Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Design Principles for Data Analysis

Lucy D'Agostino McGowan, Roger D. Peng & Stephanie C. Hicks

To cite this article: Lucy D'Agostino McGowan, Roger D. Peng & Stephanie C. Hicks (2022): Design Principles for Data Analysis, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2022.2104290

To link to this article: https://doi.org/10.1080/10618600.2022.2104290

Accepted author version posted online: 25 Jul 2022.

Design Principles for Data Analysis

Lucy D’Agostino McGowan
Department of Mathematics and Statistics, Wake Forest University
and
Roger D. Peng
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
and
Stephanie C. Hicks
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

lucydagostino@gmail.com

Abstract

The data revolution has led to an increased interest in the practice of data analysis. While much has been written about statistical thinking, a complementary form of thinking that appears in the practice of data analysis is design thinking – the problem-solving process to understand the people for whom a solution is being designed. For a given problem, there can be significant or subtle differences in how a data analyst (or *producer* of a data analysis) constructs, creates, or designs a data analysis, including differences in the choice of methods, tooling, and workflow. These choices can affect the data analysis products themselves and the experience of the *consumer* of the data analysis. Therefore, the role of a producer can be thought of as designing the data analysis with a set of design principles. Here, we introduce *design principles for data analysis* and describe how they can be mapped to data analyses in a quantitative and informative manner. We also provide data showing variation of principles within and between producers of data analyses. Our work suggests a formal mechanism to describe data analyses based on design principles. These results provide guidance for future work in characterizing the data analytic process.

Keywords: Data science, statistics, education, design

1 Introduction

The data revolution has led to an increased interest in the practice of data analysis (Box, 1976; Chatfield, 1995; Tukey, 1962; Tukey and Wilk, 1966; Wild, 1994; Wild and Pfannkuch, 1999). In the practice of data analysis, one often uses **statistical thinking** (Wild and Pfannkuch, 1999), namely the vague but intuitive process of aiming to accurately describe or understand uncertainties in a complex world using foundations from mathematics, statistics, computer science, psychology, and other fields of study (Snee, 1990; Chance, 2002; Poldrack, 2021). Statistical thinking often manifests where, for a given question or decision that needs to be made, a *producer* of a data analysis makes analytic choices, such as which methods, algorithms, computational tools, languages, or workflows to use in a data analysis that most accurately capture or describe a complex world (Grolemund and Wickham, 2014; Donoho, 2017). For example, a data analysis can consist of simply calculating the sample mean for a given set of observations. Alternatively, the producer may choose to calculate a sample median if they suspect there are outliers in the observed data. A data analysis can also be more complicated consisting of, for example, importing, cleaning, transforming, and modeling data with a goal to build a machine learning algorithm to decide which product a company should sell.

In addition to the goal of describing a complex world accurately through statistical thinking, complementary forms of thinking also appear in the practice of data analysis, including **design thinking** (Cross, 2011; Parker, 2017; Woods, 2019; Nolis and Robinson, 2020). This iterative, solutions-based, problem-solving process aims to understand and to deeply empathize with the people for whom a product is being designed (Cross, 2011). A common practice in design thinking is to employ divergent thinking, or the process of identifying and exploring many solutions (possible or impossible) (Cross, 2021). This is in contrast to convergent thinking, sometimes used in statistical thinking (Wild and Pfannkuch, 1999; Grolemund and Wickham, 2014; Watson and Callingham, 2003; Horton and Hardin, 2015). In convergent thinking the choices, which can be influenced by factors outside the control of the producer such as time or budget constraints (Peng and Parker, 2022) or the availability of appropriate data, are narrowed down to a final solution that is most accurate or correct for given a problem.

In the practice of data analysis, one way divergent thinking often manifests is through a producer of an analysis exploring the design space of (i) how information from the data is extracted, summarized, and presented (Cook and Swayne, 2007; Parker, 2017), (ii) the degree to which evidence in the data is reported or is convincing or the degree to which alternative methods or approaches are considered (Wild and Pfannkuch, 1999; Breiman, 2001), and (iii) how reproducible the data analysis is (Knuth, 1984; Stodden and Miguez, 2014). Ultimately, these design choices for a given data analysis shape the final product that is produced (Nolis and Robinson, 2020). For example, a producer of a data analysis can choose to exhaustively check a set of assumptions of a specific method instead of making a more modest effort. While this design choice often leads to a longer data analysis, it can also lead to different results (or a different interpretation of results) if the assumptions of the method are found to not be supported by the data. Previous empirical studies have found that even when using the same data to investigate the same question, there can be significant variation in how producers build data analyses, which has been shown to influence the results of the analysis (Silberzahn et al., 2018).

These design choices can not only induce variation in the data analyses themselves, but also can affect a *consumer* or ‘stakeholder’ (Nolis and Robinson, 2020) of the data analysis. Using the same example as above, when a producer chooses to exhaustively check a set of assumptions of a specific method, the experience of a consumer of the data analysis (who was expecting an exhaustive analysis) might also be changed from being less confident to more confident as the degree of exhaustively checking the assumptions increases. Alternatively, if a producer makes a design choice to summarize the results from a data analysis with only tables, then a consumer (who was expecting summaries with plots) might not understand the results without data visualizations, and therefore be skeptical of any results.

We refer to factors or characteristics that are relevant to the production of a data analysis, as a whole or its individual components, as *design principles* for data analysis. Broadly, when building a data analysis, the role of a producer can be thought of as designing the data analysis with a set of data analytic principles to serve a larger purpose, such as to be able to extract meaningful information, answer an original question, support decision-making, or address the needs or expectations of data analysis consumers. Similar to principles of art or music (Lambert, 2014), the design principles for data analysis are not meant to be used to evaluate the quality of a data analysis, but rather they are meant to be characteristics about the data analysis that can be used to induce or describe variation between data analyses.

Our primary focus in this manuscript is to (i) introduce a set of data analytic design principles (Section 2), (ii) describe an example of how the design principles can be used to measure different characteristics of a data analysis (Section 3), and (iii) present data on the variation in principles within and between producers of data analyses (Section 3). In the Discussion (Section 4), we discuss how these data analytic design principles can be implemented in practice, for example, how the design principles can be used in the classroom by practitioner-instructors (Kross and Guo, 2019) to build data analyses.

2 Design principles for data analysis

The design principles for data analysis are qualities or characteristics that are relevant to the analysis and can be observed or measured. Driven by statistical thinking and design thinking, a data analyst can use these principles to guide the choice of which data analytic *elements* to use, such as code, code comments, data visualization, non-data visualization, narrative text, summary statistics, tables, and statistical models or computational algorithms (Breiman, 2001), to build a data analysis. Briefly, the elements of an analysis are the individual basic components of the analysis that, when assembled together by the analyst, make up the entire analysis. A data analysis can be *scored* based on how well it adheres to each of these principles. The scoring is not meant to convey a value judgment with respect to the overall quality of the data analysis. Value judgments may be overlaid on to an analysis by the consumer based on how different principles are scored, but we do not consider such judgments universal characteristics. Next, we describe six principles that we hypothesize are informative for characterizing variation between data analyses.

Data Matching. Data analyses with high *data matching* have data readily measured or available to the producer that directly match the data needed to investigate a question (Figure S1). In contrast, a question may concern quantities that cannot be directly measured or are not available to the producer. In this case, data matched to the question may be surrogates for covariates that measure the underlying data phenomena. While we consider the main question and the data to be contextual inputs to the data analysis, we consider this a design principle of data analysis because the producer selects methods, tooling, or workflows that are used to investigate the question, which depend on how well the data are matched. If the data are poorly matched, the producer will not only need to investigate the main question with one set of methods, but also will need to use additional methods that describe how well the surrogate data are related to the underlying data phenomena.

It is important to note that questions can be more or less specific, which will impose strong or weak constraints on the range of data matching to the question. Highly specific questions tend to induce strong constraints to investigate which methods, tooling, or workflows are used. Less specific questions emit a large range of potential data to investigate the question. Data that can be readily measured or are available to the producer to directly address a specific question results in high data matching, but depending on the problem specificity, can result in a narrow or broad set of data to consider.

Exhaustive. An analysis is *exhaustive* if specific questions are addressed using multiple, complementary methods, tooling, or workflows (Figure S2). For example, using a scatter plot and a correlation coefficient are two different tools that could be employed to investigate whether two predictors are associated. Analyses that are exhaustive use many methods to address the same question, knowing that each given tool reveals some aspects of the data, but obscures other aspects. As a result, the combination of tools and methods used may provide a more complete picture of the evidence in the data than any single tool would. For example, a non-randomized study comparing two groups may make the groups comparable using two different methods: matching and weighting. Matching often allows for a straightforward comparison for clinicians to understand, since it is easy to conceptualize; whole individuals are either included if they have a match in the opposite group, or they are not. A clinician seeing this may be more likely to "trust" the analysis because they can better understand the process by which individuals are either explicitly included or excluded. Weighting may benefit the statistically minded, such that there is efficiency as no subjects are excluded. An exhaustive analysis may include both methods.

Skeptical. An analysis is *skeptical* if multiple, related questions are considered using the same data (Figure S3). Analyses, to varying extents, consider alternative explanations of observed phenomena and evaluate the consistency of the data with these alternative explanations. Analyses that do not consider alternate explanations have no skepticism. For example, to examine the relationship between a predictor X and an outcome Y, an analysis may choose to show results from different models containing different sets of predictors that might potentially confound that relationship. Each of these different models represents a different, but related, question about the X-Y relationship. A separate question that arises is whether the configuration of alternative explanations are relevant to the problem at hand. However, often that question can only be resolved using contextual information that is outside the data.

The need for more or less skepticism in a data analysis is typically governed by outside circumstances and the context in which the analysis sits. Analyses that may have large impacts or result in significant monetary costs will typically be subject to detailed scrutiny. In July 2000, the Health Effects Institute (HEI) published a reanalysis of the Harvard Six Cities Study, a seminal air pollution study that showed significant associations between air pollution and mortality. Due to the potential regulatory impact of the study, HEI commissioned an independent set of investigators to reproduce the findings and conduct a series of sensitivity analyses (Krewski et al., 2000). The result was a nearly 300 page volume where the data and findings were subject to intense skepticism and every alternative hypothesis was examined.

There are other instances when skepticism in the form of alternate explanations is not warranted in the analysis. For example, with an explicitly planned and rigorously-conducted clinical trial, the reported analysis will typically reflect only what was pre-specified in the trial protocol. Other data analytic elements or analyses may be presented in a paper, but they will be explicitly labeled as secondary. For example, in a large clinical trial studying the effect of a pest management intervention on asthma outcomes (Matsui et al., 2017), the reported analysis is ultimately a simple comparison of asthma symptoms in two groups. Some other secondary analyses are presented, but they do not directly address the primary question. Such an analysis is acceptable here due to the strict pre-specification of the analysis and due to the standards and practices that the community has developed regarding the reporting of clinical trials.

Second-Order. An analysis is *second-order* if it includes methods, tooling, or workflows that do not directly address the primary question, but give important context or supporting information to the analysis (Figure S4). Any given analysis will contain, for example, data visualizations that directly contribute to the results or conclusions and serve as key pieces of evidence. However, many analyses will also contain other information or data that provide background or context or are needed for other reasons (Figure S1). Second-order analyses contain more of these background and contextual elements in the analysis, for better or for worse. For example, in presenting an analysis of data collected from a new type of machine, one may include details of who manufactured the machine, why it was built, or how it operates. Often, in studies where data are collected in the field, such as in people's homes, field workers can relay important details about the circumstances under which the data were collected. Clinical studies might report information about patient intake forms. In each of these examples, these details may be of interest and provide useful background, but they are not considered primary data and may not directly influence the analysis itself. Rather, they may play a role in helping a consumer interpret the results and evaluate the strength of the evidence.

An analysis that is highly second-order does not necessarily include many secondary *analyses*. For example, an analysis may employ a machine learning algorithm and include a secondary analysis that shows the sensitivity of the results to different tuning parameter settings. Such a secondary *analysis* may play a direct role in interpreting the strength of the evidence in the primary analysis. Second-order information is less directly connected to the primary analysis and likely will not use any of the data under consideration. That said, second-order information could end up playing an important role should a data analysis produce a result that is highly unexpected. For example, in a clinical study, an unexpected data analytic result could be related to the way a patient intake form is formatted. Hence, it can sometimes be useful to have such second-order information presented. Another example would be a paper that includes a table showing how different health outcomes were derived from diagnosis and procedure codes available in administrative data (Roumie et al., 2017).

Clarity. Analyses with *clarity* summarize or visualize key pieces of evidence in the data that explain the most "variation" or are most influential to understanding the key results or conclusions (Figure S5). Clarity could be demonstrated by simply including one data visualization, or it could consist of multiple data visualizations. For example, having a data visualization that draws attention to the key message of the narrative of the analysis can focus the reader on how the data are connected to the results. Clarity can also be represented by presentations that highlight the data generation process and any uncertainties or biases introduced by that process.

In a study of 8,111 people in six industrial cities in the midwestern United States, researchers found a strong association between ambient air pollution and mortality (Dockery et al., 1993). This study examined numerous individual-level factors, assembled large datasets, and employed complex modeling to estimate the key association. Ultimately, the relationship between fine particulate matter and mortality was summarized in a single plot containing six data points in a line, demonstrating a strong linear association between the two factors (Figure S6). Distilling the complexity of the analysis into a simple scatter plot incorporating all of the relevant information is a demonstration of clarity because it summarizes the main result of the paper, while also providing some indication of the strength of the evidence.

Reproducible. An analysis is *reproducible* if someone who is not the original producer can take the published code and data and compute the same results as the original producer (Figure S7). Critical to reproducibility is the availability of a stable form for both the dataset and the analytic code. For example, an analysis might consist of interactively calculating a sample mean for a given set of observations in the console of a programming language. In this case, as there is no stable code because the analysis was performed interactively, it is not possible for another person to reproduce the analysis. These types of analyses that are not deliberately recorded happen frequently and do not necessarily imply a negative quality about the analysis. Rather such analyses are simply not reproducible. In contrast, analyses that integrate literate programming (Knuth, 1984) in an analytic compendium are more reproducible (Vassilev et al., 2016). Another consideration is that it may not be possible for businesses, such as those in the finance industry, to make available entire analytic compendia for proprietary or financial reasons. In contrast, analytic compendia that are integrated as part of the analytic product or analytic presentation are by definition more reproducible. Finally, much has been written about reproducibility and its inherent importance in science, so we do not repeat that here (Peng, 2011). We simply add that reproducibility (or lack thereof) is usually easily verified and is not dependent on the characteristics of the consumer of the analysis. Reproducibility also speaks to the coherence of the workflow in the analysis in that the workflow should show how the data are transformed to eventually become results.

3 Results

In this section, we describe two case studies exploring the variation in the principles across different data analyses. Our approach to this analysis was exploratory with the goal of summarizing the between person and within person variation across the design principles. These two case studies consist of longitudinal data collected at Wake Forest University and cross-sectional data collected at Johns Hopkins University.

3.1 Wake Forest University Case Study

The data for this case study were collected from 54 students enrolled in a Statistical Models course at Wake Forest University. This course is intended for students who have had at least one university-level statistics course. The study was approved by the Wake Forest University Institutional Review Board (IRB00023932).

Participants were taught the 6 design principles of data analysis. Throughout the course, they were given 8 data analysis assignments. These assignments consisted of completing data analytic tasks in R and had varying degrees of adherence to each of the design principles (the assignments can be found in the Supplementary Material). On each of the 8 assignments, students were asked to self-rank the analysis they completed from 1 to 10 across each of the principles, with one indicating that the analysis did not adhere to the principle, and 10 indicating that it did. We also collected data on the participants' current major.

Figure 1 shows each individual's score of the six principles across the 8 analyses. Two "profiles" are selected for demonstration purposes to illustrate both the variability in scores within a given individual across assignments and between principles, as well as the variability between individuals. For example, examining the exhaustive and skeptical principles, subjects 3 and 4 were relatively similar on assignments 2-7 and differed slightly on assignments 1 and 8. Subject 4 has less variability between assignments in both matching and reproducibility compared to subject 3, consistently giving scores of 10. On the other hand, subject 4 has more variability in clarity compared to subject 3. The second-order principle follows a similar pattern for both subjects across assignments, but at different levels.

An interesting feature of Figure 1 is the between-subject variation present in some of the principles. One possible explanation for this variability in principle scores is that these principles are related to or influenced by an individual's baseline characteristics. As an illustrative example, we examine the average principle score by the producers' declared major (Figure 2). There appears to be some variation in the mean and variability of principles by major, suggesting that major may be relevant, however the sample size is too small to draw any meaningful conclusions. For example, looking at the clarity principle, students majoring in the social sciences all had high average scores, differing from the other majors where more variation was present.

Figure 3 shows pairwise scatterplots between principles for one of the assignments; while some principle scores are more highly correlated than others, there appears to be appreciable variability, indicating that these principle scores measure different underlying characteristics of a data analysis. This can be better visualized by Supplemental Figure S8. The cumulative proportion of variance explained by principle components illustrates that not all principles are loaded in a single component, again suggesting that there is additional information added across the six principles.

3.2 Johns Hopkins University Case Study

Data from Johns Hopkins University were collected from 15 students enrolled in a graduate course titled Advanced Data Science. For a homework assignment, students were asked to score a data analysis completed by two separate authors (not any of the students in the course) using a score of 1 to 10 for each of the principles described in Section 2. The data analyses consisted of analyses of natural disasters in the United States and their economic impact. Each analysis was done by a different person, but the datasets and question addressed were the same. The students were given the output of the data analyses, but were not given the data and were not asked to analyze the data themselves. Stable links to the analyses the students were asked to evaluate are provided in the Supplementary Material. This study was approved by the Institutional Review Board of the Johns Hopkins Bloomberg School of Public Health (IRB 00012419).

When considering data from the participants at Johns Hopkins University, we see some differences across the two analyses they were asked to score, as expected (Figure 4). The differences in the scores given by the students for the two analyses shown in Figure 4 indicate that while the two analyses are similar on some principles, such as skepticism, exhaustive, and clarity, they differ somewhat on reproducibility, second-order, and data matching, albeit with wide variation. This pattern suggests that there is some variation in the analyses attributable to the choices of the analyst, but in this case only among a subset of principles.

3.3 Summary of Results

The two case studies described above provide some evidence that the principle scores assigned to different data analyses exhibit variation across individuals and across data analyses. Data from Wake Forest suggest that variation across individuals is greater than variation within individuals and therefore exploring the sources of this variation would be of keen interest. Figure 2 suggests that this variation may be associated with a student’s major, but more in-depth work needs to be done to draw a connection between principle scores and individual characteristics. Figures 3 and S8 demonstrate that these principles are measuring distinct underlying characteristics of a data analysis. The data from Johns Hopkins indicate that the scoring of principles has some ability to distinguish between analyses done by independent analysts. While there was variability across individuals in the scoring, a subset of principle scores were distinctly different on average between the analyses.

4 Discussion

In this paper, we introduce a set of data analytic design principles with the goal of describing variation between data analyses. These principles are characteristics of the data analysis made by the producer considering the needs of the consumer used in the application of design thinking (and complementary to statistical thinking) in the practice of data analysis. We also illustrated the use of these data analytic principles in two classroom settings aimed at teaching data analysis. In our analysis of the classroom data, we found that the principles defined here appear to measure underlying quantities that are reasonably uncorrelated with each other (Figures 3-S8).

The data from the Wake Forest and Johns Hopkins studies presented in Section 3 suggest that there is variation in principle scores across individuals working on the same data analysis task. While we would hypothesize that this variation in scores is attributable to the choices made by the individual analysts, the data suggest that an interesting avenue for future work would be to design studies to identify specific characteristics or qualities on the individuals that explain the variation in scores. The data on student majors in Figure 2 only hints at such an explanation.

One significant consequence of using design thinking concepts in data analysis is that it allows for the explicit separation of producers and consumers of a data analysis. The benefit of conceptually separating producers from consumers is that such a separation serves to demonstrate potential differences in priorities between the two groups. Traditional descriptions of statistical thinking generally conceive of a single analyst building data analyses and obtaining feedback on their approach from the data. There is frequently an iterative process, by which the data available informs the question able to be answered. While the notion of a consumer for that analysis may be embedded in the idea of statistical thinking, it is often not well-specified.

In general, consumers of data analyses will have certain expectations for what they see and data analysts (producers) can construct an analysis that either meets those expectations or not. One possible way to quantify a consumer's expectations for a data analysis is to assign *a priori* weights to each of the design principles described here. The distance between a consumer's weights on these principles and the scores assigned to the realized data analysis could indicate the extent to which the analysis meets the consumer's expectations. Producers may also assign *a priori* weights to the different principles that can guide the construction of an analysis. If a producer's and consumer's weights are known to be substantially different from each other, then this would be an *a priori* indication that the analysis may not meet the consumer's expectations. In such a situation, it may be valuable for the producer and the consumer to come to an agreement over the weighting of the principles before time is spent doing the analysis. Here, the design principles can provide a formal or informal articulation of how producer and consumer can agree (or disagree) on the ultimate outcome before seeing the data.

Building an analysis that satisfies the wants and needs of a consumer is an important requirement for a data analysis and draws upon skills that are rarely discussed in the statistical literature. An analyst must assess the background and priorities of the consumer and tailor the analysis accordingly, all while maintaining rigor in the application of statistical methodology. Experience working with different people can guide the analyst to build a successful analysis for a given consumer. However, more discussion could be had surrounding the approaches to take when negotiating analytic priorities and how we can train new analysts to do this well. We hope to address this further in future work.

The design thinking perspective on data analysis also has useful consequences for teaching data analysis in the classroom, where it is valuable to have a way to describe what makes data analyses differ from each other and why one type of analysis might be preferable in some circumstances to another type of analysis. In particular, in teaching about the divergent thinking phase of data analysis, it is common to encourage students to take different approaches to addressing a data analytic question. However, we often lack a structured basis for characterizing these different approaches for students. The data analytic design principles provide one way to separate different approaches and to guide students to explore various approaches to problem solving. Another useful consequence of bringing design thinking to the practice of data analysis is that it opens the door to bringing over many design concepts to the data analysis fields. Considering data analyses as designed objects raises questions about what their requirements are, whether they satisfy those requirements, and how we might build a framework for specifying and verifying data analyses.

These principles allow us to more specifically describe a given analysis and how it differs from what one may ideally prefer. For example, during an initial distribution of a vaccine in a population, quantifying how effective the vaccine is at preventing hospitalization in practice is of interest to the scientific community and general public. In order to calculate this, one would need data from all exposed individuals including their vaccine status, characteristics that would make them more or less likely to be hospitalized at baseline, and whether they were hospitalized. In practice, hospitals do not have information on all exposed individuals; rather they only have data on those who are currently hospitalized. A data analysis completed by a hospital may report the percentage of all hospitalized individuals that were vaccinated. While this does not indicate how effective the vaccine is, an astute practitioner may be able to glean some useful information from this statistic by inferring the prevalence of vaccinations in the general population. Having terminology to describe why this analysis offered by the hospital, while trying to answer a specific question, may not be perfectly suited to answer a question about vaccine effectiveness, i.e. the extent to which the data match a question of interest, could be useful.

Concepts from design thinking can serve as important complements to the traditional notion of statistical thinking. Together, these two forms of thinking provide a more complete road map for developing useful data analyses and present new ways to teach data analyses to novices. The specification of design principles for data analysis and how they may guide data analysis construction offers a rationale for negotiating qualities of a data analysis between producer and consumer before embarking on substantial data analytic work. An area for possible future work includes measuring to what extent manipulating the weighting of these principles can improve the quality of data analysis.

References

G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356): 791–799, 1976.

C. Chatfield. *Problem solving: a statistician's guide*. Chapman and Hall/CRC, 1995.

John W Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67, 1962.

W. Tukey and M. B. Wilk. Data analysis and statistics: an expository overview. In *In Proceedings of the November 7-10, 1966, fall joint computer conference*, pages 695–709, 1966.

C. J. Wild. Embracing the "wider view" of statistics. *The American Statistician*, 48(2):163–171, 1994.

C. J Wild and M. Pfannkuch. Statistical thinking in empirical enquiry. *International Statistical Review/Revue Internationale de Statistique*, 1999.

Ronald D. Snee. Statistical thinking and its contribution to total quality. *The American Statistician*, 44(2):116–121, 1990. ISSN 00031305. URL http://www.jstor.org/stable/2684144.

Beth L. Chance. Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, 10(3):null, 2002. doi: 10.1080/10691898.2002.11910677.

Russell A. Poldrack. *Statistical Thinking for the 21st Century*. Stanford University, 1 2021. URL https://statsthinking21.github.io/statsthinking21-core-site. [Online; accessed 2021-03-04].

Garrett Grolemund and Hadley Wickham. A cognitive interpretation of data analysis. *International Statistical Review*, 82(2):184–204, 2014.

David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26 (4):745–766, 2017. doi: 10.1080/10618600.2017.1384734. URL https://doi.org/10.1080/10618600.2017.1384734.

Nigel Cross. *Design Thinking: Understanding How Designers Think and Work*. Berg Publishers, Oxford, April 2011.

Hilary Parker. Opinionated analysis development. *PeerJ Preprints*, 5:e3210v1, 08 2017. ISSN 2167-9843. doi: 10.7287/peerj.preprints.3210v1. URL https://doi.org/10.7287/peerj.preprints.3210v1.

Rachel Woods. A design thinking mindset for data science, Mar 2019. URL https://towardsdatascience.com/a-design-thinking-mindset-for-data-science-f94f1e27f90.

Jacqueline Nolis and Emily Robinson. *Build a Career in Data Science*. Manning Publications, 1 edition, 2020. ISBN 978-1-61729-624-6. URL http://gen.lib.rus.ec/book/index.php?md5=7AEAD73047B0807179C0A505C10173E2.

Nigel Cross. *Engineering Design Methods: Strategies for Product Design (5th ed.)*. John Wiley & Sons, Chichester, 03 2021. URL http://oro.open.ac.uk/39439/.

J. Watson and R. Callingham. Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, 2(2):3–46, 2003.

Nicholas J. Horton and Johanna S. Hardin. Teaching the Next Generation of Statistics Students to ”Think With Data”: Special Issue on Statistics and the Undergraduate Curriculum. *The American Statistician*, 69(4):259–265, 2015. doi: 10.1080/00031305.2015.1094283. URL https://doi.org/10.1080/00031305.2015.1094283.

Roger D Peng and Hilary S Parker. Perspective on data science. *Annual Review of Statistics and Its Application*, 9:1–20, 2022.

D. Cook and D. F. Swayne. *Interactive and dynamic graphics for data analysis with R and GGobi*. Springer Publishing Company, Incorporated, 2007.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215, 2001. ISSN 08834237. URL http://www.jstor.org/stable/2676681.

Donald E. Knuth. Literate programming. *The Computer Journal*, 27:97–111, 1984.

Victoria Stodden and Sheila Miguez. Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research. *Journal of Open Research Software*, 2(1):e21, 2014. doi: http://doi.org/10.5334/jors.ay. URL https://towardsdatascience.com/a-design-thinking-mindset-for-data-science-f94f1e27f90.

R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, A. BahnÃk, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. HÃgden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope,

B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. SchlÃŒter, F. D. SchÃnbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. SpÃrlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 2018. doi: 10.1177/2515245917747646.

Philip Lambert. *Principles of Music*. New York: Oxford University Press, 2014.

Sean Kross and Philip J. Guo. Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300493. URL https://doi.org/10.1145/3290605.3300493.

D Krewski, RT Burnett, MS Goldberg, K Hoover, J Siemiatycki, M Jerrett, M Abrahamowicz, and WH White. *Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality*. The Health Effects Institute, Cambridge MA., 2000.

Elizabeth C Matsui, Matthew Perzanowski, Roger D Peng, Robert A Wise, Susan Balcer-Whaley, Michelle Newman, Amparito Cunningham, Adnan Divjan, Mary E Bollinger, Shuyan Zhai, et al. Effect of an integrated pest management intervention on asthma symptoms among mouse-sensitized children and adolescents with asthma: a randomized clinical trial. *Jama*, 317 (10):1027–1036, 2017.

Christianne L Roumie, Jea Young Min, Lucy D'Agostino McGowan, Caroline Presley, Carlos G Grijalva, Amber J Hackstadt, Adriana M Hung, Robert A Greevy, Tom Elasy, and Marie R Griffin. Comparative safety of sulfonylurea and metformin monotherapy on the risk of heart failure: a cohort study. *Journal of the American Heart Association*, 6(4):e005379, 2017.

Douglas W Dockery, C Arden Pope, Xiping Xu, John D Spengler, James H Ware, Martha E Fay, Benjamin G Ferris Jr, and Frank E Speizer. An association between air pollution and mortality in six us cities. *New England journal of medicine*, 329(24):1753–1759, 1993.

Boris Vassilev, Riku Louhimo, Elina Ikonen, and Sampsa Hautaniemi. Language-agnostic reproducible data analysis using literate programming. *PLoS One*, 11(10):e0164023, 2016. doi: 10.1371/journal.pone.0164023.

R. D. Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 12 2011.

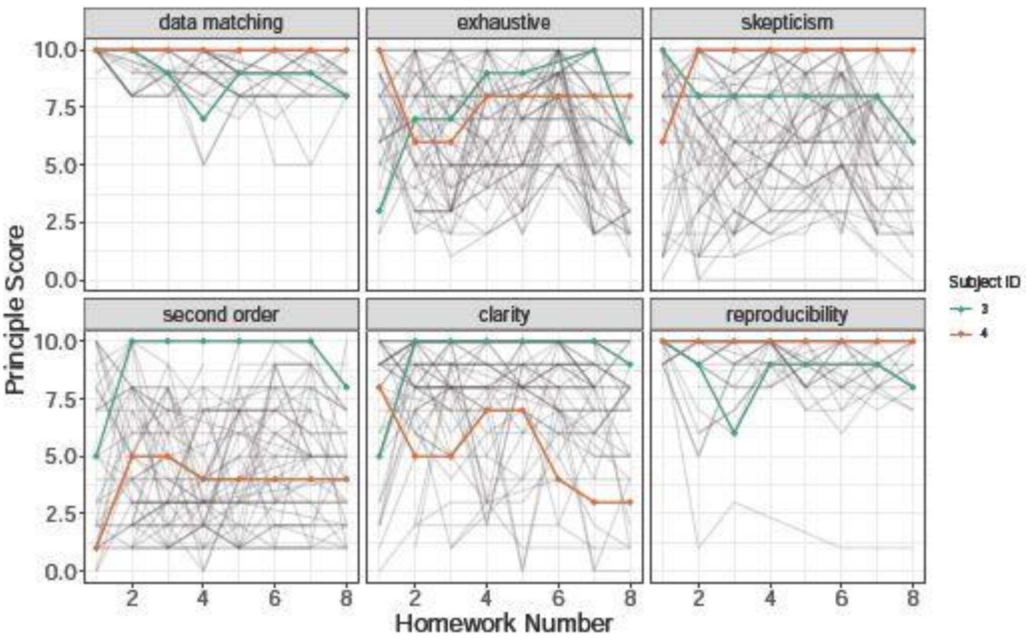

Fig. 1 **Between and within person variation of principles across assignments.** For a given homework assignment (*x*-axis), individuals scored the assignment from 1 to 10 for each data analytic principle (*y*-axis), with one indicating that the analysis did not adhere to the principle, and 10 indicating that it did. Two individuals (Subject ID 3 and 4) are highlighted in green and orange to illustrate both the variability in scores within a given individual across assignments and between principles, as well as the variability between individuals.

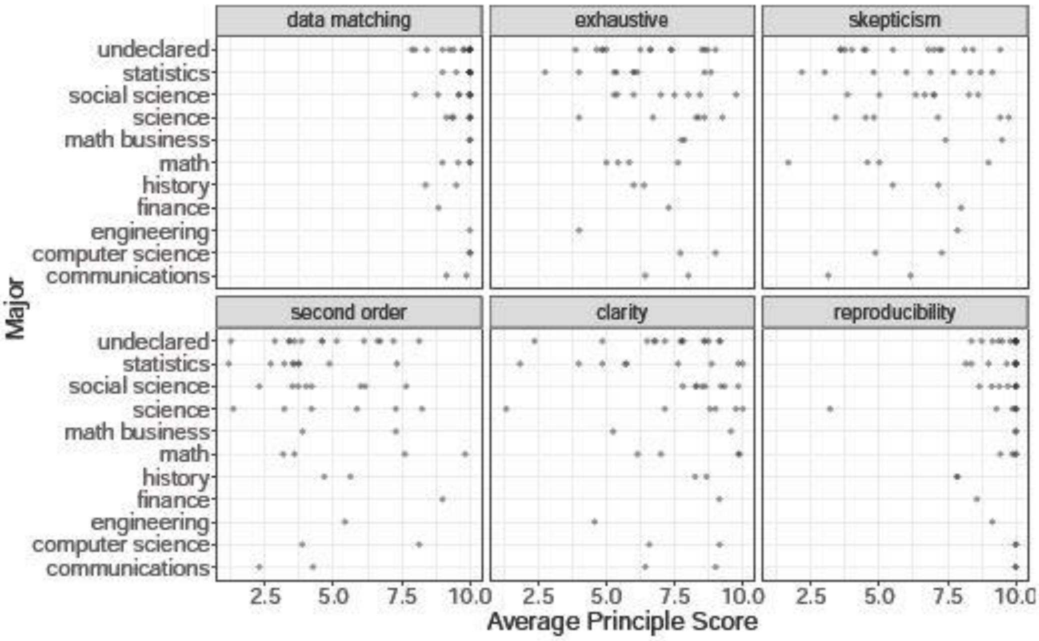

Fig. 2 **Average principle scores by declared major.** For a given major (*y*-axis), we show the scores (averaged across analyses) from the data analytic principles from individuals who declared that major (*x*-axis). The average scores are shown as faceted plots by the six principles. The numbers of students in each major are communications (2), computer science (2), engineering (1), finance (1), history (2), math (4), math business (2), science (6), social science (8), statistics (9), undeclared (14).

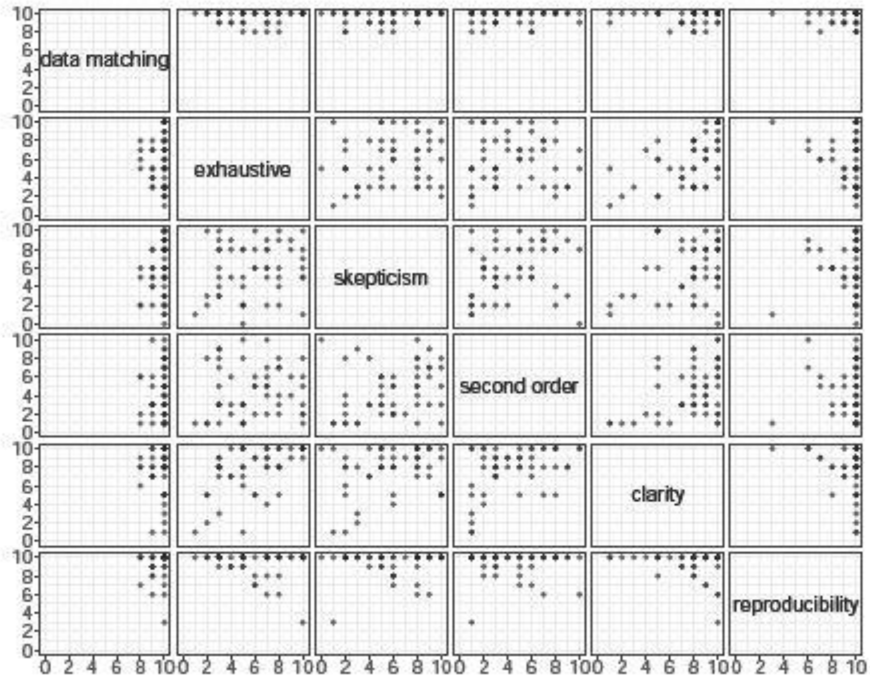

Fig. 3 **Pairwise relationship between six data analytic principles.** Using the Wake Forest data, we subset to a single assignment (Assignment 3, n = 50) and for each pair of principles, we show a pairwise scatter plot to illustrate the (dis)agreement between each principle.

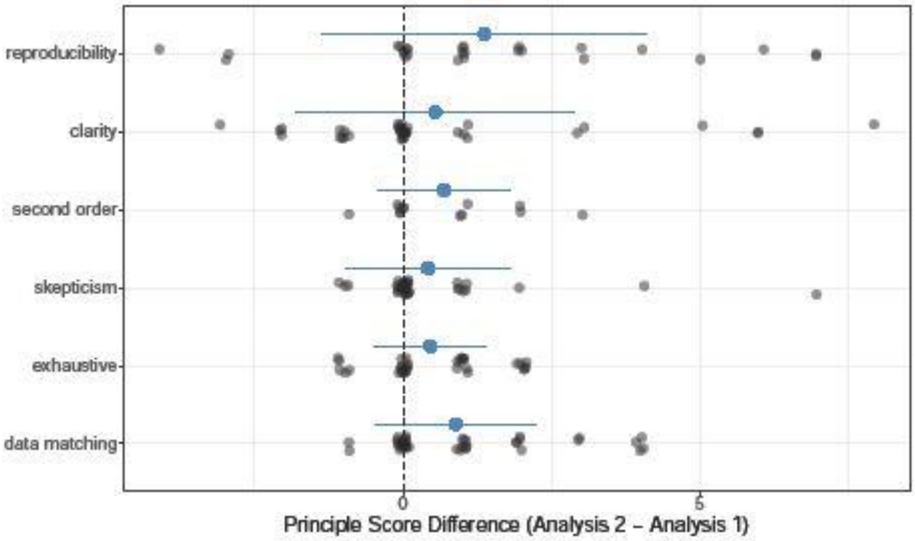

Fig. 4 **Differences in principle scores between two data analyses, by principle.** The score differences for the two data analyses (*x*-axis) for different principles (*y*-axis). In blue are the mean difference ± 1 standard deviation.