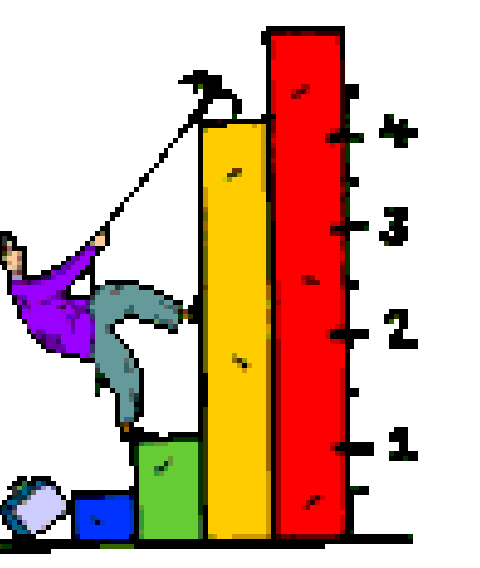


# Check Yourself: Perceived and Actual Accuracy of Data Checking Methods



Dalton J. Terry, Sarah A. Cobb, Ashley A. Anderson, and Kimberly A. Barchard  
University of Nevada, Las Vegas

## ABSTRACT

Data entry errors can have drastic effects on the quality of research. Small errors can cause significant t-tests to appear non-significant and turn moderate correlations to zero (Barchard & Pace, 2011). Therefore, researchers use various data checking methods to identify and correct data entry errors. Previous research shows that double entry is the most accurate but it takes longer than other data checking methods (Barchard & Pace, 2011; Paulsen, Overgaard, & Lauritsen, 2012). However, not everyone uses double entry. Researchers might use less effective methods because they are easier or faster, and because they think that their preferred method of data checking is highly accurate. The purpose of this study is to examine the relationship between actual and perceived accuracy.

A total of 27 undergraduates participated in this study in return for course credit. They were randomly assigned to one of four data checking methods: visual checking, double-entry, solo read-aloud, and partner read-aloud. When participants had finished checking the data, they rated the accuracy of the data checking method that they used. We compared these subjective opinions of accuracy with the actual number of correct entries.

We found a small non-significant correlation between perceived and actual accuracy ( $r(25) = .10, p = .61$ ). We conclude that while researchers have varied perceptions of the accuracy of the methods they use, these perceptions are not related to the actual effectiveness of their data checking methods. Researchers should not rely on subjective impressions of accuracy. They should use the data checking method that empirical research demonstrates is the most accurate: double entry. Several free double entry programs are available (Barchard, Verenikina, & Pace, 2013; Beatty, 1999; Harris, Taylor, Thielke, Gonzalez, & Conde, 2009; Lauritsen & Bruus, 2004).

## INTRODUCTION

There are possible sources of errors in every step of data management, and these errors may affect the results, reliability, and conclusions of a study (Kawado et al., 2003). Factors that can affect data accuracy include characteristics of the data entry personnel, fatigue, speed of entry, and the work environment (Atkinson, 2012). Errors can also have drastic effects on the quality of research. Even small data entry errors can cause t-tests to be non-significant and turn correlations to zero (Barchard & Pace, 2011). Because of the importance of accurate data, researchers use a variety of data checking methods to locate and correct data entry errors. In visual checking, a person visually compares the data that is shown on the computer screen to the original paper data sheet. In solo read-aloud, a person reads the original paper data sheet out loud and visually compares it to the data that is shown on the computer screen. In partner read-aloud, one person reads the data sheet out loud while another person visually compares it to the data that is shown on the computer screen. Finally, in double entry a person enters the data in twice; then the computer identifies mismatches between the two entries and identifies any values that outside the allowable range. Regardless of which of these methods is used to identify data entry errors, the person would correct the error by consulting the original data sheet and entering the correct value into the computer.

Double entry is considered the gold standard for transferring data to an electronic format (Paulsen, Overgaard, & Lauritsen, 2012) because it has higher accuracy rates than other data checking methods (Barchard & Verenikina, 2013). However, double entry also takes longer (Paulsen et al., 2012; Barchard & Verenikina, 2013). Researchers might use other data checking methods because they are faster and because they think that those techniques have high accuracy rates. The question is whether researchers' perceptions of accuracy match the actual accuracy of the data checking methods. The purpose of this study is to examine the relationship between actual accuracy and perceived accuracy in data checking. Previous research shows that people misjudge the accuracy of their data checking methods (Verenikina, Anang, Craun, & Barchard, 2012). The purpose of this study was to replicate that result in a new dataset, using a wider range of data checking methods. We hypothesize that there will be a small relationship between actual accuracy and perceived accuracy.

## METHOD

### Participants

A total of 27 undergraduates (18 female, 9 male) completed this study in return for course credit. They ranged in age from 18 to 46 (mean = 21.41, SD = 6.22). Participants identified themselves as follows: Caucasian (40.7%), Hispanic (18.5%), Asian (14.8%), African American (14.8%), Pacific Islander (7.4%), and other (3.7%).

## METHOD

### Measures

The actual accuracy of the data checking method was measured by the number of correct entries. An entry was considered correct if it matched the data on the paper data sheets. The perceived accuracy of the data checking method was self-reported by the participant using a five-point scale: Strongly Disagree = 1, Disagree = 2, Neutral = 3, Agree = 4, Strongly Agree = 5.

### Procedures

Participants completed one 90-minute testing session. First, participants watched a five-minute instructional video that explained how to use Microsoft Excel. Next, participants were randomly assigned to one of the four data checking methods (double-entry, visual checking, solo read-aloud, or partner read-aloud). Participants then watched another video detailing the method to which they had been assigned and practiced their assigned method by checking the data for five data sheets. During this practice session, the study administrator answered any questions the participant had.

An example data sheet is shown in Figure 1. For the Sex and Study Habits sections, participants were asked to convert the letters to numbers in order to increase the difficulty of data checking.

If a participant was assigned double entry, solo read-aloud, or visual checking, then they checked the data by themselves. If the participant was assigned partner read-aloud, the study administrator read the data to the participant. If what the administrator read matched what was on the screen, the participant was asked to say, "check." If the data did not match, the participant was asked to say, "verify." The study administrator would read the data again and if it did not match then the participant was asked to correct the mistake in the Excel file.

After completing the 5-sheet practice session, the participants checked 20 additional data sheets. This time, they were not offered any assistance. The data for this section is the data we used to determine actual accuracy. Lastly, participants took a brief survey which included information such as demographics and their perception of the accuracy of the data checking method they used.

## RESULTS

We found a weak non-significant correlation between perceived accuracy and actual accuracy ( $r(25) = .10, p = .61$ ).

Figure 1 Example Data Sheet

The Learning Study	
ID: 739925	
Sex: (M) F	
<b>Learning Style</b>	<b>Study Habits</b>
1. 1 2 3 4 5	1. SD D N A SA
2. 1 2 3 4 5	2. SD D N A SA
3. 1 2 3 4 5	3. SD D N A SA
4. 1 2 3 4 5	4. SD D N A SA
5. 1 2 3 4 5	5. SD D N A SA
6. 1 2 3 4 5	6. SD D N A SA
7. 1 2 3 4 5	7. SD D N A SA
8. 1 2 3 4 5	8. SD D N A SA
<b>Spelling Test</b>	<b>Math Test</b>
1. ACCOMMODATE	1. 156
2. AMATEUR	2. 235
3. CALENDAR	3. 485
4. CEMETERY	4. 493
5. CONSHENCE	5. 364
6. EMBARRASS	6. 327
7. EXHILARATE	7. 203
8. MAINTAINANCE	8. 347

## DISCUSSION

The purpose of this study was to examine the relationship between actual accuracy and perceived accuracy of four data checking methods. We found that the correlation was small and non-significant. We conclude that researchers are not good judges of the accuracy of their data checking methods. They should not rely upon their subjective opinions of accuracy. Instead, they should use methods for which empirical proof of accuracy is available.

These results should be considered tentative. This study did not take into account various factors that could contribute to data checking accuracy, such as prolonged work fatigue (Healy, Kole, Buck-Gengler, & Bourne, 2004) and level of data entry expertise (Atkinson, 2012). If those variables were added as covariates, it is possible that a moderate (or even strong) relationship between perceived and actual accuracy could result. Until such research is conducted, however, we must conclude that there is no evidence that researchers are able to judge the accuracy of the data checking methods they use. Researchers should limit themselves to methods with strong empirical evidence. Research has consistently shown that double entry is the most accurate (Barchard & Pace, 2011; Barchard & Verenikina, 2013; Paulsen et al., 2012). Several free double entry programs are available (Barchard, Verenikina, & Pace, 2013; Beatty, 1999; Harris, Taylor, Thielke, Gonzalez, & Conde, 2009; Lauritsen & Bruus, 2004).

Additional research is needed to determine why researchers select particular data checking methods. What factors most influence researchers to use a particular method? Accuracy (or perceived accuracy)? Convenience? Speed? Cost? Habit? By figuring out what matters most to researchers, we can perhaps encourage them to use the data checking methods that have the highest accuracy.

## 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. (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.
- 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/>
- Beatty, J. C. (1999). The PowerChecker: A visual basic program for ensuring data integrity. *Behavior, Research Methods, Innovation, & Computers, 31*, 737-740.
- 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. & Bruus M. (2004). *EpiData (version 3.02): A comprehensive tool for validated entry and documentation of data*. The EpiData Association, Odense Denmark. Available at <http://www.epidata.dk/>
- 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
- Verenikina, Y., Anang, C., Craun, E. A., Barchard, K. A. (2012, April). *Misperceptions of data checking*. Poster presented at the Western Psychological Association Annual Convention, San Francisco, CA.

