Relationship between Perceived and Actual Quality of Data Checking

Hunter Speich, Sophia E. Karas, Dan N. Erosa, Kelly E. Grob, & Kimberly A. Barchard
University of Nevada, Las Vegas

Abstract

Data quality is critical to reaching correct research conclusions. Previous research has demonstrated that some methods of data checking are better than others, but not all researchers use the best methods. The purpose of this study was to examine the relationship between perceived data quality and actual data quality. A total of 29 participants completed this study. Participants checked that letters and numbers had been entered correctly into the computer using one of three randomly assigned data checking methods. Afterwards, they rated the quality of their data checking method. The sample correlations between perceived and actual data quality were small to moderate and confidence intervals for the population correlations did not include high values. We conclude that the relationship between actual and perceived data quality is not high. Researchers should not trust their subjective evaluations data checking effectiveness: They need empirical evidence of the quality of their data checking.

Introduction

Data entry accuracy is extremely important to researchers who desire exact results. This research will attempt to reveal the extent to which subjects' perceptions of the quality of their data checking methods are flawed; in other words, the discrepancy between perceived data checking quality and actual data checking quality. It may be the case that subjects rely on faulty versions of data checking methods, mistakenly believing them to be accurate.

Accurate study results depend heavily on correcting mistakes in the relevant data entry. It is hard to overestimate the destructive impact a single entry error can have on study outcomes. For instance, significant t test results can be mistakenly interpreted to be non-significant, or a moderate correlation might be misinterpreted for zero (Barchard, Pace & Burns, 2009). Thus significant statistical results might be rendered invalid. Such problems, however, are easily preventable with the use of better data checking methods.

The comparison of different data checking methods consistently proves Double Entry to be more reliable than both Single Checking and Visual Checking (Reynolds-Haertle & McBride 1992). Though the Read Aloud method in particular reduced time expenditure, Double Entry was demonstrated to have a much higher error detection rate. And yet, despite recognizing the importance of data quality, researchers nonetheless persist in using data checking methods inferior to Double Entry.

Perhaps researchers do not use double entry because they mistakenly believe their method to be very accurate. If researchers are good judges of accuracy, their perceptions are all right. However, if actual accuracy is not highly related to perceived accuracy, researchers may be using ineffective methods of data checking. The purpose of this study is to examine the relationship between perceived data quality and actual data quality.

Method

Participants

A total of 29 (14 male, 15 female) undergraduate students participated in this study to receive course credit. Their ages ranged from 18 to 39 (mean = 22.3, SD 6.0). The participants identified themselves as: 13.8% African American, 27.6% Asian, 24.1% Caucasian, 24.1% Hispanic, 6.9% Pacific Islander, and 3.4% Other.

Measures

Perceived quality was measured with two items. Participants were asked to rate the data checking method they used in terms of its accuracy and reliability. After the participants completed the data entry, filled out a questionnaire.

Actual quality was measured using a five-point scale, where Strongly Disagree = 1, Disagree = 2, Neutral = 3, Agree = 4, Strongly Agree = 5.Higher scores indicate that subjects have greater confidence in the data checking method's reliability and accuracy Actual accuracy was measured as the number of correct entries in the database after data checking was complete. Perceived accuracy was measured using one item. Participants were asked to rate the data checking method they used in terms of its accuracy and reliability.

Procedures

This study was given as an in-person study that participants could complete at a scheduled time with an administrator. Participants completed the study in person during supervised 90-minute sessions. Before starting the data checking, participants watched a video on Excel to ensure they know how to use the program. Then participants were randomly assigned one of these data-checking three methods: Double Entry, Visual Checking, or Read-Aloud to check data in Excel. The data checked included numbers and words. The data checking was done in two groups, once for practice and once to be scored. When using Double Entry the participants compared the data on the screen to data on a piece of paper. Once the data was checked, they would type in the data on a different paper and Excel would run a comparison of the original data and the newly-typed data to check for accuracy. The data comparison is used to catch missed errors that the participant can go back and correct. With Visual Checking the participant read over the data once and visually compared it with data on a sheet of paper to sure it was correct. With Read-Aloud the Administrator read the data out loud to the participant who compared what was read to data on a monitor. The participant would then have had the reader reread the data if it seemed incorrect and then change the mistakes they caught. After all of the data checking was completed, the participants then completed a survey based on a five-point scale.

Analysis

We calculated two correlations: the correlation between perceived reliability and actual accuracy and the correlation between perceived accuracy and actual accuracy.

Results

The correlation between perceived reliability and actual accuracy was moderate and approached significance ($r(27) = .49$, $p = .008$). A 95% confidence interval for the population correlation is given by the interval [.15, .72]. The correlation between perceived accuracy and actual accuracy was very small and non-significant ($r(27) = .12$, $p = .550$) and a 95% confidence interval for the population correlation is given by the interval [-.26, .46]. Thus, neither of these correlations are strong.

Discussion

The purpose of this study was to examine the relationship between perceived and actual data quality. Although one of the two correlations was statistically significant, neither correlation was large. Our research had a 95% confidence intervals for the population correlations showed that there is no evidence that the population correlations are large. This suggests that people are poor judges of the quality of the data checking methods they are using. We conclude that subjective judgments of data quality cannot be used as a substitute for an empirical evidence of the quality of a data checking method.

Efficient data entry is essential to facilitate the integration of data. It allows us to store, process, and share information rapidly with others around the world. One may question, however, the quality of the data presented. Which, out of "Double Entry", "Visual Checking", and "Read Aloud", is the most precise in regards to accuracy and reliability? Double Entry has been found to be the most accurate method of data checking. (Johnson, H.C., et al., 2010) Though its tendency to reveal human error, in contrast to "Visual Checking" and "Read Aloud", may make it seem flawed, it is a prudent form of data entry. Therefore, we hypothesized that the Double Entry method would result in fewer errors, but the participants would rate it lower for accuracy and quality.

Due to lack of participants, the study did not produce significant data to prove or disprove our hypothesis. Had there been enough participants to allow for sufficient data, the study may have revealed differences in quality among the various data entry methods. The data at present shows the correlation between actual and perceived data quality is not strong. Therefore, researchers should not trust their subjective evaluations of the effectiveness of their data checking methods. They need empirical evidence of the quality of their data checking.

Future research will use larger sample sizes and will likely find significant results based on the three different data checking methods. This future research should determine if these results of accuracy, reliability, and perceived quality hold true for all three methods, leading to comparisons and discoveries of new methods of data entry or the improvement of existing methods.

References

Barchard, K.A., Pace, L.A., & Burns, S.S. (2009, May). *Double entry: Accurate results from accurate data.* Poster presented at the Association for Psychological Science convention, San Francisco, CA.

Johnson, H.C., Verenikina, Y., Hensley, S., Grob, K.E., & Barchard, K.A (2010). *Check it: Evaluating the accuracy of three data checking techniques.* Poster presented at the Western Psychological Association Annual Convention, Los Angeles, CA.

Kawado, M., Hinotsu, S., Matsuyama, Y., Yamaguchi, T., Hashimoto, S., & Ohashi, Y (2003). A comparison of error detection rates between the reading aloud method and the double data entry method. *Controlled Clinical Trials, 24*, 560-569. doi:10.1016/S0197-2456(03)00089-8

Reynolds-Haertle, R. A., & McBride, R. (1992). Single versus double data entry in CAST *Controlled Clinical Trials*, *13*, 487-494. doi:10.1016/0197-2456(92)90205-E