Data Checking: Which Method is Superior?

Theresa M. Ross, Ernesto H. Bedoy, Ashley A. Anderson, & Kimberly A. Barchard
University of Nevada, Las Vegas

ABSTRACT

After research data are input into the computer, researchers use various data checking methods to find and correct possible data entry errors. Not all data checking methods are equally effective. The purpose of this study was to determine which data checking method is most effective at finding and correcting errors. We examined four data checking methods: visual checking, solo read-aloud, partner read-aloud and double entry. Previous research has shown that double entry is more effective than visual checking (Barchard & Pace, 2011) or partner read-aloud (Kawado, Hinotsu, Matsuyama, Yamaguchi, Hashimoto, & Ohashi, 2003). Although no research has examined solo read-aloud (where the typist reads the data sheet out loud while entering the data), we hypothesized that double entry would be more effective than all three of the remaining methods. Before participants arrived for the study, we created a dataset that they would check. This Excel dataset was deliberately created to contain 32 errors. Participants' job was to locate and correct those errors.

Twenty-seven undergraduates participated in this study in return for course credit. These participants were randomly assigned one of the four data checking methods. Participants checked the Excel file against the data on twenty data sheets. After each participant finished data checking, we calculated the number of errors that remained in the dataset. Double entry had the fewest errors, while partner read-aloud had the most errors. These results supported our hypothesis that double entry is the most effective. Future research should replicate these results using a larger sample size, so that we can distinguish between each of the four data checking methods. In addition, future research should use more experienced data entry personnel, such as graduate students and paid professionals, to extend the generalizability of these results.

INTRODUCTION

Have you ever calculated a complex mathematical problem only to get the answer wrong because of a simple addition error? It can be quite frustrating because you would have gotten the answer correct if you only went back to check your work. When researchers or doctors input incorrect data, their results could be catastrophic to their practice. Checking data is a way to produce only the most accurate work whether it is in research or other work-related settings. For example, a researcher is about to analyze the results. The researcher enters data into a spreadsheet and accidentally enters wrong numbers for different variables then does not check the entered data afterwards. When the researcher runs the statistical analysis, the researcher will be using incorrect data which increases the possibility of incorrect significance. Another setting would be a medical setting in which patient information can be entered incorrectly thus increasing the chance of a misdiagnosis. Having incorrect data leads to the possibility of having an outlier which will skew results significantly (Wilcox, 1998). By using a data checking method, accuracy of results can be increased.

Data checking is used to verify the accuracy of the data that researchers input, analyze, and base their studies on. There are a few methods being used, which include partner read-aloud, solo read-aloud, double entry, and visual checking. Past research examines the effectiveness of methods such as double entry in comparison to other methods as well as using different or same operators. For example, double entry and visual checking were compared to each other to determine which method produces the most accurate data (Barchard & Pace, 2011). Double entry is more effective than visual checking because visual checking yields 2958% more errors than double entry (Barchard & Pace, 2011). Double entry was also compared to the partner and solo read-aloud methods. Double entry detected 88.3% of errors with the same operator while double entry with a different operator detected 69.0% of errors (Kawado et al., 2003). In comparison, the read-aloud method with the same operator detected 59.5% of errors, and the read-aloud method with a different operator detected 39.9% of errors (Kawado et al., 2003). Although double entry took longer (74.8 hours) than read-aloud (57.9 hours), double entry produced more accurate results (Kawado et al., 2003). In addition, using different operators for double entry was more effective than using same operators for double entry. When double entry was compared to visual checking, double entry still had 15 errors whereas visual checking still had 22 errors out of 10,000 data entries. Researchers have said that if double entry was to be omitted, it could have catastrophic effects on study results and conclusions (Atkinson, 2012). This research concluded that the double entry method is the more accurate data checking method.

The purpose of the current study is to examine which data checking method is the most effective. In other words, the most effective data checking method is the method that catches the most errors. It is important to find out which method works the best. We hypothesize that double entry is the most effective method to check data. A small error increases the chance of drastically changing results. Having correct data increases the validity of research because the results reflect the inputted data. Having correct data also increases the reliability of research because it would be difficult to replicate a study and compare them when the results from the first study are not valid. Thus researchers should be more careful when entering and checking data.

METHOD

Participants

Participants were 27 undergraduate students (18 female, 9 male). Participants identified as the following: African American (14.8%), Asian (14.8%), Caucasian (40.7%), Hispanic (18.5%), Pacific Islander (7.4%), and Other (3.7%). The ages ranged from 18 to 46 years old (M = 21.48, SD = 6.216).

Procedures

All sessions were supervised by a trained undergraduate research assistant. Each session was an hour and 30 minutes long. All participants were tested individually. After reading and agreeing to the informed consent, participants watched a video tutorial about Microsoft Excel.

The computer randomly assigned the participants to one of the data checking methods: visual checking, double entry, solo read aloud, or partner read aloud. Afterwards, the participants watched a video tutorial explaining the method assigned to them. For the visual checking method, the participant visually reviewed the previously entered data on Microsoft Excel. The participant then visually compared his or her data with the original sheet and corrected any discrepancies found in the data. In double entry, the data was already entered once prior to the study. The participant must enter the data a second time from the original sheet. When there was an error, the computer program highlighted the mismatch or out-of-range error in which the participant made a correction. In solo read aloud, the participant read aloud the data from the original data sheet and checked it with Microsoft Excel. For partner read aloud, the procedure was the same as solo read aloud except the participant read aloud with the administrator. The administrator read data from the original sheet while the participant checked the data on the Excel sheet.

Data Analysis

We assessed the efficiency of data checking methods and how well they detected errors in comparison to each other. The independent variable was the group each participant belonged to which included visual checking, double entry, solo read aloud, and partner read aloud. The dependent variable was the number of errors left in each data set after a participant completed checking data.

RESULTS

There were significant differences between the four data checking techniques ($F(3,23) = 3.18$, $p = .043$). Double entry had the fewest errors and partner read aloud had the most. See Table 1.

Table 1
Average Number of Errors Left in the Data Set after Data Checking

Data Checking Method	Mean	Standard Deviation
Visual Checking	1.33	1.86
Partner Read Aloud	2.33	1.52
Double Entry	0.00	0.00
Solo Read Aloud	2.00	1.73

UNLV

DISCUSSION

Data checking is vital for having accurate results. Inaccurate results could lead to improper patient treatment in the medical field or false discoveries in the scientific field. The computer randomly assigned each of the participants to implement one of four data checking methods: partner read-aloud, solo read aloud, double entry, and visual checking. Afterwards, data checking methods were examined in order to find the most effective method in regards to finding the most errors within a data set. The results supported our hypothesis by indicating double entry as the most effective method while partner read aloud was the least effective.

In this study, double entry is shown to be the best method for checking data. When participants make errors when inputting errors, the discrepancy is highlighted in the Excel file. This is the only data checking method where the participant knows exactly which error to correct because of Microsoft Excel. This leaves the participant with a high level of confidence in which they know they are able to fix the error completely. Not only does double entry highlight the exact discrepancies, but it shows the number of discrepancies and range errors. This makes double entry the most helpful and user friendly data checking methods.

Unexpectedly, partner read aloud contained the most errors left over from the data set after the participant finished data checking. There are many possible reasons for this poor performance. Perhaps the people reading the data sheets sometimes read too fast. Perhaps the participants do not want to interrupt the person reading the data sheets, because interrupting is rude. Or perhaps the participant notices the discrepancy between the Excel sheet and what they hear, but they are not confident that they have read the entry correctly; the next data piece is read before they have made their decision, and their attention is torn away from the discrepant entry. These possibilities may suggest why solo read aloud had fewer errors.

Double entry is still the only viable alternative to automated forms. Paulsen, Overgaard, and Lauritsen (2012) compared the effectiveness of double entry and automated forms. In their study, the automated forms scanned handwritten marks and inputted the marks into the computer. They found double entry to be just as effective as the automated forms. Our study shows that no other manual data checking method is as good as double entry. Therefore, if researchers cannot afford to use automated forms, they should use double entry.

On the other hand, these results should be viewed with caution. Because participants were randomly assigned to groups, and because there were very few participants, only a single participant was assigned to the double entry condition. These results should therefore be replicated with a larger sample size. With a larger sample size, we would be more confident that the results would generalize to other data checkers.

Future research should examine differences between data checking methods with more experienced data entry personnel, such as graduate students and paid professionals. Undergraduate students may have little experience or motivation with data entry. In research, those inputting data and doing data checking are more experienced than an undergraduate enrolled in an introductory psychology course. This expert population would provide more valid results.

Since data entry is part of research, researchers should incorporate the best method to check their data. Even if a method is more time consuming, it is very beneficial to do the extra work in the end.

REFERENCES

Atkinson, I. (2012). Accuracy of data transfer: double data entry and estimating levels of error. *Journal of Clinical Nursing*, 21, 2730-2735. doi:10.1111/j.1365-2702.2012.04353.x

Barchard, K. A., & Pace, L. A. (2011). Preventing human error: The impact of data entry methods on data accuracy and statistical results. *Computers in Human Behavior*, *27*(5), 834-1839.

Healy, A., Kole, J., Buck-Gengler, & C., Bourne Jr., L. (2004). Effects of prolonged work on data entry speed and accuracy. *Journal of Experimental Psychology*, *10*(*3*), 188-199.

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.

Paulsen, A., Overgaard, S., Lauritsen, J. M. (2012). Quality of data entry, double entry and automated forms processing-an example based on a study of patient-reported outcomes. *PLoS ONE*, *7*(4), 1-7. doi:10.1371/journal.pone.0035087

Wilcox, Rand R. (1998) How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*(3), 300-314. doi:10.1037/0003-066X.53.3.300