether consensus had been achieved using the interquartile range (middle 50% of the responses). The cutoff point for the interquartile range in order to determine consensus among participants required (1) a minimum sample size of 5 given the small sample (N = 13) for the study and that all participants didn't rate all skills/strategies and (2) an interquartile range $\leq 1.2$ (Alexander, 2008; Baker, 2005; Hussein, 2010).

The qualitative analysis for Delphi panel 1 consisted of identifying emerging skills and strategies for interpreting and creating visualizations to be used in Delphi panel 2. The first step in the process for identifying skills and strategies was to code all responses as either *interpreting* or *creating*. Then, using the literacy framework of the four resources model each of the *interpreting* or *creating* responses was coded at the next level as either a *navigator, interpreter, designer,* or *interrogator*. Finally, each of those codes was read through again and emerging skills/strategies were identified. For example, a response coded as *interpreting/interpreter* was "Skills or strategies to use when interpreting a data visualization are to draw meaningful comparisons between different levels or factors of variables" (ID HVPZA, Delphi panel 1 question 6) and was identified as the emerging skill of *drawing comparisons among variables*. The response of "Perhaps, to this end, the most important skill is to carefully read the caption and/or labels" (ID 68X21, Delphi panel 1 question 6) which was coded as *interpreting/navigator* and was identified as the emerging skills of *reading captions* and *reading axes.*

A response that was coded as *creating/designer* was "Understanding how to decide between variables to put in axis versus variables to put in panels for best comprehension" (ID 08LB9, Delphi panel 1 question 7) was identified as the emerging skill of *visualizing multiple variables at once.* Another example of a response that was coded as *creating/interrogator* was "We will then look at all

of these to decide which format best allows the most efficient and rapid interpretation by the reader" (ID 5UDQT, Delphi panel 1 question 7) was identified as the emerging skill of *designing visualizations with clear and efficient meaning*. The final step was reading through the emerging skills and strategies and combining some of them together such as *dynamic visualizations* and *interacting with the visualization* into the final skill/strategy as *exploring data by interacting with the visualization*. These emerging skills and strategies were used as a starting point for the consensus building process for Delphi panels 2 and 3.

RESULTS

The final results for the study after participants had rated their level of importance for their selected skills/strategies for interpreting and creating visualization (Delphi panel 2) and were provided the overall ratings in boxplot visualizations (Delphi panel 3) to adjust their ratings if desired are shown in Figures 1 and 2. The count of participants that selected each skill/strategy are provided in the y-axis label. The skills/strategies in the boxplot visualizations are ordered to indicate the level of consensus among the participants based on the results of the interquartile range. Skills/strategies starting with a "*" were determined to have reached consensus (IQR $\leq$ 1.2). A skill/strategy with a "^" indicated that that skill/strategy was added during Delphi panel 2 by a participant. There were only minor changes to the ratings of skills/strategies by the participants in Delphi panel 3 which did not result in any of the skills or strategies changing from having achieved consensus to not achieving consensus or the other way around.

The results of the first research question indicated that consensus was determined among the participants for 6 skills/strategies for interpreting visualizations. The six skills/strategies that were confirmed as important during the process of interpreting a visualization were *understanding the layout of the visualization, reading axes, constructing meaning from the visualization/gaining insight, drawing comparisons among variables, reading captions/text*, and *understanding the definition/meaning of variables displayed*. These skills were not passive actions and consisted of thinking processes that are more than just "reading" the graph and practicing interpreting graphs. However, the number of skills and strategies that were selected and were found to have achieved consensus clearly indicated that there were complex thinking processes that the participants engaged in when interpreting a visualization rather than just reading the graphical display.

The results of the second research question indicated that consensus was determined among the participants for 11 skills/strategies for creating visualizations. The 11 skills/strategies that achieved consensus as important during the process of creating a visualization were *aesthetic sense, thinking about the research questions from the study/experiment, designing visualizations with clear and efficient meaning, labeling all aspects of the visualization (axes, legend, etc.), scaling axes appropriately, facilitating comparisons among graphs in a visualization, critical thinking skills, defining the purpose of the visualization, highlighting main points/patterns (relationships/trends), using color to highlight multiple variables,* and *using story telling techniques*. The skills and strategies identified were all active actions that the participants engaged in when creating a visualization.

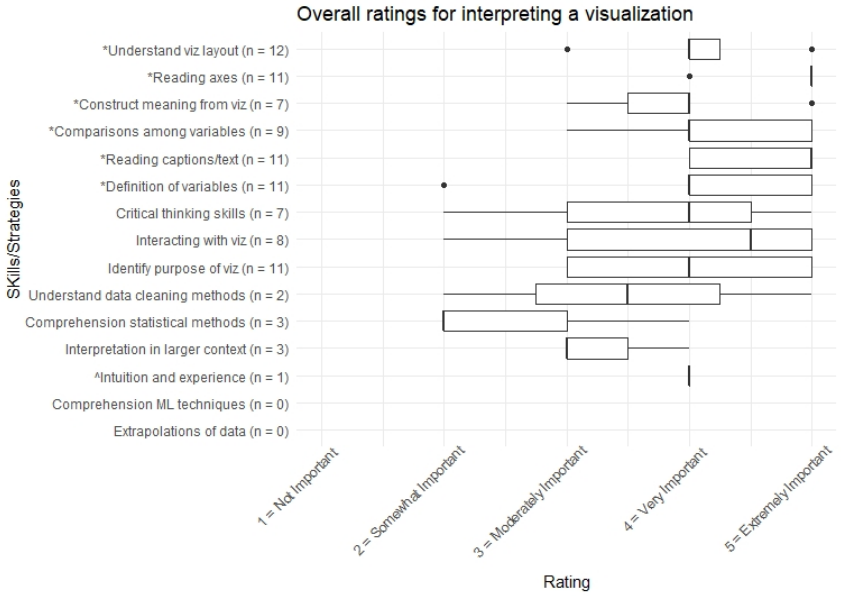


Figure 1. Overall ratings for skills/strategies about interpreting a visualization

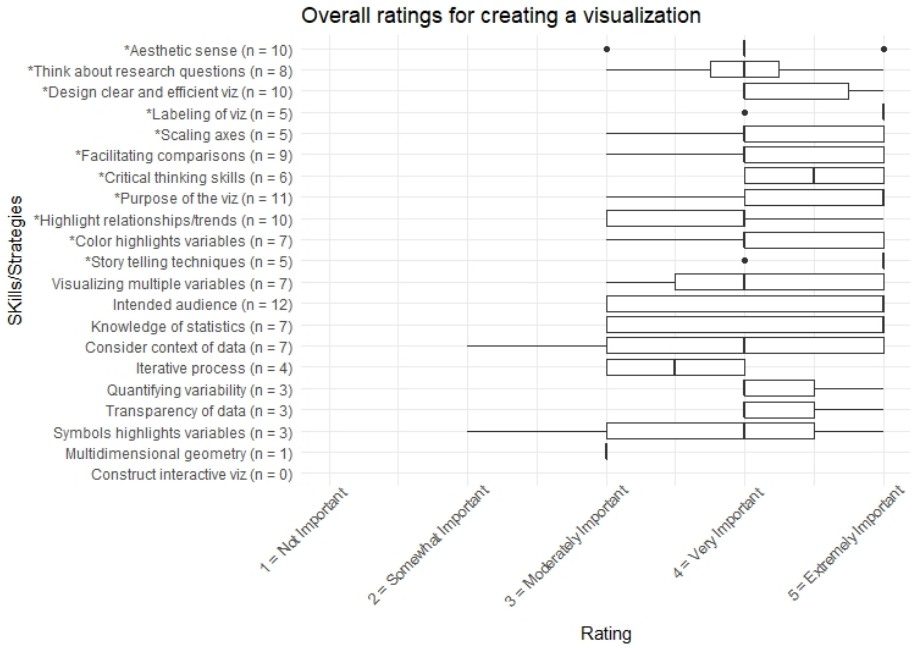


Figure 2. Overall ratings for skills/strategies about creating a visualization

## CONCLUSION

In this Delphi study, the experts identified and agreed upon common skills and strategies for interpreting and creating a visualization. The skills/strategies for data visualization that have been identified through the qualitative analysis and brought back to the panel of experts through three rounds of Delphi panels indicate that the interpretation of visualizations requires complex thinking in terms of understanding the data displayed, making meaning for those data values, and drawing comparisons among variables displayed. The process of creating visualizations requires data scientists to make decisions about how the visualization will be interpreted in terms of facilitating a clear and efficient meaning, highlighting relationships between variables, and defining the intended purpose of the visualization. The results of this study based on these participants provide evidence that there are common skills and strategies for interpreting and creating visualizations across various research fields indicating that data visualization literacy is not domain specific.

A research implication of this study is curriculum design of data visualization literacy materials. The findings of this research study provide a basis for modules that could be designed to

assist undergraduate students in understanding the skills and strategies that are used for efficiently interpreting a visualization and effectively creating a visualization. The skills and strategies identified would be able to provide a starting point for developing a rubric as well that could be used to assess students' interpretations of a visualization and a final visualization that they created. In addition, think aloud interviews with college students about their epistemic cognition when interpreting and creating a visualization could assist in defining the variability in data visualization literacy for students with various levels of experience. Think aloud interviews could also be used to support curriculum design research in providing detailed information about students' epistemic cognition about interpreting and creating visualizations before and after a curriculum intervention focused on data visualizations. The results of these think aloud interviews would help in defining the learning progression of college-level students' data visualization literacy.

The significance of this study was identifying skills/strategies involved in the interpretation and creation of visualizations and achieving consensus among experts from one university. This study provides support for the understanding of what makes an effective data visualization and the thought processes for comprehension to inform curriculum development about teaching data visualization to undergraduate students. Further research is needed to understand effective teaching practices for presenting the skills/strategies of data visualization to students and how to assess students' knowledge and application of the skills/strategies to better understand their learning progression along the spectrum of data visualization literacy.

REFERENCES

Alexander, D. R. (2008). *A Modified Delphi Study of Future Crises and Required Leader Competencies.* (Doctor of Management in Organizational Leadership), University of Phoenix,

Azzam, T., & Evergreen, S. (2013). *Data Visualization, Part 1: New Directions for Evaluation*: Jossey-Bass.

Baker, K. (2005). *A model for leading online K-12 learning environments.* University of Phoenix, Phoenix, Arizona.

Barzilai, S., & Zohar, A. (2016). Epistemic (Meta)cognition: Ways of Thinking about Knowledge and Knowing. In J. A. Greene, W. A. Sandoval, & I. Bråten (Eds.), *Handbook of Epistemic Cognition*. New York: Routledge.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. doi:10.1191/1478088706qp063oa

Chinn, C., Rinehart, R. W., & Buckland, L. A. (2014). Epistemic cognition and evaluating information: Applying the AIR model of epistemic cognition. In D. Rapp & J. Braasch (Eds.), *Processing inaccurate information*. Cambridge, MA: MIT Press.

Figueiras, A. (2013, 16-18 July 2013). *A Typology for Data Visualization on the Web.* Paper presented at the 2013 17th International Conference on Information Visualisation, London.

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi Survey Technique. *Journal of advanced nursing, 32*, 1008-1015. doi:10.1046/j.1365-2648.2000.t01-1-01567.x

Hussein, M. M. (2010). Corporate social responsibility: finding the middle ground. *Social Responsibility Journal, 6*(3), 420-432.

Kapler, T., & Wright, W. (2004). *GeoTime Information Visualization.* Paper presented at the IEEE Symposium on Information Visualization, Washington, DC.

Koparan, T., & Güven, B. (2015). The effect of project-based learning on students' statistical literacy levels for data representation. *International Journal of Mathematical Education in Science & Technology, 46*(5), 658-686. doi:10.1080/0020739X.2014.995242

Linstone, H. A., & Turoff, M. (1975). *The Delphi Method: Techniques and Applications*. Reading, MA: Addison-Wesley Publication Company.

Manizade, A. G., & Mason, M. M. (2011). Using Delphi methodology to design assessments of teachers' pedagogical content knowledge. *Educational Studies in Mathematics, 76*(2), 183-207. doi:10.1007/s10649-010-9276-z

Philip, T. M., Olivares-Pasillas, M. C., & Rocha, J. (2016). Becoming Racially Literate About Data and Data-Literate About Race: Data Visualizations in the Classroom as a Site of Racial-Ideological Micro-Contestations. *Cognition & Instruction, 34*(4), 361-388. doi:10.1080/07370008.2016.1210418

Ritchie, D., & Earnest, J. (1999). The Future of Instructional Design: Results of a Delphi Study. *Educational Technology, 39*(1), 35-42.

Serafini, F. (2012). Expanding the four resources model: reading visual and multi-modal texts. *Pedagogies: An International Journal, 7*(2), 150-164.

Shaughnessy, J. M. (1992). Research in probability and statistics: reflections and directions. In D. A. Grouws (Ed.), *Handbook of research on mathematical teaching and learning* (pp. 465-494). New York, NY: Macmillan.

Takemura, A. (2018). *New Undergraduate Departments and Programs of Data Science in Japan.* Paper presented at the Tenth International Conference on Teaching Statistics, Kyoto, Japan.

Toulmin, S. (1958). *The Uses of Argument*. Cambridge, United Kingdom: Cambridge University Press.

van Zolingen, S. J., & Klaassen, C. A. (2003). Selection processes in a Delphi study about key qualifications in Senior Secondary Vocational Education. *Technological Forecasting and Social Change, 70*(4), 317-340. doi:https://doi.org/10.1016/S0040-1625(02)00202-0