Which Data Checking Method is More Accurate?

Sarah Cobb **&** Dr. Kimberly A. Barchard
University of Nevada, Las Vegas

Reference: Cobb, S. A., & Barchard, K. A. (2014, April). *Which data checking method is more accurate?* Poster presented at the Western Psychological Association annual convention, Portland, OR.

Contact Information: Kimberly A. Barchard, Department of Psychology, University of Nevada, Las Vegas, 4505 S. Maryland Parkway, P.O. Box 455030, Las Vegas, NV, 89154-5030, USA, barchard@unlv.nevada.edu

Abstract

Researchers use multiple methods for data checking. Each method can help identify and fix errors that were introduced during the data entry process. Fixing the errors that were introduced during the data entry process increases the accuracy of the research results. Accuracy is important because if a researcher publishes inaccurate results other researchers would not be able to replicate those results and draw the same conclusions. The purpose of this study is to compare the accuracy of four different data checking methods: double entry with one person, double entry with two people, visual checking, and solo read aloud. So far, previous research has shown that double entry is more accurate than visual checking (Barchard & Pace, 2011) and partner read aloud (Kawado, Hinotsu, Matsuyama, Yamaguchi, Hashimoto, & Ohashi, 2003). Although there has not been many studies done on the comparison of these four methods and only one study has used solo read aloud, double entry has been shown to produce the highest quality data. I therefore hypothesize that the two double entry methods will have the highest accuracy.

Four hundred undergraduates will participate in this study for the return of course credit. Each participant will be randomly assigned to one of the four data checking methods. During the study, participants will first practice their assigned data checking method by entering and checking the data on a few data sheets. In the main part of the study, they will enter and check the data on additional data sheets. A total of 50 data sheets will be used. Each data sheet contains multiple sections. In each section, answers have been filled in so that they are ready to be entered into the computer. The participants will enter the data from these sheets into an Excel file. After we have gathered all of the data we will calculate the number of errors for each method. Our study will run at a future date and we will then be able to analyze our results.

Introduction

Data entry can be a tedious and daunting task where many errors can be made. Imagine having a stack of data sheets with more than fifty numbers on each of them. This is data that you have spent months collecting. Each number is an important piece of information for your study. You start to enter the data into the computer. After an hour or two of typing, you become tired and your typing becomes careless. You accidently enter a wrong number, then another. Such little mistakes can completely change your statistical results and your substantive conclusions (Barchard & Pace, 2011). The only way to be sure your results are accurate is to check the data. However, checking your data is time consuming and can cost a lot of money. You want to use the data checking method that is the most accurate. The purpose of this study is to compare the accuracy of four data checking methods.

There are four common data checking methods: single person double entry, two person double entry, read aloud, and visual checking. The single person double entry method consists of the one person entering and checking data. The two person double entry method has one person entering the data and a second person entering the data a second time and checking that they match. The read aloud method has one person entering the data and either the same person (solo read aloud) or a different person (partner read aloud) checking the data by reading it aloud. The visual checking method consists of one person entering the data and checking the data visually. One study that compared three different data checking methods found that two-person double entry produces fewer errors but takes longer than other data checking methods (Barchard & Verenikina, 2013). Through comparing the accuracy of the four different data checking methods this study will be able to identify which method produces the fewest errors.

Literature Review

Research data can help research solve many problems throughout the world. However, just one or two serious data entry errors can completely alter and invalidate a statistical analysis (Barchard & Pace, 2011). A single data entry error can make a moderate correlation turn to zero or make a significant t-test non-significant (Barchard & Pace, 2011). If a researcher produces inaccurate results, other researchers would not be able to replicate the results and come to the same the same substantive conclusions. Therefore, the need for accuracy in the construction of the data sets must be a major concern for researchers (Atkinson, 2012).

Many variables can cause the data entry errors. These include the abilities and characteristics of the data entry personnel, the format of questionnaires and database screens, and the working environment (Atkinson, 2012). Errors can occur because the enterer simply became tired or because an experienced data entry clerk was typing extremely fast and made the common mistake of hitting a wrong key. There are indeed so many variables that can cause data entry errors that it is probably impossible to prevent data entry errors entirely. Because of this, researchers have focused instead on improving methods of identifying and fixing data entry errors.

There have been many different data checking methods. These include solo read aloud, partner read aloud, visual checking, and double entry with one or two people. Double entry has consistently been shown to be the most accurate (Barchard & Pace, 2011; Gibson, Harvey, Everett, & Parmar, 1994; Paulsen, Overgaard, & Lauritsen, 2012) and has been defined as the definitive gold standard of good clinical practice (Paulsen et al., 2012). However, even though double entry may find the most errors, not all researchers have concluded that it benefited their research (Gibson et al., 2012). The biggest problem with the double entry method is that it takes up a greater amount of time. Double entry requires up to 37% more time than other data checking methods (Reynolds-Haertle & McBride, 1992).

Read aloud has produced the lowest accuracy rates (Barchard, Pace, & Verenikina, 2013). There are two possible explanations for this (Barchard et al., 2013). First, the procedures rely on attention and vigilance. If the person loses focus, they may overlook an error. Second, there is no way to know how often the data entry errors are overlooked and no way for a supervisor to identify errors that assistants overlooked.

Regardless of which data checking method is used, accuracy rates increase when the data checking person is someone different from the data entry person. One study found that read aloud detected about 60% of the errors when a different person did the checking, but only 39.9% of errors when the same person did the checking and the original entering. Similarly, double entry detected 88.3% of errors using different operators, but only 69% of the errors when the data checking person was the same as the original data entry person (Kawado et al., 2003). Because of this, we

hypothesize that double entry will be more accurate than read aloud or visual checking, and that double entry with two people will be more accurate than double entry with one person in the present study.

Several double entry data checking programs are available. PowerChecker (Beaty, 1999) works in conjunction with Microsoft Access. To use PowerChecker, a researcher will start by choosing a database to access from Microsoft Access database objects and will then be prompted to select a table containing data. The researcher will enter the data. If there is missing data, the researcher will press the F12 key. Any data entered into the PowerChecker is simultaneously compared to the data previously entered. If an error is detected, the person entering the data makes the appropriate changes and the computer will keep a time-stamped record of the change that has been made (Beaty, 1999). Poka-Yoke Double Entry System (Barchard et al., 2013) is an add-on for Microsoft Excel. It uses a four-step procedure. First, the researcher enters the data twice. Second, Poka-Yoke checks for mismatches. Third, the program checks for out-of-range values. Finally, the data checker examines the check columns to determine if there are any mismatch or out-of-range values, and fixes any errors that have been identified. In addition to these two programs, there is a free web-based system (Harris, Taylor, Thielke, Payne, Gonzalez, & Conde, 2009) and a stand-alone program (Gao et al., 2008; Lauritsen, 2000-2008). Finally, commercial statistical packages such as SPSS and SAS allow double entry. If our hypothesis is correct and double entry is indeed more accurate than the other techniques, the easy availability of high quality double entry programs will be essential for its implementation in a wide variety of research labs.

Method

Participants

There will be 100 participants for each data checking method, giving a total of 400 participants. Participants will be undergraduates at the University of Nevada, Las Vegas. Participants will be recruited from the Department of Psychology subject pool.

Materials

The participants in our study will enter and check the data that are given on the data sheets we hand to them. The data sheets will be labeled "Animal Emotions Study". Each sheet will have a three- digit ID number on the top left hand corner. After that, there are six sections. The demographic section will start with the information containing age: this will be a two-digit number. There will be a question of "What is your sex?" It will be chosen as *male* or *female*. If the sheet states *male* then the participant will be asked to enter that as a capital M. If the sheet states *female* then the participant will be asked to enter that as a capital F. There will be a question of "Which do you prefer?" The data sheet has three choices: *cats*, *dogs*, and *no preference*. If the data sheet says *cats* the participant will be asked to enter that as C. If the data sheet states *dog* the participant will be asked to enter that as D. Finally, if the data sheet says *no preference* then the participant will be asked to enter that as N.

Demographics

Age: 26

Sex: Male Female

Which do you prefer? Cats Dogs No preference

Rating Scales

How much is each emotion expressed by the following phrases?

Quivering horse

not at all happy	1 — 2 — 3 — 4 — 5	extremely happy
not at all sad	1 — 2 — 3 — 4 — 5	extremely sad
not at all angry	1 — 2 — 3 — 4 — 5	extremely angry
not at all scared	1 — 2 — 3 — 4 — 5	extremely scared

Charging elephant

not at all happy	1 — 2 — 3 — 4 — 5	extremely happy
not at all sad	1 — 2 — 3 — 4 — 5	extremely sad
not at all angry	1 — 2 — 3 — 4 — 5	extremely angry
not at all scared	1 — 2 — 3 — 4 — 5	extremely scared

Moaning seal

not at all happy	1 — 2 — 3 — 4 — 5	extremely happy
not at all sad	1 — 2 — 3 — 4 — 5	extremely sad
not at all angry	1 — 2 — 3 — 4 — 5	extremely angry
not at all scared	1 — 2 — 3 — 4 — 5	extremely scared

Bouncing kittens

not at all happy	1 — 2 — 3 — 4 — 5	extremely happy
not at all sad	1 — 2 — 3 — 4 — 5	extremely sad
not at all angry	1 — 2 — 3 — 4 — 5	extremely angry
not at all scared	1 — 2 — 3 — 4 — 5	extremely scared

Frolicking kangaroos

not at all happy	1 — 2 — 3 — 4 — 5	extremely happy
not at all sad	1 — 2 — 3 — 4 — 5	extremely sad
not at all angry	1 — 2 — 3 — 4 — 5	extremely angry
not at all scared	1 — 2 — 3 — 4 — 5	extremely scared

Over

The second section on the data sheet will have the title "Rating Scales". It starts with the question "How much is each emotion expressed by the following phrases?" Then there are five phrases that describe animals (e.g., frolicking kangaroos). Participants will rate each type of animal on four emotions: happy, sad, angry, and scared. The rating scale goes from 1 = not at all to 5 = extremely.

The third section is titled "Categorical Variables." This section is on the second side of the data sheet. This is where a color and animal have been combined (e.g., Orange cat). Next to each phrase, there will be seven different emotions: happy, sad, angry, scared, jealous, surprised, and bored. The data sheet will have one or two of these emotions circled. For example, *red beetle* may have the emotions *angry* and *scared* circled. If the data sheet has one emotion circled then participants will enter the following numbers: Happy = 1, Sad = 2, Angry = 3, Scared = 4, Jealous = 5, Surprised = 6, and Bored = 7. If the data sheet has two emotions circled then the participant will enter that as a 9.

The fourth section is titled "Open-Ended Questions". Underneath the section title is the question, "What emotions are expressed by each of the following phrases?" In this section an action is combined with an animal (e.g., leaping puppies). Each data sheet will have different words written next to these phrases. For example, one data sheet might saw "happy, excited, and eager" and another might say "The puppies are really excited." The participants will enter that data from this section by typing exactly what is written on the data sheet.

The fifth section it titled "Culture Information." There are six questions in this section. The first question is, "Country where you were born?" Next to this question, the data sheet will have the name of a country. For example, the data sheet could say United States or Zimbabwe. The second question is, "First Language." Written next to the question will be answers like English or Spanish. Next there is the subtitle, "How comfortable are you with English?" Participants will answer four questions, to say how comfortable they are with reading, writing, speaking, and listening in English. The rating scale is 1 to 10, with 1 being not at all comfortable and 10 being very comfortable.

The final section of the data sheet is titled "Follow-up." This section asks, "May we contact you for a follow up study?" The data sheet will have either *yes* or *no* circled. If *yes* is circled then the participant will be asked to enter in a 1. If *no* is circled the participant will enter a 2. The final question on the data sheet is, "If yes, please provide an email address". There will be an email address given below the question. The participant will be asked to enter the email exactly how it is written.

There will be a total of 50 data sheets. Of these, 25 have been designed to be easy to enter and 25 have been designed to be hard to enter. Easy data sheets will have short to medium responses without spelling or grammar errors. These responses will make sense, given the question on the data sheet. Questions that have numbers will be short and easy to identify. Difficult data sheets will have long responses with grammar, punctuation, and spelling errors. These responses will not necessarily make sense to the participant. Numbers will be long and multiple answers may be chosen for the Rating Scales and Categorical Variables sections. All sections that have written responses (i.e., Open-Ended Questions, country where you were born, first language, email address) will be typed in fonts. Easy data sheets will have easy fonts, such as ones that look like printing. Difficult data sheets will have more challenging fonts, such as ones that look like script.

Group Code G77-MA-FC

Categorical Variables

Which emotion is expressed by each of the following phrases?

Black dog	(happy)	sad	angry	scared	jealous	surprised	bored
White dove	(happy)	sad	angry	scared	jealous	surprised	bored
Yellow duckling	(happy)	sad	angry	scared	jealous	surprised	bored
Grey wolf	happy	(sad)	angry	scared	jealous	surprised	bored
Golden monkey	(happy)	sad	angry	scared	jealous	surprised	bored
Red beetle	happy	sad	(angry)	scared	jealous	surprised	bored
Green lizard	(happy)	sad	angry	scared	jealous	surprised	bored
Pink flamingo	(happy)	sad	angry	scared	jealous	surprised	bored
Orange cat	(happy)	sad	angry	scared	jealous	surprised	bored
Ivory parrot	(happy)	sad	angry	scared	jealous	surprised	bored
Blue frog	happy	(sad)	angry	scared	jealous	surprised	bored
Purple butterfly	(happy)	sad	angry	scared	jealous	surprised	bored

Open-Ended Questions

What emotions are expressed by each of the following phrases?

Roaring bull	angry, mad
Trembling rabbit	scared, shaking
Leaping puppies	playful, excited
Sleeping pony	tired, sad, calm
Flying frog	joy
Running pig	all fired up
Jumping fox	glad

Culture Information

Country where you were born: Chile

First language: Spanish

How comfortable are you with English?

Reading	Not at all	0 — 1 — 2 — 3 — 4 — 5 — 6 — 7 — 8 — 9 — (10)	Very comfortable
Writing	Not at all	0 — 1 — 2 — 3 — 4 — 5 — 6 — 7 — 8 — 9 — (10)	Very comfortable
Speaking	Not at all	0 — 1 — 2 — 3 — 4 — 5 — 6 — 7 — 8 — 9 — (10)	Very comfortable
Listening	Not at all	0 — 1 — 2 — 3 — 4 — 5 — 6 — 7 — 8 — 9 — (10)	Very comfortable

Follow-up

May we contact you for a follow up study? (Yes) No

If yes, please provide your email address theman@gmail.com

Procedures

Participants will be tested individually and will be supervised by a trained research assistant. Participants will start by reading the consent form. When the participant agrees to the consent form then the study is able to proceed. All participants will watch a video on how to use Excel. Participants will then watch a video tutorial of how to enter easy data. Participants will practice entering data using five data sheets. Next, each participant will be assigned to one of four data checking methods. Participants will then watch a video on how to check data using their assigned method, and practice checking data using five different data sheets. Now that participants know how to enter and check data, they will complete part 1 of the study by entering and checking 15 easy data sheets. Participants will be given a five minute break. After the break, participants will all watch a video on how to enter difficult data. Participants will practice entering difficult data using five data sheets. Then participants will practice checking difficult data using five different data sheets. Finally, in part 2 of the study, participants will enter and check 15 difficult data sheets.

The four data checking methods that a participant can be assigned to are one-person double entry, two-person double entry, read aloud, and visual checking. First, in one-person double entry, a participant will enter the data into Excel and then enter the data a second time. Excel will identify data mismatches between the two entries and any values that are outside the allowable range. Second, in two-person double entry, one participant will enter the data and then a different participant will enter the data a second time. The second person will then check the data, using the same procedures as were used in one-person double entry. Third, in read aloud, a participant will enter first the data. To check the data, the participant will read the data sheet out loud and then visually check the data in the Excel file. Lastly, in visual checking, the participant will enter the data. The participant will check the data by looking back and forth between the data sheet and Excel file. When errors are found using any of the methods, the participant will change the error to the correct information.

Measures

Accuracy will be measured by the numbers of errors on the participant's final Excel sheet. An error is defined as a discrepancy between the Excel sheet and what was actually on the data sheets. The data checking method that produces the greatest number of errors will be considered the least accurate. The method that produces the least number of errors will be considered the most accurate data checking method.

At the end of this study, participants will be given an evaluation form. The participant will rate the data checking method they used on 16 different emotion words, such as boring, calming, and frustrating. For each emotion word, the participant will be asked to use a five-point scale, to indicate how much they agree that this adjective describes the data checking method. The rating scale they will use is SD = Strongly Disagree, D = Disagree, N = Neutral, A = Agree, and SA = Strongly Agree.

Data Analysis

To compare the accuracy of the four data checking methods, an ANOVA will be calculated. The independent variable will be the group each participant belongs to (one-person double entry, two-person double entry, solo read aloud, or visual checking). The dependent variable will be the number of errors left in the Excel sheet after the participant has completed entering and checking data.

Discussion

Unlike previous studies, this study will compare four data checking methods simultaneously. One method that this study includes is solo read aloud method, for which there has been very little published research. Only a single study has examined solo read aloud and in that study, it was

only compared to only one other data checking method: double entry (Kawado et al., 2003). Moreover, that study used only two participants. In our study, we will be comparing 100 participants in solo read aloud to 300 participants in the other three data checking methods.

This study is also unique in that it will compare two double entry methods: one-person double entry and two-person double entry. Little research has compared these two methods. The majority of research has looked at one-person double entry compared with read aloud and visual checking. Previous research has found double entry to take longer; however, double entry has also resulted in providing data entry with fewer errors (Barchard & Verenikina, 2013). This study will add to what is already known about double entry by identifying which double entry method is the most accurate.

Participants in this study will be similar to the types of people who typically enter data in psychology research studies. Participants in other studies used different qualifications. In Kawado et al, their participants were called operators. One operator had one year of experience and the other had two years of experience (2003). The typical data entry person in a psychology research lab will not have this kind of experience. Psychology research labs usually consist of undergraduate students who are learning data entry or are at least new to the data entry process.

One weakness of our study is that we are not including every possible data checking method: We are excluding partner read aloud. Partner read aloud is a data checking method similar to solo read aloud. In partner read aloud, there are two people checking the data. One person reads from the original data sheet while the other person visually checks the data in the Excel file. Partner read aloud has been excluded in order to simplify administration procedures and reduce the time it will take to complete the study. However, because we are excluding partner read aloud, we are not comparing all data checking methods that are available to researchers. A better study would be to compare the accuracy of all possible data checking methods.

In any study, there are likely to be differences between participants that might be relevant to the research question. Such differences could be confounded with group assignments and could influence the study conclusions. The standard solution is to use random assignment to groups: With a large sample size, the groups are likely to be equal on average, both on characteristics that are known to influence the dependent variables (such as previous data entry experience) and variables that the researchers have not identified as being relevant. In this study, we will use simple random assignment, wherein every participant is independently assigned to a group. We decided to use simple random assignment because this can be done easily by the computer. Because we are using simple random assignment, not all groups will have exactly 100 participants. This will result in less statistical power than a study that has exactly 100 participants in each condition.

In this study, we are including people who do have previous data entry experience and those who do not. If one method gets more of the people who aren't experienced, then that method will probably have more errors. We might conclude there is a difference in the data checking methods, when it might be a difference in the participants' previous experience. This is the rationale behind the use of random assignment in this study. However, a better method would be to include only people who have no data entry experience. This would make it easier to interpret our results because we could state more clearly who our participants were. Also, it would increase our statistical power, because it would control a random source of error. Alternatively, we could collect data from 100 people without experience and 100 people with experience in each of the four conditions – but the disadvantage of that is that it would be time-consuming to collect that data.

We are still in the process of designing this study. So far, we have almost finished designing the data sheets that participants will enter and check. We have created scripts for the Adobe Captivate videos that will be used to train the participants in each method. Next, we need to finalize all data sheets and all Captivate videos, and create a Qualtrics website that includes the consent form, links to the relevant videos and Excel files, and the evaluation form. Then we will be able to write an IRB proposal, print the 50 data sheets and place them in the two testing rooms, and train research assistants to administer the study. We expect to begin data collection this fall. We hope to finish data collection and calculate our results within the next four semesters.

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*, 1834-1839. doi:10.1016/j.chb.2011.04.004.

Barchard, K. A., Verenikina, Y., & Pace, L. A. (2013, Aug). *Poka-Yoke Double Entry System Version 2.1.43*. Excel file that allows double-entry data entry with checking for mismatches, out-of-range values, and out-of-list values. Available at http://faculty.unlv.edu/barchard/doubleentry/ or from Kimberly A. Barchard, Department of Psychology, University of Nevada, Las Vegas, Barchard@unlv.nevada.edu.

Barchard, K. A., & Verenikina, Y. (2013). Improving data accuracy: Selecting the best data checking technique. *Computers in Human Behavior, 29*, 1917-1922. doi:10.1016/j.chb.2013.02.021

Beaty, J. C. (1999). The PowerChecker: A Visual Basic program for ensuring data integrity. *Behavior, Research Methods, Innovation, & Computers, 31*, 737-740.

Day, S., Fayers, P., Harvey, D. (1998). Double data entry: What value, what price? *Controlled Clinical Trials, 19*, 15-24.

Gao, Q-B., Kong, Y., Fu, Z., Lu, J., Wu, C., Jin, Z-C., & He, J. (2008). EZ-Entry: A clinical data management system. *Computers in Biology and Medicine, 38*, 1042-1044. doi:10.1016%2Fj.compbiomed.2008.07.008

Gibson, D., Harvey, A. J., Everett, V., & Parmar, M. K. B. (1994). Is double data entry necessary? The CHART trials. *Controlled Clinical Trials, 15*, 582-488.

Harris, P. A., Taylor, R. Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*. 377-381. doi:_10.1016/j.jbi.2008.08.010

Healy, A. F., Kole, J. A., Buck-Gengler, C. J., & Bourne, L. E. (2004). Effects of prolonged work on data entry speed and accuracy. *Journal of Experimental Psychology: Applied, 10*, 188-199. doi:10.1037/1076-898X.10.3.188

Kawado, M., Hinotsu, M.D., 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

Lauritsen J. M. (Ed.) (2000-2008). *EpiData Data Entry, Data Management and Basic Statistical Analysis System*. EpiData Association: Odense Denmark. http://www.epidata.dk Accessed January 14, 2013.

Paulsen, A., Overgaard, S., & Lauritsen, J. (2012). Quality of data entry using single entry, double entry and automated forms processing--an example based on a study of patient-reported outcomes. *Plos ONE, 7*, 1-6. doi:10.1371/journal.pone.0035087

Reynolds-Haertle, R. A., & McBride, R., (1992). Single versus double data entry in CAST *Controlled Clinical Trials, 13*, 487-494. doi:10.1016/01972456(92)90205-E