Review of
Introducing Very Large Data Sets into the Classroom:
A Graphical User Interface for Teaching with Databases

The paper raises an important issue – that while interesting and increasingly available, the complexity and sometimes sheer size of large data sets makes them difficult to deal with both technically (because of potential computational breakdowns) and conceptually (because of the number of variables, often mashed or linked in smaller tables). Unfortunately, rather than pursuing the implications of how to deal with such data sets broadly, the paper focuses on the very narrow solution of extracting samples from large data sets to address computational (but not conceptual) complexity, and then explores a specific graphical user interface for doing so. In quickly moving to the GUI "solution", the paper loses sight of the much more interesting larger issues, and presents itself more as a research report on a small technical innovation rather than as a position paper. As such, many of the guiding questions suggested for review of a position paper are not relevant.

After correspondence with the TISE editor, we've decided to review the paper as a Technology Innovation rather than as a Position Paper. Though this doesn't resolve all of this reader's concerns with the paper, however, as noted below, it seems a better fit with the goals and approach of the paper.

The GUI innovation described in the paper is intended to solve the problem of providing technical access to complex large data sets, though it doesn't address how students can be using the tool to think about the larger data set and what it's saying. The authors state (p.1, ¶2) that "Even when only partial information from the database would be sufficient, there is often no, or at least no easy way for sub-setting or aggregating the data at particular levels." This seems to be the core of the problem they are addressing, but it is only partially addressed with their GUI.

The authors describe the components of the R packages used to access the off-site database, and provide specific steps for doing so. They walk the reader through use of the dialogs and tabs that make up the GUI. They don't describe how to set up the GUI for use with other data sets (though presumably it's designed for flexible use), and especially how to make links among conceptually related databases that were not collected together. Such descriptions would increase the utility of the paper.

Though the authors describe their innovation sufficiently well, it is not clear how this example of a query tool is particularly innovative. It allows for searches of subsets of variables linked by Boolean "AND" and "OR" operators – it's not clear if there is also a "NOT" operator to exclude specific classes or cases. It allows users to limit the size of the result set (but it's not clear if this is the *first* N items matching the query, or a random subset, or something else.) It then translates the search options entered using drop-down menus and numerical input into SQL syntax. On one hand, it seems the intention is to use the more intuitive interface to help students learn to write SQL syntax directly, though the display of the syntax is limited to a single line (presumably scrollable) which, however, de-emphasizes the importance of this aspect of the tool. That is, if the purpose of the GUI is to learn SQL, it's not clear that it's designed very well as a teaching and learning tool.

The authors state four Learning Objectives for the tool. It clearly facilitates #2, Obtaining access to data stored in databases; and #4, Providing a gentle introduction to SQL. It does allow #3, Searching and extracting specific data records of interest, but doesn't really support readers/ users to think well about what might be interesting, or how to frame questions in large complex data sets. And it's not at all clear how the tool accomplishes Objective #1, Familiarizing students with databases and exploring the utility of databases for efficient data storage beyond merely working with a database — the GUI doesn't seem to address any back-end issues let alone explore different possibilities for structuring data.

The authors' implicit instructional (and research) approach seems to be very procedural, and this undermines the possibility of true exploration and supporting deeper conceptual understanding. For example, just above heading 4, they say "instructors can provide students with the final sample size of the desired data subset or subset related summary statistics to verify whether students submitted the correct SQL query." Is the goal just getting the answer the instructor was looking for? Are there things that students could learn *about* sample size, or relationships among variables, by exploring more on their own? These are not pursued.

One good aspect of the paper is that they actually test and report the results of an empirical study of usability by a mix of undergraduate and graduate statistics students rather than just making a claim about their GUI without any evidence to support it. However, there are also weaknesses in the usability study and the report on the study.

The authors claim that the study assessed student understanding, but there's no real evidence of this beyond performing correctly – e.g., there were no interviews or open-ended feedback sections of the study to actually check with students about what they thought was going on with the GUI, their impressions of what works well about it and/ or what's confusing and could be improved. Looking through the questions in Table 2 confirms this reader's sense about the limited claims that should be made from this study: the first 3 questions are about reading information from the GUI correctly; the next 4 are about reading more than one piece of information and comparing them; and the last 3 (the table is missing #8) are about following instructions with a bit of interpretation – but not about actually understanding SQL. (I'm also confused about how more undergrads could have gotten #11 correct than #10, since #11 depends on #10.)

The paper doesn't report on how errors were coded (including some examples so we can understand your scheme would help) and this seems essential for us to make sense of the findings and what they mean.

Many of the findings are reported in the form of graphs that are essentially unreadable because the type size is outrageously too small — I can just barely read these if I view them at 200%; 400% makes them easily readable, but then the graphs only barely fit on my screen. Furthermore, the graphs don't have a consistent scale so they seem to suggest the same proportions of students get each question correct when there are, in fact, large differences between them.

While the graphs show undergraduate and graduate responses in different colors, it's not clear if there are differences between these groups on any of these ratings – did you conduct a statistical test (if so, report it). And I'm skeptical about the one difference you do report, that "graduate students show a more clear understanding of the importance of database knowledge as compared to the undergraduate students." To me, what you're noting is that graduates *agree* with you about the importance of this, but we don't know anything about their understanding, and we don't know if they're agreeing because of your GUI or other experiences or cultural expectations, etc. While it does seem from the graphs that there's a pre- post- difference in students' assessment of the difficult of doing the SQL task, again this finding would benefit from a more rigorous analysis/ statistical test.

The suggested next steps about joining data across several tables, allowing stratified sampling, and allowing random sampling sound like important and good directions.

Finally, there are a number of errors in the copy – missing words, wrong words, mis-spellings, etc. I would include specifics, but I believe this paper needs a more thorough revision before it could be published, so such copy-edits seem premature.

In general, the GUI only addresses one small aspect of problems associated with working with large complex data sets, and doesn't seem a very innovative approach to doing so. If its primary goal is to teach potential data analysts how to construct SQL queries, it does that a bit more effectively, but could be improved by providing a larger window in which to review the generated SQL query all at once, or by providing real-time links between changes in the interactive GUI parameters and the query text. As it currently stands, it doesn't seem very compelling.