Data Management Plan

Expected Data Types Data used to generate research stimuli may be a product of pre-existing data sets or simulation; in either case, the data necessary to generate the stimuli will be made available in a standard text format (CSV or JSON). In addition, research stimuli will be separately saved as standard image files (JPG, PNG, SVG, PDF) in the case of two-dimensional stimuli or STL files in the case of three-dimensional stimuli. Three-dimensional stimuli will also be accompanied by a markdown file describing the software, settings, and hardware used to convert the STL into a physical object; these are theoretically not necessary to reproduce the experiment as STL is a standard, open format, but will provide an additional measure of documentation in case the file format changes in the future. Experiments planned as part of this project will collect de-identified data by design, and only de-identified data will be included in version control repositories. Version control repositories are used to ensure that the provenance of the data is recorded and publicly available while preserving the anonymity of participants as much as possible. Identifiable records (such as those connecting names to participation) required by IRB or institutional requirements to maintain payment records will be stored securely and separately from experimental data. Participant data will be saved digitally in either plain-text or as SQL/SQLite databases and will be provided for public use in standard text formats (CSV, JSON).

This project will also generate curricula for use in introductory statistics courses as well as graduate courses for middle and high school educators. These resources will be created in plain text documents (using literate programming formats such as Quarto documents, with embedded code and results) and then rendered to the published format (Docx, PDF, HTML). Teaching and presenter notes will be included in the plain text documents as comments to facilitate use and adaptation of the generated materials as well as to maintain a record of the effectiveness of the materials over time. All curricula will be published under a CC-by 4.0 license.

Software Products This program will generate R packages supporting the development of multimodal graphics experiments as well as the analysis of multimodal experiment data. All packages will be submitted to CRAN (The Comprehensive R Archive Network), ensuring that they are available to the community and meet minimal standards for documentation and security. Packages will make use of `pkgdown`, which creates documentation websites that can be hosted on GitHub pages, and each package will include multiple vignettes demonstrating how the package should be used. As the project experiments conclude, each experiment will be used to create a separate vignette demonstrating data analysis functions, in order to ensure that there are several full analyses in the package documentation, along with standard examples which document core features of each function. We will also leverage `testthat`, an R package for unit testing, to ensure that all code is adequately tested. Where possible, packages will conform to tidy standards for package development. Development versions of all packages will be made available on GitHub.

Data formatting standards All research products (stimuli generation code and results, data collection procedures, de-identified data, data analysis code, literate programming manuscripts) will be made available online via GitHub or a similar version control service, with accompanying documentation describing the environment necessary to run the code, instructions to generate the results presented in the manuscript. A markdown or plain text README file will be provided at the top level to indicate authorship, provide author contact information, acknowledge sources of support, describe the folder organization and file structure, and explain how to set up a software environment which is capable of running the provided code. Data files will additionally be accompanied by a markdown formatted data dictionary explaining the format, variables, units, and any relevant pre-processing procedures used for e.g. de-identification.

Period of data retention This project will conform to UNL data management policies with regard to the necessity of data sharing agreements for personally identifiable information, however, the experiments

in this project are carefully designed in order to ensure that the de-identified data is sufficient to computationally replicate the results of all experiments; identifiable information is only collected in order to e.g. ensure compliance with subject payment documentation requirements, and the key connecting responses to these records will be destroyed as soon as allowed by UNL records retention policies (3 years after the study is closed). All de-identified data will be tracked using a version control program and mirrored to a remote repository automatically during data collection; an additional copy of the data will be maintained on UNL servers in compliance with university data security policies. This process ensures transparency and guards against data deletion: at any time, the data is stored in at least 5 separate physical locations, with at least two separate cloud services providing backups. Once experimental results have been submitted for publication, the version control repository containing data, analysis code, and manuscript (written using literate programming), along with README and documentation files, will be made publicly visible. As part of the publication process, a snapshot of the experimental stimuli, data collection procedures, de-identified data, and code will be uploaded to the UNL Data Repository, which ensures preservation for a minimum of 20 years and provides a DOI for the supplementary material. This complies with the supplementary material requirements of target journals for project publications (Journal of Computational and Graphical Statistics, IEEE Transactions on Visualization and Computer Graphics, R Journal, Journal of Statistical Software), in addition to good practices for reproducible computing and data analysis.

Policies for dissemination and data-sharing

Licensing We will license all collected data and generated curricula under a Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license, which encourages re-use, re-distribution, and creation of derivatives.

We will license all software created during this process using an open license chosen to conform to the conditions of any software dependencies, with GPL-3.0 being the preferred license in absence of dependency restrictions. All production code written by the PI and any students associated with the project will be made available on GitHub or other version control platforms for reproducibility and reference purposes.

Dissemination Targeted journals for the results of this project (Journal of Computational and Graphical Statistics, IEEE Transactions on Visualization and Computer Graphics, R Journal, Journal of Statistical Software) provide downloadable PDF copies of manuscripts on the web. All targeted journals also require reproducible code and data be provided along with the published manuscript; this requirement is typically satisfied by providing a GitHub repository or persistent university data storage library. As part of the publication process, a snapshot of the experimental stimuli, data collection procedures, de-identified data, and code will be uploaded to the UNL Data Repository, which ensures preservation for a minimum of 20 years and provides a DOI for the supplementary material.

We will also submit conference papers to Computer-Human Interaction and InfoVis conferences as appropriate, and will meet the same standards for reproducibility detailed above for journal publications.

Curricula created as part of this project will be made available on GitHub and published to the web. Snapshots of developed materials will be uploaded to UNL Data Repository or another persistent data storage site such as FigShare upon publication of papers describing and evaluating the success of the pedagogical approaches taken.