Learning by teaching program tracing using virtual pedagogical agents: The role of non-cognitive factors (Dalhousie University Generic Plan)

A Data Management Plan created using DMP Assistant

Creator: Eric Poitras

Affiliation: Dalhousie University

Funder: Social Sciences and Humanities Research Council of Canada

Template: Dalhousie University Generic Plan

ORCID iD: 0000-0003-0563-7978

Grant number: N/A

Project abstract:

Students in large-scale introductory programming course acquire knowledge by studying worked examples and solving practice problems with the benefit of feedback. Solving a programming problem requires conceptual knowledge of programming but students who face difficulties to reinstate it from memory are most likely to fail in their efforts to trace and infer program states during execution (Qian & Lehman, 2017; Sorva, 2013). Previous research in the context of intelligent programming tutors has addressed this issue by relying on visual representations such as memory diagrams and tables to scaffold program tracing (Sorva, Karavirta, & Malmi, 2013). However, students may still fail to build a valid mental model of the program and retain the requisite knowledge without feedback and self-explaining certain aspects of the algorithms (Xie et al., 2019). This design-based study examines the effects of intelligent programming tutors designed according to a learning by teaching paradigm, where students are prompted to self-explain and receive feedback while teaching virtual pedagogical agents playing the roles of tutors and tutee. The participants in this study include undergraduate students recruited from CSCI 1105 Introduction to Computer Programming during the Fall 2021 term. The task consists of studying worked examples interleaved with a set of practice problems under one of three different experimental conditions, where students either (1) learn by teaching a virtual pedagogical agent playing the role of a student to trace the execution of a program (i.e., learning by teaching with open-ended learner model), (2) trace the execution of the program without an agent but with a skills meter that shows their progress (i.e., learning with open-ended learner model), or (3) trace the execution of the program without a skills meter (i.e., learning without open-ended learner model). In each condition, an agent playing the role of a tutor prompts them to self-explain, provides hints on request, and delivers feedback. Learning processes are measured by logging in the system student responses, hint requests, self-explanations, attempts, and elapsed time in addition to self-reported experiences and interactions with agents featured in the system (e.g., self-efficacy, cognitive load, value/enjoyment, effort/importance, and interest/enjoyment). Learning outcomes are captured through practice problems delivered after the learning session that differ in terms of (1) values and cover story (i.e., isomorphic and contextual properties), (2) procedure required to attain the task goal (i.e., procedural properties), and (3) performance on quiz/assignments delivered in the course (i.e., temporal and functional properties). The specific objectives of this research are two-fold: to broaden theory and model the cognitive and metacognitive components of student self-regulated learning in computing

contexts by examining non-cognitive factors, and to provide empirical foundations from which to test the effectiveness of visual representations to improve self-regulatory skills. We expect our findings to broaden our understanding of self-regulatory processes in the context of program comprehension during learning and gain novel insights into the use of visual representations for instructional purposes with intelligent programming tutors.

Last modified: 21-04-2021

Learning by teaching program tracing using virtual pedagogical agents: The role of non-cognitive factors (Dalhousie University Generic Plan)

Data Collection

What types of data will you collect, create, link to, acquire and/or record?

Textual data will be collected in this project. The specific textual data include the interactions logged in the intelligent programming tutor, student performance on practice problems, responses to surveys, code for the software, curricular material, as well as information about the study. The data describing student-tutor interactions is the most extensive of these three data types. It describes individual student actions and the responses of the tutoring system, including time stamps, student input, tutor hints, and correctness. The code of the software will be uploaded in the form of a link to a repository that includes all files with open source license that allow them to be freely used, modified, and shared. Screenshots in addition to help documentation for getting started will be included to give additional context and for those files to be understandable.

What file formats will your data be collected in? Will these formats allow for data re-use, sharing and long-term access to the data?

This research project is collecting a variety of types of data stored in non-proprietary file types to ensure ease and flexibility of reuse. Examples of these include Comma Separated Values (.csv), Text (.txt), Joint Photographic Experts Group (.jpeg), and JavaScript Object Notation (.json).

What conventions and procedures will you use to structure, name and version-control your files to help you and others better understand how your data are organized?

Each file will be named with a short description/acronym to reflect its content, date of creation, and an identifier to record different versions. For example, the format below:

FileNm_20210422_v01.docx

Where FileNm denotes the name of the file, 20210422 describes the date (22nd of April, 2021), and v01 refers to the 1st uploaded version of the file separated by the "_" underscore symbol.

A README file will explain naming conventions as well as short descriptors or acronyms used in file names.

Documentation and Metadata

What documentation will be needed for the data to be read and interpreted correctly in the future?

We will describe the study rationale, methods, and findings as well as explain how team members transformed the data, including steps taken to collect the raw data, coding and scoring rubrics applied in extracting features from the raw data, and any issues affecting data quality or any pertinent background information to assist other researchers in understanding the data. All data fields and type will be defined and explained in the data dictionary.

How will you make sure that documentation is created or captured consistently throughout your project?

A log file will be created to document any steps made by team researchers leading to the final results. The log file will be updated on ongoing meetings where any issues/problems that might occur during the research process is discussed; for example, how to deal with interrater disagreements, segmenting units for assigning codes, or incomplete/missing responses. Any decision made will be recorded in the log file to ensure it captures the changes that were agreed upon by research team members.

If you are using a metadata standard and/or tools to document and describe your data, please list here.

The Tutor Message format is used to describe student-tutor interactions (http://pslcdatashop.org/dtd/guide), a standard data format for student-tutor interaction data. This textual format provides information about student-computer interaction at a fine-grained level (the level of individual mouse clicks and typing), and includes information about the context of the interactions (e.g., the location in the curriculum in which the tutored problem appears). The data format for test results will include columns for student ID, whether the test was a pretest or posttest, a description of the item, the form of the test, whether the student was correct or incorrect, and optionally the answer the student gave. Additional data not captured by the Tutor Message format but similarly descriptive of student-tutor interactions will be stored as custom fields.

The data logged by the system follows the Experience API (xAPI) to ease the ability to share data between eLearning applications, which is stored following the JavaScript Object Notation (JSON) format using the RESTful approach to communicate with the Learning Record Store. The conversion to the Tutor Message format is done during the data analysis stage in order to ease the usability of the data for the purposes of research within the broader community.

Storage and Backup

What are the anticipated storage requirements for your project, in terms of storage space (in megabytes, gigabytes, terabytes, etc.) and the length of time you will be storing it?

Storage space is anticipated to be approximately 20-50 Megabytes based on the data collected in a previous study.

How and where will your data be stored and backed up during your research project?

The 3-2-1 backup rule will be followed for data storage and backup. Team members will transfer identifiable data stored in the Dalhousie University's institutional OneDrive, a secure cloud based storage developed by Microsoft, onto two encrypted external storage hard drives. OneDrive can be easily accessed and limited to the neutral third-party leading the consent and data collection efforts. This process will also be repeated for the de-identified data stored in the Learning Record Store, our database solution for the intelligent programming tutor. One of the external storage hard drive will be stored offsite.

How will the research team and other collaborators access, modify, and contribute data throughout the project?

Access to the identifiable data will be limited to the neutral third-party who will be responsible for interactions with study participants during the content and data collection stages of this research. This is done in order to distance the lead investigators from study participants due to their dual role as both researchers and instructors. Once the course taught by the lead investigator is concluded, the data will be checked by the researcher to ensure its accuracy and completeness, and will be combined into a master file to be backed up and encrypted.

The Dalhousie University's institutional OneDrive is used to store, share, and work with data. All transformations to the data performed by team members will be uploaded to OneDrive following the file naming convention described above, and subsequently deleted from their local machine.

Preservation

Where will you deposit your data for long-term preservation and access at the end of your research project?

Data collected in this study will be deposited on the Scholars Portal Dalhousie University Dataverse (https://dataverse.library.dal.ca/), a database that is free to access/use as well as hosted on servers housed by Canadian universities. Along with the recommendations of the Social Sciences and Humanities Research Council of Canada, identifiable information for participants or data deemed sensitive by researchers or the Research Ethics Board involved in the study will be deleted to ensure respect for individuals' right to privacy. The de-identified data and meta data will be preserved and made available for secondary analysis through Dataverse. These data will be stored indefinitely and uploaded within a two year period when the study will be completed.

Indicate how you will ensure your data is preservation ready. Consider preservation-friendly file formats, ensuring file integrity, anonymization and de-identification, inclusion of supporting documentation.

To ease the sharing and reusability of the data, all data files include a description of team members responsible for creating the data, how the data was collected, as well as coding and scoring rubrics and protocols, and a log that identifies any issues affecting data quality to facilitate understanding. Student interaction data will be converted from the xAPI standard to the Tutor Message format to ease usability by other researchers. Efforts will be made by the research team to ensure that all column names are easily understood by others, and defined in a data library. The steps followed in data analysis will be captured through syntax files to ease replicability and the final results will be documented and saved. No identifying information of participants may be included in data files. The metadata will also include information on how to navigate to the code repository for the software, relevant publications, and the funding agency and grant name.

Sharing and Reuse

What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final).

The analyzed, de-identified datasets will be stored in the Scholars Portal Dalhousie University Dataverse.

Have you considered what type of end-user license to include with your data?

We will have the Creative Commons Attribution CC BY license for the data, which allows others to distribute, reuse, share, and build upon the data as long as the original data creators are credited.

What steps will be taken to help the research community know that your data exists?

Data deposited in Dataverse is assigned a Digital Object Identifier (DOI), a unique and persistent code that can be used to locate and access the data. Attracting researchers to perform secondary analysis on this data set is likely due to Dataverse's existing user base and the presence of publications which cite Dataverse data. Metadata is harvested by the Federated Research Data Repository, a Canada wide research repository, where data can be discovered, and then shared nationally. We also link our dataset in all publications arising from a study.

Responsibilities and Resources

Identify who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible.

The project lead, Dr. Eric Poitras, is responsible for ensuring that team members as well as undergraduate/graduate research assistants follow this data management plan. If two or more team members are working jointly on a project that involves this dataset, they will determine, at the outset of their work, which member is responsible for implementing the data management plan.

How will responsibilities for managing data activities be handled if substantive changes happen in the personnel overseeing the project's data, including a change of Principal Investigator?

One of the research team members will fulfill the responsibilities of the lead investigator in the unlikely event that substantive changes occur in personnel, including a change of PI.

What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?

Dalhousie University libraries offer Dataverse services for the university at no cost to researchers. Storage of data in external drives should cost approximately 300$.

Ethics and Legal Compliance

If your research project includes sensitive data, how will you ensure that it is securely managed and accessible only to approved members of the project?

Only analyzed, de-identified data will be made available once the project is complete.

If applicable, what strategies will you undertake to address secondary uses of sensitive data?

No sensitive data will be shared; furthermore, all identifiable information will be deleted within the period of study completion or within two years. De-identified and non sensitive data will be made available on Dataverse.

How will you manage legal, ethical, and intellectual property issues?

Approval by the Dalhousie University Research Ethics Board is required to perform the tasks outlined in this study. Participants fill an informed consent agreement and may cease participation in the study voluntarily and without penalty.