Prevent Data Entry Catastrophes: Use Double Entry
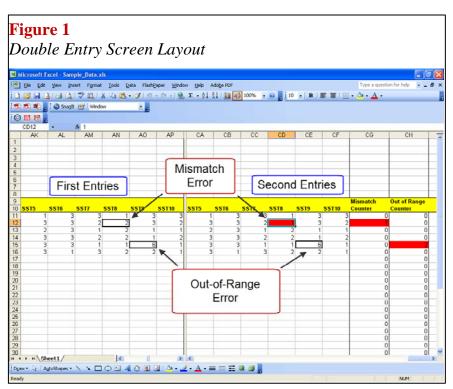
Kimberly A. Barchard
Department of Psychology
University of Nevada, Las Vegas

Larry A. Pace
Department of Psychology
Anderson University

Reference: Barchard, K.A. & Pace, L.A. (2009, April). *Prevent data entry catastrophes: Use double entry.* Paper presented at the Western Psychological Association Conference, 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

Figure 1
Double Entry Screen Layout

Abstract

Data entry errors can have catastrophic effects on the results of a statistical analysis. A single data entry error can make a moderate correlation turn to zero and make a significant t test non-significant. The purpose of this paper was to compare the accuracy of three data entry methods. A total of 197 undergraduates were randomly assigned to one of three conditions: double entry with automatic checking for mismatches and out-of-range values, visual checking of the entries against the original paper sheets, and single entry (a control condition). After receiving training in their assigned technique, participants entered 30 data sheets that each contained six types of data. Double entry was significantly more accurate than visual checking overall and for three of the six data types and resulted in 28 times fewer errors. Future research should compare double entry done by one person with double entry done by two people and with visual checking done by two people. For now, researchers should abandon visual checking done by a single person, given its high error rate. A free double-entry system that includes checking for mismatches and out-of-range values will be available during the poster session.

Introduction

Data entry errors can have catastrophic effects on study results and conclusions. A single data entry error can make a moderate correlation turn to zero or make a significant t test non-significant. Just one or two serious data entry errors can completely alter (and invalidate) a statistical analysis (Kruskal, 1960; Velleman & Hoaglin, 1995; Wilcox, 1998). Because data entry errors can be so devastating, researchers sometimes spend considerable effort to identify and correct the most severe errors. Preventative efforts include doing all data entry oneself, entering data twice, and checking entries visually (Beaty, 1999; Cummings & Masten, 1994; Winkler, 2004); corrective efforts including using graphs and diagnostic statistics to identify outliers (Tukey, 1977). The purpose of this paper is to compare two data entry methods that are intended to eliminate data entry errors at their source.

There are two common methods of preventing and catching data entry errors. In single entry with visual checking, the data entry person enters the data once. Afterwards, the same person visually compares entries with the original paper measures. In double entry with checking for mismatches and out-of-range values, data

are entered twice. The computer compares these entries to identify mismatches, and also identifies values outside the allowable range. The data entry person then corrects the errors.

The purpose of this study is to compare these two techniques to each other and to a control condition in which data are entered only once. Small-sample medical research has shown that double entry is more accurate than single entry (Reynolds-Haertle & McBride, 1992) and visual checking (Kawado et al., 2003). The current study extends that research by using a large sample of data entry personnel who are similar to the volunteers used in academic research, and by using six types of data that are commonly encountered in psychological research.

Method

Participants

A total of 197 undergraduate students participated in this study in return for course credit. None of these students had done data entry before.

Procedures

Data were collected during 90-minute one-on-one supervised sessions. Because data entry was completed using Microsoft Excel, participants first watched a short video on how to use Excel. Next, the computer randomly assigned participants to one of the data entry methods, and showed participants a video on that method. The first group (double-entry) was taught to enter the data twice and to locate and correct their errors using mismatch and out-of-range counters built into the worksheet. See Figure 1. The second group (visual checking) was taught to enter the data once and to check the data visually by comparing the typed entries with the original paper sheets. The third group (single entry) was taught to enter the data once; they were told that accuracy was more important than speed and to please be as accurate as they could. Next, all participants completed a practice session where they entered five data sheets, and the study administrator corrected any procedural errors. Finally, participants completed the main data entry, which consisted of 30 data sheets. Afterwards, participants evaluated the data entry technique using an 11-adjective scale.

To mimic the data entry tasks that research assistants complete, each data sheet contained six types of information: an ID number for the hypothetical participant, Sex, and four 10-item measures that used different response scales (letters or numbers, with 3 or 5 possible responses). To increase the difficulty of the data entry task for some of these scales, participants were instructed to type only numbers. See the example data sheet.

Results

Time

Double entry took 28% longer than visual checking, which took 25% longer than single entry. Specifically, double entry took 48.7 minutes on average; visual checking took 38.1 minutes; and single entry took 30.4 minutes.

Accuracy

Double entry was more accurate than visual checking and single entry. As shown in Table 1, there were significant differences overall and for four of the six types of data. Furthermore, for Sex, Family Background, School Experiences, and the overall accuracy, Tukey's HSD showed that double entry was significantly more accurate than visual checking ($p < .05$). For one additional type of data (Social Skills Test), the differences approached significance ($p = .062$) and, once again, double entry was more accurate than the other methods.

All three data entry techniques had high accuracy rates, which could obscure differences between them. We therefore calculated the average number of errors that participants made across the 30 data sheets. Participants in the double entry condition made an average of 0.38 errors. In visual checking, participants made an average of 11.09 errors. In single entry, participants made an average of 11.97 errors. Thus, visual checking resulted in 28 times more errors than double entry. See Figure 2.
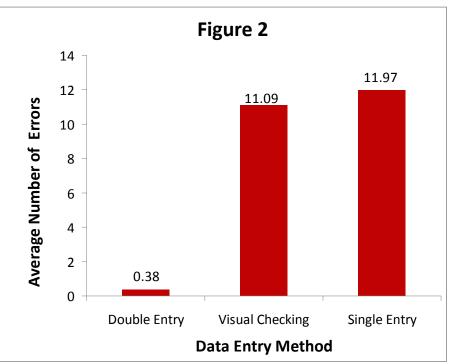
Visual checking was slightly more accurate than single entry, but this difference did not reach statistical significance (Tukey's HSD $p = .971$). We conclude that visual checking is no more accurate than single entry. Therefore, we do not recommend that researchers use visual checking by a single person, given that it takes more time than single entry and has no apparent benefit.

Figure 2

Table 1

Average Accuracy of the Three Data Entry Methods

Data Type	Double Entry	Visual Checking	Single Entry	ANOVA
ID	1.0000	.9986	.9968	$F(2, 135) = 0.87, p = .423$
Sex	1.0000	.9915	.9962	$F(2, 135) = 4.27, p = .016$
FB 5 letters	.9997	.9885	.9849	$F(2, 135) = 5.43, p = .005$
Ex 5 numbers	1.0000	.9913	.9909	$F(2, 135) = 3.29, p = .040$
SE 3 letters	.9992	.9894	.9896	$F(2, 135) = 4.69, p = .011$
SST 3 numbers	.9999	.9950	.9956	$F(2, 135) = 2.83, p = .062$
Overall	.9997	.9912	.9905	$F(2, 135) = 4.93, p = .009$

Note. FB = Family Background. Ex = Extraversion. SE = School Experiences. SST = Social Skills Test.

Catastrophic Errors

Next we examined the effect of catastrophic data entry errors and low accuracy rates on research results. The 197 participants in our study are taking the role of research assistants, each of whom is entering the complete data set for an imaginary study with 30 participants. When our participants make data entry errors, this is mimicking a situation where a research assistant makes data entry errors and the published results are wrong. Of the 197 participants, three made catastrophic errors such as entering the scales in the wrong order (these participants were excluded from the main analyses), and six additional participants had accuracy rates of 95% or less.

We examined the effect of catastrophic errors on three statistics: internal consistency, correlations, and an independent sample *t* test. For each statistic, we calculated the "true" values of the statistics when the correct data were used – the data that was actually given on the data entry sheets. We then compared these to the "observed" values of the statistics that were calculated using the data entered by participants.

Table 2
Effect of Data Entry Errors on Internal Consistency, Selected Participants

Participant ID		Family Background 5 letters	Extraversion 5 numbers	School Experiences 3 letters	Social Skills Test 3 numbers
Correct Values		.67	.63	.54	.55
172439	Scales in wrong order	.62+	-.24++	-.17++	-.03++
27578	87% accuracy	.24++	.50++	.39++	.50+
188413	94% accuracy	.60+	.63	.23++	.46+

+ Observed value differs from true value by at least .05.
++ Observed value differs from true value by at least .10.

Table 3
Effect of Data Entry Errors on Correlations, Selected Participants

Participant ID		Correlation	
		E and SE	SE and SST
Correct Values		.67**	.41*
172439	Scales in wrong order	.58**+	-.00++
27578	87% accuracy	.45*++	.12++
188413	94% accuracy	.49**++	.39**

* *p* < .05. ** *p* < .01.
+ Observed value differs from true value by at least .05.
++ Observed value differs from true value by at least .10.

Table 4
Effect of Data Entry Errors on Independent Sample t-test, Selected Participants

Participant ID	Family Background 5 letters		Extraversion 5 numbers		School Experiences 3 letters		Social Skills Test 3 numbers	
	t test	Effect size	*t* test	Effect size	*t* test	Effect size	*t* test	Effect size
Correct Values	3.05**	3.41	3.13**	2.90	2.07*	2.17	2.04	2.89
172439	0.22	0.10++	-1.52	-2.39++	0.04	0.03++	1.74	1.22++
27578	2.29	2.11++	1.22	0.96++	2.33*	2.61+	2.64*	3.76+
188413	3.66**	4.45++	3.07**	2.96	2.62*	2.91+	1.50	2.23+

* *p* < .05. ** *p* < .01.
+ Observed effect size differs from true value by at least .50.
++ Observed effect size differs from true value by at least 1.00.
Note. Effect size = (mean for men – mean for women) / pooled variance.

Table 5

Frequency of Catastrophic Errors for Each Data Entry Method

	Data Entry Method		
Error?	Double Entry	Visual Checking	Single Entry
None	61	59	68
Entered incorrect ID numbers	0	1	0
Entered scale in the wrong order	0	2	0
Accuracy rate 95% or less	0	4	2

These errors had strong effects on internal consistencies, correlations, and independent-samples *t* tests. For example, internal consistencies were sometimes negative (see Table 2 for selected results) and correlations were sometimes reduced to zero (see Table 3). In two cases, a strong positive effect size was changed into a strong negative effect size (see Table 4). Such data entry errors would invalidate the results of a study. Seven of these nine error-prone participants were in the visual-checking condition; none in the double-entry condition (see Table 5).

Only two of the nine error-prone participants entered a large number of values that were outside the allowable ranges for those variables. The two participants who reversed the order of the Extraversion and School Experiences scales – participants 61321 and 172439 – entered 129 and 132 out-of-range values, respectively. The other error-prone participants entered no more than 3 out-of-range values. If a supervisor corrected all out-of-range values in these datasets, most of the data entry errors would remain.

Subjective Opinions

Subjective opinions of the three data entry methods were significantly different on two adjectives. Single entry was considered more pleasant than double entry (single entry mean 3.46; double entry mean 2.75; Tukey's HSD $p < .001$) and visual checking was considered more frustrating than both single entry and double entry (single entry mean 2.49; double entry mean 2.54; visual checking mean 3.10; Tukey's HSD $p < .05$ for both comparisons). When we controlled statistically for the amount of time to complete the data entry (see Table 6), the differences in pleasantness disappeared; however, visual checking remained more frustrating than double entry (Tukey's HSD $p = .011$).

Table 6

Average Unstandardized Residuals for Subjective Opinions, Time Partialled Out

Adjective	Double Entry	Visual Checking	Single Entry	ANOVA
Accurate	.01	.03	-.04	$F(2, 161) = 0.12, p = .883$
Reliable	-.03	.01	.02	$F(2, 161) = 0.04, p = .965$
Enjoyable	-.15	.06	.06	$F(2, 161) = 0.67, p = .513$
Fun	-.09	.04	.03	$F(2, 161) = 0.22, p = .802$
Pleasant	-.23	.03	.15	$F(2, 161) = 2.05, p = .132$
Relaxing	-.05	-.02	.06	$F(2, 161) = 0.16, p = .849$
Satisfying	.02	.02	-.03	$F(2, 161) = 0.04, p = .963$
Boring	-.02	.04	-.02	$F(2, 161) = 0.05, p = .952$
Frustrating	-.29	.36	-.12	$F(2, 161) = 4.86, p = .009$
Painful	-.01	.07	-.06	$F(2, 161) = 0.20, p = .817$
Tedious	.01	-.15	.15	$F(2, 161) = 1.20, p = .303$
Total Eval	-.01	-.02	.03	$F(2, 162) = 0.10, p = .903$

Discussion

Some methods of identifying and correcting data entry errors are better than others. Visual checking was not significantly more accurate than single entry, despite the extra time involved. In contrast, double entry resulted in significantly fewer errors than visual checking for three of the six types of data examined. Furthermore, the differences between these techniques were large: visual checking had 28 times more errors than double entry. Thus, although double entry took 25% longer than visual checking, we conclude that the substantial increase in accuracy is easily worth the additional time.

Future research should compare double entry completed by one person (which was examined here) with other data entry techniques. For example, data could be entered by one person but then visually checked by someone else, or it could be visual checked by having one person read the entries out loud while another person reads the original sheets. These techniques might result in higher accuracy rates than visual checking by the same person. Also, data could be entered twice by two different people and then compared. This might result in similar accuracy levels compared to double entry done by one person but have higher subjective ratings, because each data entry task will be more like single entry, which was rated as the most pleasant and least frustrating.

Unless future research shows that some form of visual checking performs substantially better than it did here, it should be abandoned. Studies that have examined data quality (Kawado, et al., 2003; Reynolds-Haertle & McBride, 1992) have unanimously found that double entry is the most accurate method. The subjective opinion of researchers and data entry personnel that visual checking is a highly accurate method is contradicted by every empirical study on this topic. Double entry systems should be employed in every research lab.

Commercial double entry systems are available from SPSS and SAS, and free double entry systems are available as a stand alone program (Lauritsen & Bruus, 2004) or as free add-ons for Access (Beaty, 1999) and Excel (Barchard & Pace, 2008; in press). The Barchard and Pace double-entry system will be available for free during the poster session.

References

Barchard, K.A. & Pace, L.A. (in press). Meeting the challenge of high quality data entry: A free double entry system. *International Journal of Services and Standards*.

Barchard, K.A. & Pace, LA. (2008). *Nearly Perfect Data Entry Template Version 1.12.* Excel file that shows how to set up a double-entry data entry system for any number of measures and items. Available from barchard@unlv.nevada.edu.

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

Cummings, J., & Masten, J. (1994). Customized dual data entry for computerized data analysis. *Quality Assurance*, *3*, 300-303.

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.

Kruskal, W. H. (1960). Some remarks on wild observations. *Technometrics, 2.* Available at http://www.tufts.edu/~gdallal/out.htm Accessed June 25, 2008.

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/ Accessed on June 25, 2008.

Reynolds-Haertle, R. A., & McBride, R. (1992). Single versus double data entry in CAST *Controlled Clinical Trials*, *13*, 487-494.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Velleman, D.C. & Hoaglin, P.F. (1995). A critical look at some analyses of major league baseball salaries. *American Statistician, 49*, 277-285.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods. *American Psychologist, 53*, 300-314.

Winkler, W. E. (2004). Methods for evaluating and creating quality data. *Information Systems*, *29*, 531-550.