Journal of the Midwest Association for Information Systems (JMWAIS)

Volume 2018 | Issue 1 Article 2

# Safely Using Real-World Data for Teaching Statistics: A Comparison of Student Performance and Perceived Realism between Dataset Types

Andy Luse
*Oklahoma State University*, andyluse@okstate.edu

Jim Burkman
*Oklahoma State University*, jim.burkman@oksyaye.edu

## Recommended Citation

Luse, Andy and Burkman, Jim () "Safely Using Real-World Data for Teaching Statistics: A Comparison of Student Performance and Perceived Realism between Dataset Types," *Journal of the Midwest Association for Information Systems (JMWAIS)*: Vol. 2018 : Iss. 1 , Article 2.
DOI: 10.17705/3jmwa.000037
Available at: http://aisel.aisnet.org/jmwais/vol2018/iss1/2

Journal of the Midwest Association for Information Systems

Volume 2018 | Issue 1 Article 2

**Date: 01-31-2018**

# Safely Using Real-World Data for Teaching Statistics: A Comparison of Student Performance and Perceived Realism between Dataset Types

**Andy Luse**

*Oklahoma State University, andyluse@okstate.edu*

**Jim Burkman**

*Oklahoma State University, jim.burkman@okstate.edu*

## Abstract

Academics strive to bring real-world experiences and examples into the classroom thereby creating a richer experience for the instructor and student. This goal of relevance is particularly challenging in the instruction of statistics where the instructor often must choose between "canned" simulated datasets that lack richness and relevance versus using their own research data. Real-world research datasets offer familiarity, storytelling opportunities, and an intimate understanding of the dataset providing a fuller understanding for the student; however, the public release of research data could be problematic. This article examines a solution that offers the richness and relevance of real-world datasets while safeguarding the integrity of the data and research. Experimental results support the use of these derived datasets.

**Keywords:** Simulation, real-world data, derived data

DOI: 10.17705/3jmwa.000037

## 1. Introduction

*I've been imitated so well I've heard people copy my mistakes.*
Jimi Hendrix

Statistics education is a vital piece in the overall puzzle of information systems (IS) instruction, providing practical expertise for managers in the workplace and necessary skills for future researchers (Russell, Noble, Carter, Currier, & Wiseman, 2011). The Journal of the Midwest Association for Information Systems (JMWAIS) has demonstrated an interest in improving IS education through pieces designed to aid in instruction in areas such as online learning (Hadidi & Power, 2017), project management (Harb, Noteboom, & Sarnikar, 2015; Klein, Davis, & Kridli, 2015), decision making (Klein, 2016), and doctoral education (Deokar, 2015; Hosack & Sagers, 2015). This research adds to the trend of instructional improvement present in JMWAIS by providing a method for improved instruction in statistics.

In order to facilitate optimal statistical instruction, research suggest that data should come from an authentic study consisting of an uncleansed dataset with sufficient background information (Willett & Singer, 1992). Russel et al. state the matter directly:

> *Recent work has shown that using real-life, regularly updated data to teach econometrics and related social science statistical skills has a number of benefits over using fictional or pre-configured datasets that have been developed purely for specific learning activities.* (Russell et al., 2011)

These real-world datasets hold many advantages over traditional “canned” datasets created strictly for educational purposes by providing the student with a much richer understanding of the complexity and nuances of an actual real-world data analysis.

There can be disadvantages for the instructor when using real-world datasets for instruction including preparation time, relevance to the course area, messiness of the statistical output, etc. (Kuiper & Sturdivant, 2015; Neumann, Hood, & Neumann, 2013; Willett & Singer, 1992). Researchers may also face copyright restrictions (Morgan, 2001; Neumann, Neumann, & Hood, 2010), ethical considerations (Neumann et al., 2013), and the possibility of a loss of intellectual property and research opportunities whereby the researcher reveals too much about the data prior to publication or grant work.

Simulated datasets have long been used by instructors to help better convey statistical concepts to students. These simulated datasets have been used to illustrate specific statistical properties such as random sampling, long run patterns, and other statistical phenomena (Blejec, 2002; Chance & Rossman, 2006; Chang, Lohr, & McLaren, 1992). The anonymous nature of the simulated datasets typically bypasses most copyright and intellectual property issues, but unfortunately these datasets lack the beneficial qualities of real-world datasets because they are confined to an artificial scenario with data that is typically cleansed to show an almost perfect association to the topics being covered in the course.

This research shows how real-world dataset advantages can be combined with the anonymity of simulated datasets to provide a novel approach for statistical instruction for IS undergraduate, graduate, or doctoral students. This is achieved by creating datasets derived from the distributional properties of their related real-world datasets. These new datasets are referred to in this study as “derived datasets,” in contrast to “real-world datasets” and the canned “simulated datasets.” An experiment is used to measure student performance and perceptions of realism between derived datasets and their related real-world datasets in hopes of supporting the use of the derived datasets in the classroom.

## 2. Background

Statistics education has been an important part of the educational experience for quite some time both for applied areas (Minton & Freund, 1977) as well as graduate research programs (Cockerill & Fried, 1991), but the problem remains that statistical literacy skills are lacking. Furthermore, many students have negative attitudes and anxiety towards statistics (Onwuegbuzie & Wilson, 2003; Tremblay, Gardner, & Heipel, 2000), which can severely limit instruction. This has led to a call by many to improve the teaching of statistical concepts (MacInnes, 2009). Several outlets for research and discussion pertaining to statistics education point towards the overarching goal of improving education in this area.

These include journals (*Journal of Statistics Education*, *Statistics Education Research Journal*, *Technology Innovations in Statistics Education*, etc.), conferences (International Conference on Teaching Statistics, National Conference on Education Statistics, etc.) as well as numerous other outlets that publish work on statistical instruction.

One method for improving the educational experience of students studying statistics calls for the use of real-world datasets (Russell et al., 2011). While statistics courses have been criticized for using data and concepts that seem too abstract to the student (Hogg, 1991), real-world datasets alleviate this concern by providing data of greater interest and relevance (Scheaffer, 2001). Recent research shows that students view the use of real-world datasets as relevant to learning, interesting, easier to learn and remember, motivating, promoting greater involvement and engagement, and lending itself towards greater understanding (Neumann et al., 2013). Furthermore, the use of real-world datasets gives the learning experience a more personal nature that increases interest in learning (Chottiner, 1991).

In addition to providing direct student benefits, real-world datasets that come from the instructor's research will not only increase the learning experience of the student, but also provide a more personalized feeling for the instructor that can lead to better engagement of the instructor with the material. Research has suggested that unengaged teachers who are bored themselves, are not able to properly engage their students (Newmann, 1992; Powell, Farrar, & Cohen, 1985). Conversely, teachers that are actively engaged with the subject matter try to create more active, engaging learning environments for their students (Louis & Smith, 1991). Since the instructor is likely to deeply understand real-world data collected through his or her own research, there are increased learning outcomes when a faculty member introduces students to a problem they themselves are deeply engaged in and guides them to a solution (inductive teaching) (Prince, Felder, & Brent, 2007). To this end, providing a data set that has all the same nuances as the original data allows the instructor to show students the actual process he or she took – pitfalls and successes – during that research effort. Purely simulated data, or data from some random external source, lacks this nuanced richness through personal research engagement.

Faculty also face pressure from their institutions and accrediting bodies to integrate their research into the classroom. As one example, the Association to Advance Collegiate Schools of Business (AACSB) explicitly ties the integration of faculty research to classroom learning as essential to the academic legitimacy of both a school and the related discipline (AACSB, 2012). Land grant universities, as outlined by the Morrill Act of 1862, are also given a mission of providing relevant practical education as opposed to abstract concepts sometimes used in non-land grant institutions. Combine this with the push by deans and department heads to provide more relevant instruction, and the use of personal data in the classroom becomes important.

### 2.1 Disadvantages of using real-world datasets in the classroom

While the use of real-world datasets for statistical education holds tremendous promise, there are still some pitfalls including increased preparation time, "messy" data, and ethical and copyright considerations (Kuiper & Sturdivant, 2015; Morgan, 2001; Neumann et al., 2013; Neumann, Neumann, & Hood, 2010; Willett & Singer, 1992). Many researchers utilize their own work to combat some of these limitations and take advantage of the richness of their own research knowledge but this too can present its own problems. Copyright considerations can arise with datasets that have been used for publication, grant work, or, if not yet published, the researcher must trust the students to not make the dataset available to others outside the course. This can put the instructor in the uncomfortable position of evaluating the risk of using his or her own dataset versus providing a superior level of educational quality for the students.

### 2.2 Simulations for teaching statistics

In recent years technology has allowed for novel methods for teaching statistical concepts using simulated datasets (see Mills (2002) for an overview of some of these methods). These simulated datasets allow for the instructor to more easily convey concepts such as random sampling (Simon, 1994), statistical inference (Tintle et al., 2015), long-run patterns within data, distributional assumptions (Chance & Rossman, 2006; delMas, Garfield, & Chance, 1998; Lane, 2015; Lee, Angotti, & Tarr, 2010), even financial risk (Foster & Stine, 2006). Simulated datasets provide data with pre-defined properties (Blejec, 2002) that are beneficial for some applications but which do not provide the real-world properties that are desired to convey everyday practical data analysis.

One way to provide the advantages of real-world data and the anonymity of simulated data is to combine the two methods. By deriving datasets based on distributional assumptions of an associated real-world dataset, students are introduced to the messy, practical datasets that are like what they will encounter in their work lives while maintaining the confidentiality of the actual real-world dataset to stem concerns over copyright, ethics, and loss of intellectual property. Given the similarities in the distributional attributes of the derived dataset as compared to the associated real-world dataset, we postulate that there will not be any noticeable difference in either objective outcomes or subjective measures of data "realness" when analyzing statistics from the two different datasets. Stated more formally:

H1: No significant difference will exist in the number of questions answered correctly for individuals analyzing data from a real-world dataset as compared to those analyzing data from a derived dataset.
H2: No significant difference will exist in the level of perceived realism of the dataset used for individuals analyzing data from a real-world dataset as compared to those analyzing data from a derived dataset.

## 3. Data Collection

Subjects for this research were students from both a doctoral course introducing multilevel statistical modeling and an undergraduate course where statistics was not the subject area, both at a large Midwestern university. The doctoral students were in their second year of the program and had already taken several statistical method courses. Conversely, the undergraduates were in a non-statistics related course and had minimal exposure to statistics. This provides greater generalizability of the results as both statistics novices and intermediate students were used to test the method.

The activities were completed at the beginning of the class session. The procedures are summarized in Table 1. After the instructor was introduced, the students were asked to participate in an activity for the class. The activity was designed to reintroduce basic multiple regression topics to the students. The use of a simple multiple regression exercise alleviated the potential for confounding effects of students not understanding more complex material. The students were asked to complete two activities, back to back, on two different datasets. After completing each activity, the students filled out a questionnaire. The questionnaire included questions designed with both objective and subjective measures, and the same exact questions were asked at the completion of each activity. The second questionnaire also included four questions designed to measure experience with statistics in general. The questionnaire can be viewed in the Appendix A.

1. Introduction of the first dataset
2. Individual completes first activity (using either real-world or derived dataset)
3. Individual fills out first questionnaire
4. Introduction of the second dataset
5. Individual completes second activity (using dataset opposite from first activity)
6. Individual fills out second questionnaire

Table 1. Research Procedures

The activities focused on analyzing the output of a multiple regression analysis from two separate datasets. Both datasets included three independent variables and one dependent variable for consistency. The output was standard output from SAS running the PROC REG procedure. Before each activity, the instructor gave a basic overview of the study and also displayed a model consisting of the three independent variables in boxes with arrows drawn to the dependent variable, also in a box. After this, the students were instructed to fill out the survey and wait for everyone to finish before repeating the process for the second dataset.

The questions on the survey were designed with both objective and subjective measures. The objective measures were designed to assess the ability of the student to correctly answer questions about the statistical output. The subjective measures were designed to measure the subject's impression of the "real-worldness" of the data. Two separate sets of questions were used to measure these impressions. The first set of six questions were adapted from a set of measures used to measure visual realism in an immersive virtual environment (Slater, Khanna, Mortensen, & Yu, 2009). The second set of three questions were adapted from a set of measures used to measure perceived reality of television (Huston et al., 1995). Both these sets of measures helped to gauge the perception of the real-world nature of the data and are used in the test of Hypothesis 2.

The purpose of the experiment was to examine the effect of a real-world dataset versus a derived dataset. To accomplish this, for the first activity, half the students analyzed output from a real-world dataset and the others from a derived dataset. The students were randomized into each condition. Then, for the second activity, the students analyzed a second dataset from a completely new study using the opposite real-world or derived dataset as they did for the first dataset. This created a counterbalanced, crossover, within-subjects experimental design where each student analyzed both a real-world dataset and a derived dataset from two different studies and were randomly assigned as to the order of the type of dataset (real-world vs. derived) to control for learning effects (Heppner, Wampold, & Kivlighan Jr, 2007).

The derived data was created using the R statistical package with a function designed to derive data. The function reads the input of a real-world dataset then creates a derived dataset based on the distributional properties of the original dataset. The derived dataset can also be tailored to output a specific number of subject lines. This allowed for the two real-world datasets in this study to be used as inputs and two derived datasets with the same number of subjects as the original datasets to be created. The derived data points were completely different from the real-world data points, but the statistical output was extremely similar due to the derived dataset having the same distributional properties as the original real-world dataset. The script used to create the derived datasets as well as the link to the R function are included in Appendix B.

Given the nature of the hypothesized insignificance for this study, an a-priori power analysis was run to determine the number of subjects needed to find a difference if one were to exist. Typical planning values of $\alpha = 0.05$ and $\beta = 0.2$ (i.e. Power of 0.8) were used. Given the nature of the scales (0 to 5 for the objective measure and 1 to 5 for the Likert scale) a mean difference of 0.5 was identified by the researchers as a desired metric: anything less than this value would not account to a noticeable effect for those using a derived dataset as compared to a real-world dataset in an educational setting. Finally, a very conservative value of one was also used as the planning value of the standard deviation of the mean difference. Given these values, the necessary sample size needed to find this difference is 33 (Faul, Erdfelder, Lang, & Buchner, 2007).

## 4. Results

In total, 45 students completed the research activities. The within-subject experiment examined one within-subject independent variable and three dependent variables. The independent variable consisted of whether the activity was performed on a real-world or a derived dataset, making this a 2-level within-subjects design. The first dependent variable consisted of the number of answers out of five they answered correct when asked about the statistical output. The other two variables of realism and perceived reality measured their sense of the activity and the data used as actual data from actual subjects. The realism measure consists of six items and has a high reliability ($\alpha = 0.85$) while the perceived reality measure consists of three items and also has a high reliability ($\alpha = 0.81$). Each dependent variable was analyzed separately, culminating in three separate analyses.

A paired-samples t-test was used to analyze the data. The amount of correct answers when using the real-world dataset as compared to the derived dataset was not significant ($t_{(44)} = 0.00$, $p = 1.00$, CI=[-0.18, 0.18]). The confidence interval shows an interval width of around 0.36 (well below the desired width of 0.5) with the equal upper and lower bounds showing extreme insignificance. This indicates that the use of a derived dataset does not significantly impact the number of questions an individual answer correctly, supporting H1. Also, the level of realism that the subject perceives with regard to the analysis and data is not significantly different when using a real-world dataset as opposed to a derived dataset ($t_{(44)} = 1.45$, $p = 0.15$, CI=[-0.04,0.22]) and the level of perceived reality of the subject with regard to the analysis and data is also not significantly different when using a real-world dataset as opposed to a derived dataset ($t_{(44)} = 0.96$, $p = 0.34$, CI=[-0.10, 0.28]), thus supporting H2. See Table 2 for an overview of the statistics.

| | real-world | | derived | | | | | |
|---|---|---|---|---|---|---|---|---|
| | mean | stdev | mean | stdev | Cronbach | t-value | p-value | 95% CI |
| Correct | 4.36 | 1.00 | 4.36 | 0.91 | - | 0.00 | 1.00 | -0.18, 0.18 |
| Realism | 3.72 | 0.52 | 3.63 | 0.71 | 0.85 | 1.45 | 0.15 | -0.04, 0.22 |
| Perceived Reality | 3.56 | 0.61 | 3.47 | 0.66 | 0.81 | 0.96 | 0.34 | -0.10, 0.28 |

Table 2. Descriptive statistics, reliability values, and test values for differences between real-world dataset and derived dataset groups.

## 5. Discussion

This paper discusses the use of datasets derived from real-world data in IS instructional contexts. Research has shown that the use of real-world datasets in statistical education provides many advantages to students in terms of relevance, motivation, involvement, and engagement (Neumann et al., 2013), but these advantages can be overshadowed by concerns over copyright considerations as well as the potential loss of intellectual property. This research presents a method for deriving datasets based off the distributional nature of real-world datasets. Results from an experimental analysis show no noticeable differences in objective student outcomes using derived datasets as well as no noticeable differences in student perceptions of the real-world nature of the derived datasets.

This study provides two important contributions. First, the results demonstrate that the use of derived datasets modeled from real-world datasets provide similar student performance outcomes to actual real-world datasets in educational settings. This frees instructors to use their own personal datasets without fear of compromise or legal limitations. This also holds the potential for an excellent classroom experience wherein the instructor can speak to the context of the data collection and environment, provide rich storytelling opportunities about the associated real-world scenario, and overall bring a rich, engaging learning opportunity to the classroom.

Second, Appendix B provides an easy mechanism for instructors to derive their own datasets from personal real-world datasets. The R statistical package is free to use and the code and referenced package in appendix B provide an easy to use mechanism for anyone to create a derived dataset based on their own data. This offers a practical, time-conscious method of helping faculty bring their research into the classroom without adding extensively to their course prep time.

In addition, the use of this method could also be utilized in the burgeoning instructional area of analytics as the same advantages and pitfalls with regard to using real-world datasets are the same for both. Given this, our method could be used to mimic the distributional assumptions of datasets used for analytics instruction to allow for the same benefits of real-world data without the drawbacks associated with releasing one’s data.

## 6. Limitations and Future Work

The current study provides numerous avenues for future research by focusing on the subjects, the tasks, and/or the dataset types. First, this study uses both students from a doctoral course and an undergraduate course. While legitimate, these students may not be indicative of other students taking a statistics course such as statistics majors, other business students, etc. Future research should investigate the use of derived datasets in other samples of students and also among working professionals.

Second, given the novelty of the proposed solution, this study used a very introductory statistical technique to gauge outcomes in using derived datasets. Now that the foundation has been laid, future researchers should investigate the use of derived datasets for teaching more advanced statistical concepts and techniques, including courses in data analytics. The application of derived datasets in case-study courses is also interesting, as a balance may be found between the realism of the case information and the privacy of the actual data.

Third, this study looked at differences between real-world datasets and derived datasets, but intentionally did not include the “canned” simulated datasets. While the advantages of the richness associated with real-world datasets seem clear, additional types of datasets should be compared for domain completion.

Finally, while both objective and subjective measures were used to evaluate the use of derived datasets, other outcome measures –in both the positivist and the interpretivist traditions - should certainly be considered in future work.

## 7. Conclusion

This research provides a novel method to enable the use of real-world datasets in statistics courses through the use of derived simulated datasets based on the distributional properties of real-world data. The experimental analysis shows that there is no significant difference between objective outcomes and subjective perceptions of these derived datasets and their real-world counterparts. This provides a useful tool to enable statistics instruction using datasets of which instructors have intimate knowledge and that provide the preferred educational experience that will better prepare students for real-world data analysis.

## References

AACSB. (2012). *Impact of Research: A Guide for Business Schools*. Retrieved from http://www.aacsb.edu/-/media/aacsb/publications/research-reports/impact-of-research-exploratory-study.ashx?la=en

Blejec, A. (2002). Teaching statistical concepts with simulated data. *The publishing of this booklet is a part of the Tempus project "Master programme in applied statistics" MAS 511140-Tempus-1-2010-1-RS-Tempus-JPCR*, 1.

Chance, B., & Rossman, A. (2006). *Using simulation to teach and learn statistics.* Paper presented at the Proceedings of the Seventh International Conference on Teaching Statistics.

Chang, T. C., Lohr, S. L., & McLaren, C. G. (1992). Teaching Survey Sampling Using Simulation. *The American Statistician, 46*(3), 232-237. doi:10.1080/00031305.1992.10475892

Chottiner, S. (1991). Using real (intimate) data to teach applied statistics. In (Vol. 45, pp. 169-169): American Statistical Association.

Cockerill, R., & Fried, B. (1991). Increasing public awareness of statistics as a science and a profession—reinforcing the message in universities. *The American Statistician, 45*(3), 174-178.

delMas, R., Garfield, J., & Chance, B. (1998). *Assessing the effects of a computer microworld on statistical reasoning.* Paper presented at the Proceedings of the Fifth International Conference on Teaching Statistics.

Deokar, A. V. (2015). A Reply to Hosack and Sagers "Applied Doctorates in IT: A Case for Designing Data Science Graduate Programs". *Journal of the Midwest Association for Information Systems*(1), 69.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods, 39*(2), 175-191.

Foster, D. P., & Stine, R. A. (2006). Being Warren Buffett. *The American Statistician, 60*(1), 53-60. doi:10.1198/000313006X90378

Hadidi, R., & Power, D. (2017). Implications of the Sharing Economy for Online and Blended Education. *Journal of the Midwest Association for Information Systems| Vol, 2017*(1), 1.

Harb, Y., Noteboom, C., & Sarnikar, S. (2015). Evaluating Project Characteristics for Selecting the Best-fit Agile Software Development Methodology: A Teaching Case. *Journal of the Midwest Association for Information Systems*(1), 33.

Heppner, P., Wampold, B., & Kivlighan Jr, D. (2007). *Research design in counseling*: Cengage Learning.

Hogg, R. V. (1991). Statistical education: Improvements are badly needed. *The American Statistician, 45*(4), 342-343.

Hosack, B., & Sagers, G. (2015). Applied Doctorate in IT: A Case for Designing Data Science Graduate Programs. *Journal of the Midwest Association for Information Systems, 1*(1), 61-68.

Huston, A. C., Wright, J. C., Alvarez, M., Truglio, R., Fitch, M., & Piemyat, S. (1995). Perceived television reality and children's emotional and cognitive responses to its social content. *Journal of Applied Developmental Psychology, 16*(2), 231-251.

Klein, B. D. (2016). Developing an Applied, Integrated MBA Managerial Decision Making Course. *Journal of the Midwest Association for Information Systems| Vol, 2016*(2), 61.

Klein, B. D., Davis, T. A., & Kridli, G. (2015). Building a Rube Goldberg Machine in an Undergraduate Business School Course to Learn Principles of Project Management and Leadership Skills. *Journal of the Midwest Association for Information Systems| Vol, 2015*(2), 53.

Kuiper, S., & Sturdivant, R. X. (2015). Using Online Game-Based Simulations to Strengthen Students' Understanding of Practical Statistical Issues in Real-World Data Analysis. *The American Statistician, 69*(4), 354-361. doi:10.1080/00031305.2015.1075421

Lane, D. M. (2015). Simulations of the Sampling Distribution of the Mean Do Not Necessarily Mislead and Can Facilitate Learning. *Journal of Statistics Education, 23*(2).

Lee, H. S., Angotti, R. L., & Tarr, J. E. (2010). Making comparisons between observed data and expected outcomes: students' informal hypothesis testing with probability simulation tools. *Statistics Education Research Journal, 9*(1), 68-96.

Louis, K. S., & Smith, B. (1991). Restructuring, Teacher Engagement and School Culture: Perspectives on School Reform and the Improvement of Teacher's Work. *School Effectiveness and School Improvement, 2*(1), 34-52.

MacInnes, J. (2009). Proposals to support and improve the teaching of quantitative research methods at undergraduate level in the UK. *Economic and Social Research Council*.

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education, 10*(1), 1-20.

Minton, P. D., & Freund, R. (1977). Organization for the conduct of statistical activities in colleges and universities. *The American Statistician, 31*(3), 113-117.

Morgan, B. L. (2001). Statistically Lively Uses for Obituaries. *Teaching of Psychology, 28*(1), 56-58.
Neumann, D. L., Hood, M., & Neumann, M. M. (2013). Using real-life data when teaching statistics: Student perceptions of this strategy in an introductory statistics course. *Statistics Education Research Journal, 12*(2), 64-75.

Neumann, D. L., Neumann, M. M., & Hood, M. (2010). The development and evaluation of a survey that makes use of student data to teach statistics. *Journal of Statistics Education, 18*(1), 1-19.

Newmann, F. M. (1992). *Student engagement and achievement in American secondary schools*: ERIC.

Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics Anxiety: Nature, etiology, antecedents, effects, and treatments--a comprehensive review of the literature. *Teaching in Higher Education, 8*(2), 195-209
.
Powell, A. G., Farrar, E., & Cohen, D. K. (1985). *The shopping mall high school*. Boston: Houghton Mifflin.

Prince, M. J., Felder, R. M., & Brent, R. (2007). Does faculty research improve undergraduate teaching? An analysis of existing and potential synergies. *Journal of Engineering Education, 96*(4), 283-294.

Russell, C., Noble, S., Carter, J., Currier, S., & Wiseman, R. (2011, July 4-6). *Real World, Real Stories: Teaching Quantitative Methods with Real Life Data.* Paper presented at the 3rd International Conference on Education and New Learning Technologies (EDULEARN11), Barcelona, Spain.

Scheaffer, R. L. (2001). Statistics education: Perusing the past, embracing the present, and charting the future. *Newsletter for the section on statistical education, 7*(1).

Simon, J. L. (1994). What Some Puzzling Problems Teach about the Theory of Simulation and the Use of Resampling. *The American Statistician, 48*(4), 290-293. doi:10.1080/00031305.1994.10476083

Slater, M., Khanna, P., Mortensen, J., & Yu, I. (2009). Visual realism enhances realistic response in an immersive virtual environment. *Computer Graphics and Applications, IEEE, 29*(3), 76-84.

Tintle, N., Chance, B., Cobb, G., Roy, S., Swanson, T., & VanderStoep, J. (2015). Combating Anti-Statistical Thinking Using Simulation-Based Methods Throughout the Undergraduate Curriculum. *The American Statistician, 69*(4), 362-370. doi:10.1080/00031305.2015.1081619

Tremblay, P. F., Gardner, R., & Heipel, G. (2000). A model of the relationships among measures of affect, aptitude, and performance in introductory statistics. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 32*(1), 40.

Willett, J. B., & Singer, J. D. (1992). Providing a Statistical "Model": Teaching Applied Statistics Using Real-World Data. In F. Gordon & S. Gordon (Eds.), *Statistics for the twenty-first century* (pp. 83-98). Washington, DC: Mathematical Association of America.

## Appendix A: Questionnaire

***After running the analysis of the dataset as directed by the instructor, please answer the questions on this page ONLY.***

***The following questions are about the regression analysis. Please answer these to your best ability.***

| | | | |
|---|---|---|---|
| How many subjects are in the dataset? | ______________________ | | |
| Is the overall regression model statistically significant? | Yes | No | |
| How much variance is being explained in the dependent variable? | ______________________ | | |
| Which independent variable is having the greatest impact on the dependent variable? | Process | Trust | Ability |
| Which independent variable is having the least impact on the dependent variable? | Process | Trust | Ability |

***Please rate your reaction to working with this data. We would like to have you rate this data on a scale from Strongly Disagree on the left to Strongly Agree on the right.***

| | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
|---|---|---|---|---|---|
| During this experience, I had a sense of working with real-world data. | □ | □ | □ | □ | □ |
| There were times during the experience when the dataset was real for me. | □ | □ | □ | □ | □ |
| During the experience I was thinking that I was working with real-world data. | □ | □ | □ | □ | □ |
| I responded as though the data I was working with were real. | □ | □ | □ | □ | □ |
| I deliberately worked as if the data were real. | □ | □ | □ | □ | □ |
| My thoughts while working with the data were the same as if it had been a real statistical analysis. | □ | □ | □ | □ | □ |
| The data used was real-world data. | □ | □ | □ | □ | □ |
| The data was collected from actual subjects. | □ | □ | □ | □ | □ |
| The data represents the attitudes of actual subjects. | □ | □ | □ | □ | □ |

| | None | Little | Some | Quite a bit |
|---|---|---|---|---|
| Rate your experience level with statistics. | □ | □ | □ | □ |
| Please rate your experience level with SAS. | □ | □ | □ | □ |
| Please rate your experience level with other statistical packages. | □ | □ | □ | □ |

## Appendix B: R Script

Below is the R script used to read a dataset from an Excel workbook, create the derived dataset, and save the new derived dataset to a new Excel workbook. The most up-to-date version of the fakeData function can be downloaded from the following location: http://openmx.psyc.virginia.edu/wiki/generating-simulated-data.

```
setwd("C:\\openMx")

install.packages("rJava")
install.packages("xlsx")
require(rJava)
require(xlsx)

df = read.xlsx("dataSCCT_Reg.xlsx",sheetName="Sheet1")

source("FakeData.R")

SCCT_sim = fakeData(df)
write.xlsx(SCCT_sim, "dataSCCT_Reg_sim.xlsx", "Sheet1", TRUE, FALSE)
```

## Author Biographies

**Andy Luse** received a B.A. degree in Computer Science from Simpson College, M.S. degrees in Information Assurance, Computer Engineering, Business Administration, and Psychology, and Ph.D. degrees in Human Computer Interaction, Computer Engineering, and Information Systems from Iowa State University. He is currently an Assistant Professor in Management Science and Information Systems at Oklahoma State University. Andy's research has focused on computer security, technology acceptance, and research methods. He has been published in the Journal of Management Information Systems, IEEE Transactions on Visualization and Computer Graphics, ACM Transactions on Computing Education, IEEE Transactions on Education, Decision Sciences Journal of Innovative Education, Computers and Human Behavior, and several other outlets.

**Jim Burkman** holds a Ph.D. in Management Information Systems from Indiana University and is currently a Clinical Associate Professor at Oklahoma State University. Jim is the program coordinator for the Master of Science in Information Assurance degree and teaches applied and theory courses in information assurance, digital forensics and database. His research focuses on the social psychology aspects of technology, to include issues of trust and expectations in the areas of secure computing and technology adoption. He has published in the Journal of the AIS, European Journal of Information Systems and other journals and conferences.