Comparing the accuracy and speed of four data-checking methods

Kimberly A. Barchard¹ · Andrew J. Freeman¹ · Elizabeth Ochoa¹ · Amber K. Stephens¹

Published online: 11 March 2019
© The Psychonomic Society, Inc. 2019

Abstract

Double entry locates and corrects more data-entry errors than does visual checking or reading the data out loud with a partner. However, many researchers do not use double entry, because it is substantially slower. Therefore, in this study we examined the speed and accuracy of solo read aloud, which has never before been examined and might be faster than double entry. To compare these four methods, we deliberately introduced errors while entering 20 data sheets and then asked 412 randomly assigned undergraduates to locate and correct these errors. Double entry was significantly and substantially more accurate than the other data-checking methods. However, the double-entry participants still made some errors. Close examination revealed that whenever double-entry participants made errors, they made the two sets of entries match, sometimes by introducing new errors into the dataset. This suggests that double entry can be improved by focusing attention on making entries match the original data sheets (rather than each other), perhaps by using a new person for mismatch correction. Solo read aloud was faster than double entry, but not as accurate. Double entry remains the gold standard in data-checking methods. However, solo read aloud was often substantially more accurate than partner read aloud and was more accurate than visual checking for one type of data. Therefore, when double entry is not possible, we recommend that researchers use solo read aloud or visual checking.

Keywords Data checking · Data entry · Double entry · Read aloud · Visual checking

Introduction

In the last 20 years, technological advances such as optical mark recognition and online surveys have allowed much data entry to be computerized, which increases both efficiency and accuracy. However, not all data entry can be automated: Behavioral observations, children's data, and scores from open-ended questions still need to be manually entered into a computer. Moreover, many of the data that theoretically could be entered automatically are not. For both financial and practical reasons, field data, surveys, and classroom exams are often completed on paper and then later entered into a computer.

Manual data entry inevitably leads to data-entry errors. In medical settings, data-entry errors can have catastrophic consequences for patients and are thankfully rare (Gibson,

Harvey, Everett, Parmar, & on behalf of the CHART Steering Committee, 1994): Estimates put the error rate around 0.20% (Reynolds-Haertle & McBride, 1992) or between 0.04% and 0.67%, depending upon the type of data (Paulsen, Overgaard, & Lauritsen, 2012). However, in research contexts, data-entry errors are more common. Error rates typically range from 0.55% to 3.6% (Barchard & Pace, 2011; Bateman, Lindquist, Whitehouse, & Gonzalez, 2013; Buchele, Och, Bolte, & Weiland, 2005; Kozak, Krzanowski, Cichocka, & Hartley, 2015; Walther et al., 2011), although error rates as high as 26.9% have been found (Goldberg, Niemierko, & Turchin, 2008). Even if only 1% of entries are erroneous, if a study contains just 200 items, manually entering the data could result in data-entry errors for almost every participant.

Simple data-entry errors, such as typing an incorrect number or skipping over a line, can drastically change the results of a study (Barchard & Pace, 2008, 2011; Hoaglin & Velleman, 1995; Kruskal, 1960; Wilcox, 1998). For example, they can reverse the direction of a correlation or make a significant *t* test nonsignificant (Barchard & Verenikina, 2013). In one study, data-entry errors sometimes increased sample means tenfold, made confidence intervals 17 times as wide, and increased correlations by more than .20 (Kozak et al., 2015). In clinical research, data-entry errors can impact study

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13428-019-01207-3>) contains supplementary material, which is available to authorized users.

✉ Kimberly A. Barchard
kim.barchard@unlv.edu

¹ University of Nevada Las Vegas, Las Vegas, NV, USA

conclusions, thus swaying the standard care of thousands of patients (Goldberg et al., 2008). Researchers therefore use a variety of strategies to prevent data-entry errors, such as numbering the items, using data-entry software that shows all items simultaneously, and entering data exactly as shown on the page (Schneider & Deenan, 2004). However, whatever steps are taken to ensure the initial data accuracy, researchers cannot know whether entries were entered correctly unless they check them.

Some researchers have advocated holistic data-checking methods, such as examining scatterplots and histograms (Tukey, 1977) and calculating univariate and multivariate statistics to detect outliers and other influential data points (Osborne & Overbay, 2004; Tabachnick & Fidell, 2013). However, these methods may not detect errors that fall within the intended ranges for the variables (Stellman, 1989). Therefore, item-by-item data checking is necessary to ensure that all data-entry errors are identified.

A variety of item-based data-checking methods can be used. In visual checking, the data checker visually compares the original paper data sheets with the entries on the computer screen. In solo read aloud, the data checker reads the original paper data sheets aloud and visually checks that the entries on the screen match. In partner read aloud, two data checkers are required: One reads the data sheets aloud, while another checks that the entries on the computer screen match.

Finally, in double entry, the data checker enters the data into the computer a second time, and the computer compares the two entries and flags any discrepancies (see Fig. 1); the data checker then examines the original data sheet to determine which entry is correct. The purpose of all these data-checking procedures is to identify and correct data-entry errors. It is therefore important to study the effectiveness and efficiency of these methods.

Double entry is recommended by many sources (Burchinal & Neebe, 2006; Cummings & Masten, 1994; DuChene et al., 1986; McFadden, 1998; Ohmann et al., 2011) because it is more accurate than visual checking and partner read aloud. For example, among experienced data enterers checking medical data, double entry detected 73% more errors than partner read aloud (Kawado et al., 2003), and among university students checking psychological data, double entry was three times as likely as visual checking and partner read aloud to correct every data-entry error (Barchard & Verenikina, 2013). In fact, double entry is as accurate as optical mark recognition and intelligence character recognition (Paulsen et al., 2012). However, double entry is more time-consuming than visual checking or partner read aloud (Barchard & Pace, 2011; Barchard & Verenikina, 2013; Kawado et al., 2003; Reynolds-Haertle & McBride, 1992), so researchers have sought alternatives. For example, visual checking can be augmented by graphical representations of the numbers being

Learning Style								Learning Style								Mismatch Counter	Out of Range Counter
LS 1	LS 2	LS 3	LS 4	LS 5	LS 6	LS 7	LS 8	LS 1	LS 2	LS 3	LS 4	LS 5	LS 6	LS 7	LS 8		
1	3	4	3	4	3	2	3	1	4	3	4	3	2	3	4	0	0
2	5	3	3	2	2	5	4	2	5	2	3	2	2	5	4	1	0
4	2	1	5	4	5	2	1	4	2	1	5	4	5	2	1	0	0
1	5	2	3	4	4	3	1	1	5	2	3	4	4	3	1	0	0
3	1	5	1	3	7	1	5	3	1	5	1	3	7	1	5	0	1
5	4	1	2	5	1	4	2	5	4	1	2	5	1	4	2	0	0

Fig. 1 Double-entry screen layout

checked: One study showed that this reduced initial data-entry errors by 60%, but unfortunately it had no noticeable effect on data-checking accuracy (Tu et al., 2016). Some researchers have created dynamic data-entry systems that only ask a portion of the values to be reentered: those entries that Bayesian analyses suggest are likely to be data-entry errors (Chen, Chen, Conway, Hellerstein, & Parikh, 2011). Although such systems save time and are better than no data checking, one study showed that they still left 22% of the errors in the dataset (Chen, Hellerstein, & Parikh, 2010).

In this study we examined the effectiveness of a new data-checking method: solo read aloud. Although substantial research has demonstrated that double entry is superior to visual checking and partner read aloud, no previous research has examined the accuracy of solo read aloud or compared its accuracy to that of other data-checking methods. Therefore, the purpose of our study was to compare solo read aloud to double entry, visual checking, and partner read aloud. On the basis of previous research, we predicted that double entry would be more accurate than visual checking and partner read aloud, but also more time-consuming. We made no predictions regarding solo read aloud, which had never been empirically tested.

The present study went beyond previous research in two additional ways. First, this study included a detailed examination of the types of errors that data-checking participants left in datasets, to provide guidance for future improvements to data-checking systems. As you will see, this novel analysis led to insights regarding further improvements in double-entry systems. Second, this study compared the accuracy of data-checking participants with previous data-entry experience to those without. Surprisingly, no previous research has examined the effect of experience on data-checking accuracy. This study examined whether data-entry experience increases speed, reduces errors, and changes subjective opinions of the data-checking system.

Method

Participants

A total of 412 undergraduates (255 female, 153 male, 4 unspecified) participated in return for course credit. They ranged in age from 18 to 50 ($M = 20.98$, $SD = 5.09$). They identified themselves as Caucasian (32.2%), Hispanic (26.7%), Asian (21.6%), African-American (9.7%), Pacific Islander (2.7%), and Other (0.5%).

We originally planned to collect data from about 400 participants, roughly 100 using each method. After 4.5 years of data collection, we had 412 participants. Because participants were assigned to the data-checking methods completely at random, there were slightly more participants using some

methods than others: double entry, 94; visual checking, 98; partner read aloud, 119; and solo read aloud, 101.

Of these 412 participants, 90 had previous data-entry experience. These 90 participants had between 4 h of data-entry experience and more than 2 years of full-time work (40+ h per week). About two-thirds of them (61) had more than 100 h of data-entry experience.

Materials

The 412 participants in our study were taking the role of research assistants, each of whom was checking the complete dataset for an imaginary study with 20 subjects. Before participants arrived for the study, the data sheets were entered into the computer. These data sheets contained six types of data (see Fig. 2): a six-digit ID code, a letter (M or F; labeled *Sex*), five-point numerical rating scales (labeled *Learning Style*), five-point alphabetical scales (SD D N A SA; labeled *Study Habits*), words in capital letters (often with spelling errors; labeled *Spelling Test*), and three-digit whole numbers (labeled *Math Test*).

When we entered these 20 data sheets, we deliberately introduced 32 errors (see Table 1). Thirteen of these errors

The Learning Study

ID: 346525

Sex: M (F)

<u>Learning Style</u>					<u>Study Habits</u>				
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

<u>Spelling Test</u>		<u>Math Test</u>	
1.	ACOMADATE	1.	165
2.	AMETURE	2.	234
3.	CALANDAR	3.	450
4.	CEMETARY	4.	473
5.	CONSCIENCE	5.	360
6.	EMBARRAS	6.	350
7.	EXILARATE	7.	230
8.	MAINTAINANCE	8.	324

Fig. 2 Example data sheet

Table 1 Errors researchers inserted into the Excel file for participants to locate and correct

Type of error	Number of instances	Example
Easy-to-Find		
Blank	1	It says nothing, when the data sheet says 260
Two responses in one cell	1	It says 290384 instead of 290
Letter instead of number	1	It says <i>n</i> instead of 1
Word instead of number	1	It says ACCOMADATE instead of 3
Entirely wrong word	7	It says CEMETARY instead of CALENDAR
Wrong number (out of range)	2	It says 7 instead of 2 (when correct numbers range from 1 to 5)
Hard-to-Find		
Reordered digits	4	It says 152 instead of 125
Repeated wrong digit	1	It says 496691 instead of 496991
Misspelled word	2	It says CALENDER instead of CALANDER
Wrong number (in range)	12	It says 5 instead of 2 (when correct numbers range from 1 to 5)

Adapted from Barchard and Verenikina (2013)

would be easy for data checkers to identify later: for example, entering a word when a number was expected or entering two numbers in a single cell. These entries were so obviously wrong that they could be detected by simply looking at the Excel sheet: A data checker would not even need to look at the data sheet. The other 19 errors were less obvious, and thus would be difficult for data checkers to identify from a superficial examination of the data: for example, entering an incorrect number that was still within the range for that variable. These entries were plausible values for those variables, so the data checker would only know that they were wrong by noticing that they did not match the data sheet.

These 32 errors represented 5% of the entries. This error rate is higher than the rates found in most previous research (e.g., Barchard & Pace, 2011). Using a larger number and variety of errors increased our ability to detect differences between the data-checking methods.

Procedures

Participants completed this study individually during single 90-min sessions supervised by trained research assistants. Because participants used Excel 2007 to check the data, they started the study by watching a video on how to use Excel. The video explained how to scroll vertically and horizontally, move between cells, enter and delete cell contents, and save and close the spreadsheet. To view this video (and the other videos used in this study), see the supplementary materials.

Next, the computer randomly assigned participants to one of the four data-checking methods and showed them a video on using that method to identify and correct data-entry errors. These videos provided an overview of the imaginary study. They showed an example paper data sheet with its 34 variables (ID, sex, and eight items for each of the four scales) and explained that the participant would use Excel to check data

entered by someone else. The videos also reviewed the Excel data file, explaining that each row contained the data for one subject and that each column provided a different piece of information about that subject.

The Excel sheet in the videos differed depending upon the data-checking method. The Excel sheet in the visual-checking, solo read-aloud, and partner read-aloud videos showed data for five subjects. The Excel sheet in the double-entry video did not show any data initially. This video showed participants how to enter the data themselves. Then it showed them the first set of entries, along with the mismatch counter and the out-of-range counter. The mismatch counter gives the number of nonidentical entries between the first and second entries. The out-of-range counter gives the number of entries that are outside the range for the variable (e.g., an entry of 8 on a five-point scale).

Most critically, the data-checking videos showed different methods of identifying and correcting data-entry errors. The visual-checking video told participants to check the data in Excel by visually comparing them to the paper data sheets. If the entry did not match the paper data sheet, participants were to correct the entry. The solo read-aloud video told participants to read the paper data sheets out loud and to check that the data in Excel matched. The partner read-aloud video told participants to listen as the researcher read the paper data sheet out loud; the participant was told to say *check* when the entries matched, and *verify* when they did not, in which case the researcher would read the data point a second time. During the two read-aloud videos, the computer took the roles of the participant and researcher: It read the data out loud and said *check* and *verify* as needed. During all three of these videos, the computer demonstrated how to check the data for an entire subject. If the Excel entries did not match what was on the data sheet, the computer demonstrated how to correct them.

Similarly, the double-entry video showed participants how to check and correct the data for an entire subject. First, the video told participants to enter the data themselves. Next, the video showed participants the mismatch counter and out-of-range counter and explained what they were. If either of these counters was nonzero, participants were told to check the original paper data sheet to determine the correct value. Finally, the video demonstrated correcting whichever Excel entry was incorrect (one example corrected the original entry, and another example corrected the second entry).

Several variations of double entry are used in real research (Kawado et al., 2003). In one method, a single person enters the data twice, and then that person (or another person) compares the two entries to each other and resolves discrepancies. In another method, *two* people enter the data into the computer, and then one of those two people (or another person) compares the entries to each other and resolves discrepancies. In this study, we used the second method: The research team entered the data into the computer before participants arrived; the double-entry participants then entered the data a second time. We used this method so that all participants—regardless of whether they were using double entry, visual checking, solo read aloud, or partner read aloud—would be responsible for identifying and correcting exactly the same errors.

After participants had finished watching the videos, they started checking the data. During the training phase, participants checked the data for five data sheets, and the researcher answered any questions they had and corrected any procedural errors they had made. During the testing phase, participants checked the data for 20 additional sheets without further interactions with the researcher.

After participants had completed the data checking, they provided their subjective opinions of the data-checking method they had used by rating their assigned method on 16 adjectives (i.e., *satisfying*, *comfortable*, *frustrating*, *pleasant*, *painful*, *boring*, *relaxing*, *accurate*, *enjoyable*, *tedious*, *uncomfortable*, *fun*, *annoying*, *calming*, *depressing*, and *reliable*) using a five-point scale from (1) *strongly disagree* to (5) *strongly agree*.

Statistical analyses

We examined the effect of data-entry experience and data-entry method on three criterion variables: time, number of data-checking errors, and subjective opinions. Time was the difference between the time when the participant loaded the Excel file whose data they were checking and the time when they closed that file and moved on to the follow-up questionnaire. The number of data-checking errors was the number of discrepancies between the participant's completed data file and the correct data. If the participant corrected each of the 32 original errors and did not introduce any new errors, the participant was

further coded as having completed perfect data checking. Subjective opinions were the participants' ratings on the 16-adjective measure. We used the total score from the 16 adjectives, after reverse coding the seven items containing negative adjectives.

We began by examining the effect of data-entry experience on these three criterion variables. We hypothesized that previous data-entry experience would improve both the speed and accuracy of data checking. We made no prediction about the effect of data-entry experience on subjective opinions of the data-checking methods.

We used different types of models for the different criterion variables. We selected these types of models on the basis of theoretical considerations (e.g., continuous vs. dichotomous criterion variables) and then checked to ensure that model assumptions had been met. For the criterion variable of time, which was continuous with a roughly normal distribution, we used multiple linear regression. For the number of data-checking errors, which was a count variable with a large number of zero values, we used negative binomial regression (as recommended by Cameron & Trivedi, 1998). For the dichotomous variable of whether the participant had perfect data checking, we used logistic regression with Nagelkerke's pseudo- R^2 as a measure of effect size. Finally, for the criterion variable of subjective opinions, which was a nearly continuous variable with a roughly normal distribution, we used multiple linear regression.

Next we examined the effect of data-checking method. We used the same three criterion variables and the same models. To take into account data-entry experience, we fit a hierarchical series of models. In the first model, the only predictor was data-entry experience. In the second model, data-checking method was added as a second predictor. In the third model, the interaction between data-entry experience and data-checking method was added as a third predictor. However, for each of the criterion variables, the interaction term was nonsignificant (all $ps > .05$). Therefore, we do not report the results of the third models.

For each criterion variable, we present the results in two tables. In the first table, we provide means and 95% confidence intervals for each combination of data-entry experience and data-checking method. The confidence intervals were constructed using percentile bootstrapping with 10,000 replications. In the second table, we present the hierarchical models that we used to examine the effects of data-entry experience and data-checking method. For convenience, we have indicated the results of these significance tests in the first table (with the means). Thus, the first table shows what we found, and the second table shows how we found it.

For each criterion variable, we fit the hierarchical models twice, to compare the data-checking methods to each other. First, because double entry is the current gold standard in data

checking, we compared double entry to the remaining three data-checking methods. These first analyses included all participants. Second, because solo read aloud has never been examined empirically, we compared it to partner read aloud and visual checking (double entry was excluded from this analysis because this comparison had already been completed). These second analyses included only the participants who used the solo read-aloud, partner read-aloud, and visual-checking methods.

Most of these models converged upon solutions without incident, providing useable results for testing the hypotheses above. However, when we were predicting the number of errors for variables with very few errors (i.e., sex), these models did not converge. For this variable, we are unable to make conclusions about the relative frequency of errors across data-checking methods, because so few errors were made.

Finally, to determine whether the data-checking participants were able to judge the accuracy of their work, we completed two analyses. First, we correlated subjective judgments of accuracy and reliability with actual accuracy. These subjective judgments of accuracy and reliability were obtained as part of the 16-item measure of subjective opinions. Actual accuracy was calculated as the number of correct entries in the Excel sheet after the participant had finished checking it. If participants were good judges of the quality of their data-checking efforts, these correlations should be large and positive. Second, we examined the effect of data-entry experience and data-checking method on subjective opinions of accuracy and reliability. If subjective opinions are a good way of evaluating data-checking quality, then the effects of data-entry experience and data-checking methods on subjective opinions should mirror their effects on actual errors. We therefore used the same models as we had used to compare actual errors: Model 1 examined the effect of data-entry experience on subjective opinions; Model 2 added the predictor of data-checking method; and Model 3 added the interaction term (which was not significant, so Model 3 is not reported).

We conducted a power analysis before beginning data collection. However, our analytic plan changed during

peer review, making those earlier calculations irrelevant. Therefore, we conducted a sensitivity analysis after the fact to determine what size of effect we would have been able to detect, given our sample size of 412 participants. Using the R statistical package *lmSupport* (Curtin, 2017), we found that we had power of at least .80 to detect small differences between the data-checking methods in terms of the total number of errors. Specifically, when predicting the total number of errors on the basis of data-entry experience (Model 1), we had 80% power to find effects as small as $\eta^2 = .0188$. When predicting the total number of errors on the basis of data-checking method while controlling for data-entry experience (Model 2), we had 80% power to find effects as small as partial $\eta^2 = .0261$. Finally, when predicting the total number of errors from the *interaction* between data-checking method and data-entry experience, while controlling for data-checking method and data-entry experience (Model 3), we had 80% power to find effects as small as partial $\eta^2 = .0263$.

Results

Data-entry experience

Time

Data-entry experience had no significant effect on the time it took participants to check the data. Across the four data-checking methods, participants with data-entry experience took 1.8% less time than participants without data-checking experience (see Table 2). However, this small effect for data-entry experience was nonsignificant [$F(1, 409) = 0.73$, $p = .395$; see Table 3, Model 1, for all participants).

Perfect data checking

Data-entry experience had no significant effect on the proportion of participants who created an error-free dataset (by correcting each of the original errors without introducing

Table 2 Means [95% confidence intervals] for time to complete data checking (in minutes)

Data-entry experience	Data-checking method			
	Double entry	Solo read aloud	Partner read aloud	Visual checking
No	44.84 [42.63–47.14]	35.50 ^a [33.98–37.05]	30.78 ^{ab} [29.59–32.02]	31.97 ^{ab} [30.35–33.65]
Yes	45.54 [40.32–52.15]	33.77 ^a [30.99–36.94]	28.40 ^{ab} [26.43–30.55]	32.82 ^{ab} [29.80–36.06]

Data-entry experience had no significant effect ($p > .05$) on time to complete data checking. See Table 3 for the significance tests. None of the interactions between data-entry experience and data-checking method were significant for any of the models, all $ps > .05$. ^a As compared to double entry, this data-checking method had a significantly different ($p < .05$) mean, after we controlled for data-entry experience. See Table 3 for the significance tests. ^b As compared to solo read aloud, this data-checking method had a significantly different ($p < .05$) mean, after we controlled for data-entry experience. See Table 3 for the significance tests

Table 3 Hierarchical multiple linear regressions predicting time to complete data checking

Participants	Model	Predictor	<i>b</i> [95% CI]	β [95% CI]	ΔR^2	<i>F</i> -test for ΔR^2
All	1	Exp	− 0.98 [− 3.25 to 1.28]	− .04 [− .14 to .05]	.00	$F(1, 409) = 0.73, p = .395$
	2	Exp	− 0.66 [− 2.53 to 1.21]	− .03 [− .11 to .05]	.33	$F(3, 406) = 67.05, p < .001$
		DE vs. SRA*	− 9.83 [− 12.06 to − 7.59]	− .44 [− .54 to − .34]		
		DE vs. PRA*	− 14.68 [− 16.84 to − 12.52]	− .69 [− .79 to − .59]		
		DE vs. VC*	− 12.73 [− 14.98 to − 10.47]	− .56 [− .66 to − .46]		
SRA, PRA, & VC	1	Exp	− 1.09 [− 2.98 to 0.81]	− .06 [− .17 to .05]	.00	$F(1, 315) = 1.28, p = .260$
	2	Exp	− 1.04 [− 2.86 to 0.79]	− .06 [− .17 to .05]	.08	$F(2, 313) = 13.65, p < .001$
		SRA vs. PRA*	− 4.85 [− 6.69 to − 3.02]	− .33 [− .45 to − .21]		
		SRA vs. VC*	− 2.87 [− 4.79 to − 0.95]	− .19 [− .31 to − .06]		

Data-entry experience had no significant effect ($p > .05$) on time to complete data checking. None of the interactions between data-entry experience and data-checking method were significant for any of the models, all $ps > .05$. Therefore, the interactions were eliminated from the models; Model 2 represents a main effects model only. Exp = data-entry experience, DE = double entry, SRA = solo read aloud, PRA = partner read aloud, VC = visual checking. * $p < .05$ for this predictor

any new errors). Across the four data-checking methods, the participants with data-entry experience were 14.6% more likely to create an error-free dataset (see Table 4). However, this small effect was not statistically significant (see Table 5).

Number of errors

Data-entry experience significantly reduced the number of errors. Across the four data-checking methods, participants with data-entry experience had 41% fewer total errors (see Table 6, first section). This moderate main effect was statistically significant ($p = .007$; see Table 7, first section). In addition, the participants with data-entry experience had 64% fewer errors when checking the spelling test ($p = .002$) and 43% fewer errors while checking the math test ($p = .046$; see Tables 6 and 7, later sections).

The 412 participants had a total of 743 errors at the end of the experiment. These errors can be divided into two types: original errors that participants left in the dataset, and new errors that participants introduced while checking the data. In Table 8, these are referred to as *left original error* and *introduced new error*. Using Table 1 as

a guide, we further subdivided these errors into those that would be easy to find during data checking, such as cells that contained entirely wrong words, and errors that would be hard to find, such as cells in which the digits of a number had been reordered. We found that only 70 of the 743 errors (9.4%) were easy-to-find errors that could have been detected by a systematic examination of histograms or frequency tables during preanalysis data screening. Most of the data-checking errors that participants made (90.6%) would be hard to detect during data screening.

Participants with data-entry experience were 40% less likely to leave hard-to-find errors in the dataset ($p = .039$; see Table 9). There were no significant differences in the number of hard-to-find errors that they introduced into the dataset ($p = .120$) or the number of easy-to-find errors that they left in the dataset ($p = .053$) or introduced into the dataset ($p = .206$).

Subjective evaluations

Data-entry experience was not associated with subjective evaluations of the data-checking methods. Across the four

Table 4 Proportions of participants [95% confidence intervals] with perfect data checking

Data-entry experience	Data-checking method			
	Double entry	Solo read aloud	Partner read aloud	Visual checking
No	.83 [.73–.91]	.44 ^a [.35–.56]	.19 ^{ab} [.12–.28]	.32 ^a [.21–.44]
Yes	.84 [.68–1.00]	.50 ^a [.30–.70]	.29 ^{ab} [.12–.50]	.41 ^a [.22–.59]

Data-entry experience had no significant effect ($p > .05$) on the proportion of participants with perfect data checking. See Table 5 for the significance tests. None of the interactions between data-entry experience and data-checking method were significant for any of the models, $ps > .05$. ^a As compared to double entry, this data-checking method had a significantly different ($p < .05$) proportion, after we controlled for data-entry experience. See Table 5 for the significance tests. ^b As compared to solo read aloud, this data-checking method had a significantly different ($p < .05$) proportion, after we controlled for data-entry experience. See Table 5 for the significance tests

Table 5 Hierarchical generalized linear models predicting perfect data checking

Participants	Model	Predictor	log(<i>b</i>) [95% CI]	Odds Ratio [95% CI]	ΔR^2	χ^2 for ΔR^2
All	1	Exp	0.23 [− 0.24 to 0.70]	1.26 [0.79 to 2.01]	.00	$\chi^2(1) = 0.93, p = .335$
	2	Exp	0.34 [− 0.19 to 0.87]	1.41 [0.83 to 2.38]	.27	$\chi^2(3) = 92.83, p < .001$
		DE vs. SRA*	− 1.77 [− 2.46 to − 1.12]	0.17 [0.09 to 0.33]		
		DE vs. PRA*	− 2.92 [− 3.65 to − 2.25]	0.05 [0.03 to 0.11]		
		DE vs. VC*	− 2.25 [− 2.96 to − 1.59]	0.11 [0.05 to 0.20]		
SRA, PRA, & VC	1	Exp	0.36 [− 0.19 to 0.91]	1.44 [0.83 to 2.48]	.01	$\chi^2(1) = 1.67, p = .196$
	2	Exp	0.38 [− 0.19 to 0.94]	1.46 [0.83 to 2.55]	.07	$\chi^2(2) = 15.37, p < .001$
		SRA vs. PRA*	− 1.15 [− 1.76 to 0.57]	0.32 [0.17 to 0.57]		
		SRA vs. VC	− 0.49 [− 1.07 to 0.09]	0.61 [0.34 to 1.09]		

None of the interactions between data-entry experience and data-checking method were significant for any of the models, all $ps > .05$. Therefore, the interactions were eliminated from the models; Model 2 represents a main effects model only. Exp = data-entry experience, DE = double entry, VC = visual checking, PRA = partner read aloud, SRA = solo read aloud. * $p < .05$ for this predictor

data-checking methods, the participants with data-entry experience provided ratings that were 1.3% higher than the ratings provided by participants without data-checking experience (see Table 10). However, this small effect was not statistically significant (see Table 11).

Summary

This is the first empirical study of the effect of data-entry experience. We found that data-entry experience significantly and substantially reduces the number of errors.

Data-checking methods

Next we compared the four data-checking methods: double entry, solo read aloud, partner read aloud, and visual checking. Because participants were assigned to data-checking methods completely at random, participants with data-entry experience were not evenly distributed between the four data-checking methods. This represents a potential confound, which could cause spurious differences between the methods. To be conservative, we therefore controlled for data-entry experience in all comparisons of the data-checking methods.

Table 6 Mean numbers of errors per participant [95% confidence intervals]

Data type	Data-entry experience	Original errors	Data-checking method			
			Double entry	Solo read aloud	Partner read aloud	Visual checking
Total errors	No	32	0.65 [0.20–1.32]	1.40 ^a [0.96–1.88]	3.54 ^{ab} [2.64–4.54]	1.97 ^a [1.28–2.83]
	Yes*		0.58 [0.00–1.58]	0.90 ^a [0.40–1.50]	2.21 ^{ab} [1.12–3.79]	0.85 ^a [0.52–1.30]
ID 6 digit	No	3	0.08 [0.00–0.19]	0.20 ^a [0.11–0.30]	0.22 ^a [0.14–0.32]	0.34 ^a [0.20–0.51]
	Yes		0.05 [0.00–0.16]	0.30 ^a [0.05–0.65]	0.17 ^a [0.04–0.33]	0.11 ^a [0.00–0.22]
Sex [†]	No	1	0.00 [0.00–0.00]	0.00 [0.00–0.00]	0.03 [0.00–0.07]	0.01 [0.00–0.04]
	Yes		0.00 [0.00–0.00]	0.00 [0.00–0.00]	0.04 [0.00–0.12]	0.00 [0.00–0.00]
Learning style	No	5	0.21 [0.09–0.37]	0.20 [0.09–0.33]	0.24 [0.10–0.44]	0.18 [0.06–0.35]
	Yes		0.53 [0.00–1.53]	0.20 [0.10–0.45]	0.12 [0.00–0.25]	0.11 [0.00–0.22]
Study habits	No	7	0.15 [0.00–0.35]	0.46 ^a [0.23–0.74]	0.93 ^{ab} [0.62–1.27]	0.23 [0.06–0.45]
	Yes		0.00 [0.00–0.00]	0.25 ^a [0.00–0.55]	0.71 ^{ab} [0.29–1.25]	0.15 [0.00–0.41]
Spelling test	No	9	0.08 [0.01–0.19]	0.20 ^a [0.10–0.31]	1.30 ^{ab} [0.87–1.79]	0.69 ^{ab} [0.35–1.11]
	Yes*		0.00 [0.00–0.00]	0.10 ^a [0.00–0.25]	0.58 ^{ab} [0.21–1.08]	0.15 ^{ab} [0.04–0.30]
Math test	No	8	0.13 [0.01–0.33]	0.35 ^a [0.21–0.51]	0.81 ^{ab} [0.53–1.15]	0.52 ^a [0.32–0.75]
	Yes*		0.00 [0.00–0.00]	0.05 ^a [0.00–0.15]	0.58 ^{ab} [0.21–1.12]	0.33 ^a [0.15–0.52]

None of the interactions between data-entry experience and data-checking method were significant for any of the models, $ps > .05$. [†] Due to the low number of errors for this type of data, the regression models did not converge. Thus, the numbers of errors could not be compared across the four data-checking methods. *Data-entry experience had a significant ($p < .05$) influence on the number of errors for this type of data. See Table 7 for the significance tests. ^a As compared to double entry, this data-checking method had a significantly different ($p < .05$) mean on this type of data, after we controlled for data-entry experience. See Table 7 for the significance tests. ^b As compared to solo read aloud, this data-checking method had a significantly different ($p < .05$) mean on this type of data, after we controlled for data-entry experience, $p < .05$. See Table 7 for the significance tests.

Table 7 Hierarchical generalized linear models predicting numbers of errors

Dependent variable	Participants	Model	Predictor	$\log(b)$ [95% CI]	Odds ratio [95% CI]	ΔR^2	χ^2 for ΔR^2
Total errors	All	1	Exp*	–0.53 [–0.91 to –0.14]	0.59 [0.40 to 0.87]	.02	$\chi^2(1) = 6.88, p = .009$
		2	Exp*	–0.51 [–0.87 to –0.13]	0.60 [0.42 to 0.87]	.13	$\chi^2(3) = 55.99, p < .001$
			DE vs. SRA*	0.70 [0.24 to 1.16]	2.01 [1.27 to 3.20]		
			DE vs. PRA*	1.62 [1.19 to 2.06]	5.07 [3.30 to 7.82]		
			DE vs. VC*	0.96 [0.50 to 1.42]	2.61 [1.65 to 4.15]		
			Exp*	–0.59 [–0.96 to –0.21]	0.56 [0.38 to 0.81]	.03	$\chi^2(1) = 9.08, p = .003$
ID errors	SRA, PRA, VC	1	Exp*	–0.58 [–0.94 to –0.21]	0.56 [0.39 to 0.81]	.09	$\chi^2(2) = 29.62, p < .001$
		2	Exp*	0.93 [0.58 to 1.28]	2.53 [1.78 to 3.58]		
			SRA vs. PRA*	0.27 [–0.11 to 0.65]	1.31 [0.89 to 1.91]		
			SRA vs. VC	–0.29 [–0.95 to 0.31]	0.75 [0.39 to 1.37]	.00	$\chi^2(1) = 0.86, p = .355$
		1	Exp	–0.32 [–0.98 to 0.27]	0.72 [0.37 to 1.32]	.04	$\chi^2(3) = 10.80, p = .013$
		2	Exp	1.08 [0.23 to 2.04]	2.94 [1.26 to 7.71]		
Sex errors	SRA, PRA, VC		DE vs. SRA*	1.04 [0.21 to 1.99]	2.82 [1.23 to 7.31]		
			DE vs. PRA*	1.32 [0.50 to 2.27]	3.76 [1.65 to 9.72]		
		1	Exp	–0.30 [–0.97 to 0.30]	0.74 [0.38 to 1.35]	.00	$\chi^2(1) = 0.91, p = .340$
		2	Exp	–0.32 [–0.99 to 0.28]	0.73 [0.37 to 1.32]	.00	$\chi^2(2) = 1.09, p = .579$
			SRA vs. PRA	–0.04 [–0.64 to 0.57]	0.96 [0.53 to 1.77]		
			SRA vs. VC	0.25 [–0.35 to 0.86]	1.28 [0.71 to 2.36]		
Learning style errors	All	1	Exp	Model did not converge.			
		2	Exp	Model did not converge.			
			DE vs. SRA				
			DE vs. PRA				
			DE vs. VC				
			Model did not converge				
Learning style errors	SRA, PRA, VC	1	Exp	0.05 [–0.72 to 0.84]	1.05 [0.49 to 2.33]	.00	$\chi^2(1) = 0.02, p = .898$
		2	Exp	0.02 [–0.76 to 0.82]	1.02 [0.47 to 2.26]	.00	$\chi^2(3) = 1.24, p = .744$
			DE vs. SRA	–0.33 [–1.26 to 0.59]	0.72 [0.28 to 1.81]		
			DE vs. PRA	–0.23 [–1.13 to 0.65]	0.79 [0.32 to 1.91]		
			DE vs. VC	–0.53 [–1.49 to 0.42]	0.59 [0.23 to 1.53]		
		1	Exp	–0.40 [–1.33 to 0.50]	0.67 [0.26 to 1.65]	.00	$\chi^2(1) = 0.76, p = .384$
Learning style errors	SRA, PRA, VC	2	Exp	–0.38 [–1.31 to 0.52]	0.68 [0.27 to 1.69]	.00	$\chi^2(2) = 0.35, p = .840$
			SRA vs. PRA	0.08 [–0.76 to 0.93]	1.09 [0.47 to 2.54]		
			SRA vs. VC	–0.18 [–1.11 to 0.74]	0.83 [0.33 to 2.09]		

Table 7 (continued)

Dependent variable	Participants	Model	Predictor	$\log(b)$ [95% CI]	Odds ratio [95% CI]	ΔR^2	χ^2 for ΔR^2
Study habits errors	All	1	Exp	−0.48 [−1.17 to 0.22]	0.62 [0.31 to 1.25]	.01	$\chi^2(1) = 1.82, p = .177$
		2	Exp	−0.53 [−1.20 to 0.14]	0.59 [0.30 to 1.16]	.09	$\chi^2(3) = 29.83, p = .000$
			DE vs. SRA*	1.28 [0.43 to 2.19]	3.61 [1.54 to 8.92]		
			DE vs. PRA*	2.05 [1.24 to 2.91]	7.74 [3.45 to 18.37]		
			DE vs. VC	0.61 [−0.31 to 1.57]	1.85 [0.73 to 4.82]		
SRA, PRA, VC	SRA, PRA, VC	1	Exp	−0.44 [−1.11 to 0.25]	0.65 [0.33 to 1.29]	.01	$\chi^2(1) = 1.56, p = .212$
		2	Exp	−0.39 [−1.05 to 0.28]	0.68 [0.35 to 1.32]	.06	$\chi^2(2) = 17.72, p < .001$
			SRA vs. PRA*	0.75 [0.15 to 1.36]	2.13 [1.17 to 3.89]		
			SRA vs. VC	−0.68 [−1.43 to 0.05]	0.51 [0.24 to 1.05]		
			Exp*	−1.01 [−1.66 to −0.38]	0.36 [0.19 to 0.68]	.03	$\chi^2(1) = 9.92, p = .002$
Spelling test errors	All	2	Exp*	−1.06 [−1.69 to −0.47]	0.35 [0.19 to 0.63]	.19	$\chi^2(3) = 74.67, p < .001$
			DE vs. SRA*	1.04 [0.08 to 2.12]	2.82 [1.08 to 8.34]		
			DE vs. PRA*	2.94 [2.10 to 3.94]	18.86 [8.20 to 51.42]		
			DE vs. VC*	2.18 [1.31 to 3.21]	8.84 [3.70 to 24.74]		
			Exp*	−1.01 [−1.65 to −0.39]	0.36 [0.19 to 0.68]	.03	$\chi^2(1) = 10.11, p = .001$
SRA, PRA, VC	SRA, PRA, VC	2	Exp*	−1.02 [−1.64 to −0.42]	0.36 [0.19 to 0.66]	.13	$\chi^2(2) = 39.51, p < .001$
			SRA vs. PRA*	1.90 [1.30 to 2.53]	6.66 [3.67 to 12.58]		
			SRA vs. VC*	1.14 [0.49 to 1.82]	3.13 [1.64 to 6.16]		
			Exp*	−0.56 [−1.14 to −0.01]	0.57 [0.32 to 0.99]	.01	$\chi^2(1) = 3.98, p = .046$
			Exp*	−0.65 [−1.21 to −0.11]	0.52 [0.30 to 0.89]	.10	$\chi^2(3) = 36.86, p < .001$
Math test errors	All		DE vs. SRA*	0.99 [0.20 to 1.83]	2.68 [1.23 to 6.26]		
			DE vs. PRA*	1.99 [1.28 to 2.78]	7.28 [3.59 to 16.14]		
			DE vs. VC*	1.54 [0.79 to 2.37]	4.68 [2.21 to 10.69]		
		1	Exp	−0.52 [−1.08 to 0.01]	0.59 [0.34 to 1.01]	.01	$\chi^2(1) = 3.64, p = .057$
		2	Exp*	−0.56 [−1.11 to −0.03]	0.57 [0.33 to 0.97]	.05	$\chi^2(2) = 14.31, p = .001$
SRA, PRA, VC	SRA, PRA, VC		SRA vs. PRA*	0.99 [0.48 to 1.53]	2.69 [1.61 to 4.60]		
			SRA vs. VC	0.54 [−0.02 to 1.12]	1.72 [0.98 to 3.08]		

None of the interactions between data-entry experience and data-checking method were significant for any of the models, all $ps > .05$. Therefore, the interactions were eliminated from the models; Model 2 represents a main effects model only. Exp = data-entry experience, DE = double entry, VC = visual checking, PRA = partner read aloud, SRA = solo read aloud. * $p < .05$ for this predictor

Table 8 Mean numbers of easy-to-find errors and hard-to-find errors per participant [95% confidence intervals]

Error type	Data-entry experience	Data-checking method			
		Double entry	Solo read aloud	Partner read aloud	Visual checking
Left original error					
Easy-to-find errors	No	0.01 [0.00–0.04]	0.01 [0.00–0.04]	0.28 ^{ab} [0.09–0.52]	0.31 ^{ab} [0.06–0.63]
	Yes	0.00 [0.00–0.00]	0.00 [0.00–0.00]	0.04 ^{ab} [0.00–0.13]	0.04 ^{ab} [0.00–0.11]
Hard-to-find errors	No	0.42 [0.03–1.04]	0.57 [0.38–0.79]	1.20 ^{ab} [0.82–1.63]	1.12 ^{ab} [0.70–1.63]
	Yes*	0.00 [0.00–0.00]	0.40 [0.10–0.75]	1.05 ^{ab} [0.42–1.79]	0.44 ^{ab} [0.26–0.63]
Introduced new error					
Easy-to-find errors [†]	No	0.00 [0.00–0.00]	0.01 [0.00–0.04]	0.13 ^b [0.04–0.23]	0.03 [0.00–0.07]
	Yes	0.00 [0.00–0.00]	0.00 [0.00–0.00]	0.04 ^b [0.00–0.13]	0.00 [0.00–0.00]
Hard-to-find errors	No	0.23 [0.08–0.41]	0.80 ^a [0.46–1.22]	1.93 ^{ab} [1.39–2.56]	0.52 [0.28–0.83]
	Yes*	0.58 [0.00–1.58]	0.50 ^a [0.15–1.00]	1.09 ^{ab} [0.50–1.83]	0.37 [0.11–0.70]

Originally, there were 32 errors in the Excel sheet: 13 easy-to-find and 19 hard-to-find errors. None of the interactions between data-entry experience and data-checking method were significant for any of the models, $p > .05$. [†] Due to the low number of easy-to-find errors, the regression model that compared double entry to other methods did not converge. Thus, the number of easy-to-find errors for double entry could not be compared. *Data-entry experience had a significant main effect ($p < .05$) for this type of error. See Table 9 for the significance tests

^a As compared to double entry, this data-checking method had a significantly different ($p < .05$) mean on this type of error, after we controlled for data-entry experience. See Table 9 for the significance tests. ^b As compared to solo read aloud, this data-checking method had a significantly different ($p < .05$) mean on this type of error, after we controlled for data-entry experience, $p < .05$. See Table 9 for the significance tests

Time

Double entry was the slowest method, as expected. Overall, double entry took 45 min, solo read aloud took 35 min, visual checking took 32 min, and partner read aloud took 30 min (see Table 2). After we controlled for data-entry experience, double entry was 28% slower than visual checking, 39% slower than partner read aloud, and 48% slower than solo read aloud, all $ps < .001$ (see Table 3, Model 2 for all participants).

Solo read aloud was the second slowest method. After we controlled for data-entry experience, solo read aloud was 16% slower than visual checking and 16% slower than partner read aloud, all $ps < .01$ (see Table 3, Model 2 for comparisons of solo read-aloud, partner read-aloud, and visual-checking participants).

Perfect data checking

Originally, there were 32 errors in the Excel file. Double-entry participants were the most likely to correct all of these errors without introducing any new errors. Overall, the probabilities of creating an error-free dataset were .83 for double entry, .46 for solo read aloud, .21 for partner read aloud, and .35 for visual checking (see Table 4 and Fig. 3). After we controlled for data-entry experience, double entry was 83% more likely than solo read aloud to result in perfect checking, 95% more likely than partner read aloud, and 89% more likely than visual checking, all $ps < .001$ (see Table 5).

Solo read aloud was the next most likely to result in perfect data. After we controlled for data-entry experience, solo read aloud was 68% more likely than partner read aloud to result in perfect data checking ($p < .001$), although it was not significantly more likely than visual checking ($p = .098$; see Table 5).

Number of errors

Double-entry participants had the fewest total errors. These participants averaged 0.64 errors, as compared to 1.30 for solo read aloud, 3.27 for partner read aloud, and 1.66 for visual checking (see Table 6 and Fig. 4). After we controlled for data-entry experience, solo read aloud had 2.01 times as many total errors as double entry, partner read aloud had 5.07 times as many, and visual checking had 2.61 times as many, all $ps < .01$ (see Table 7, total errors).

Solo read aloud had the next fewest errors. After we controlled for data-entry experience, partner read aloud resulted in 2.53 times as many errors as solo read aloud ($p < .001$). Visual checking resulted in 1.31 times as many errors as solo read aloud, but this ratio was not statistically significant, $p = .168$ (see Table 7). Thus, solo read aloud was not as accurate as double entry, but it was more accurate than partner read aloud.

To determine whether the number of errors depended on the type of data being checked, we fit separate models for each of the six types of data on the data sheets. Most of these models converged without incident. For those models, the differences between data-checking methods largely reflected the patterns seen above for total number of errors (see Tables 6

Table 9 Hierarchical generalized linear models predicting the numbers of easy-to-find errors and hard-to-find errors

Type of error	Participants	Model	Predictor	log(<i>b</i>) [95% CI]	Odds Ratio [95% CI]	Δ <i>R</i> ²	χ ² for Δ <i>R</i> ²
Left original error							
Easy-to-find errors	All	1	Exp	− 1.96 [− 4.14 to 0.07]	0.14 [0.02 to 1.07]	.02	χ ² (1) = 3.59, <i>p</i> = .058
		2	Exp*	− 2.08 [− 4.26 to − 0.13]	0.12 [0.01 to 0.88]	.09	χ ² (3) = 15.00, <i>p</i> = .002
			DE vs. SRA	− 0.08 [− 3.54 to 3.38]	0.93 [0.03 to 29.51]		
			DE vs. PRA*	3.12 [0.98 to 6.21]	22.58 [2.67 to 498.90]		
			DE vs. VC*	3.18 [0.99 to 6.31]	23.98 [2.70 to 547.96]		
	SRA, PRA, VC	1	Exp	− 1.97 [− 4.17 to 0.11]	0.14 [0.02 to 1.11]	.02	χ ² (1) = 3.48, <i>p</i> = .062
		2	Exp*	− 2.05 [− 4.24 to − 0.04]	0.13 [0.01 to 0.96]	.06	χ ² (2) = 9.17, <i>p</i> <.010
			SRA vs. PRA*	3.19 [1.08 to 6.28]	24.30 [2.93 to 533.99]		
			SRA vs. VC*	3.25 [1.09 to 6.37]	25.77 [2.96 to 585.94]		
Hard-to-find errors	All	1	Exp*	− 0.51 [− 1.00 to − 0.02]	0.60 [0.37 to 0.98]	.01	χ ² (1) = 4.21, <i>p</i> = .040
		2	Exp*	− 0.61 [− 1.09 to − 0.13]	0.54 [0.34 to 0.87]	.07	χ ² (3) = 25.20, <i>p</i> <.001
			DE vs. SRA	0.52 [− 0.08 to 1.13]	1.68 [0.92 to 3.10]		
			DE vs. PRA*	1.32 [0.76 to 1.88]	3.72 [2.14 to 6.57]		
			DE vs. VC*	1.09 [0.51 to 1.68]	2.96 [1.66 to 5.36]		
	SRA, PRA, VC	1	Exp	− 0.42 [− 0.88 to 0.04]	0.66 [0.42 to 1.04]	.01	χ ² (1) = 3.23, <i>p</i> = .072
		2	Exp	− 0.44 [− 0.89 to 0.01]	0.64 [0.41 to 1.01]	.04	χ ² (2) = 11.97, <i>p</i> = .003
			SRA vs. PRA*	0.79 [0.34 to 1.23]	2.19 [1.41 to 3.43]		
			SRA vs. VC*	0.56 [0.09 to 1.04]	1.76 [1.10 to 2.82]		
Introduced new error							
Easy-to-find errors	All	1	Exp	− 1.43 [− 4.43 to 0.48]	0.24 [0.01 to 1.61]	.02	χ ² (1) = 2.03, <i>p</i> = .154
		2	Exp	Model did not converge.			
			DE vs. SRA				
			DE vs. PRA				
			DE vs. VC				
	SRA, PRA, VC	1	Exp	− 1.46 [− 4.45 to 0.44]	0.23 [0.01 to 1.55]	.02	χ ² (1) = 2.15, <i>p</i> = .142
		2	Exp	− 1.44 [− 4.43 to 0.44]	0.24 [0.01 to 1.56]	.09	χ ² (2) = 9.04, <i>p</i> = .011
			SRA vs. PRA*	2.42 [0.63 to 5.38]	11.22 [1.87 to 217.05]		
			SRA vs. VC	0.79 [− 1.67 to 3.93]	2.20 [0.19 to 50.77]		
Hard-to-find errors	All	1	Exp	− 0.40 [− 0.89 to 0.11]	0.67 [0.41 to 1.12]	.01	χ ² (1) = 2.35, <i>p</i> = .125
		2	Exp	− 0.20 [− 0.68 to 0.28]	0.82 [0.51 to 1.33]	.11	χ ² (3) = 44.84, <i>p</i> <.001
			DE vs. SRA*	0.89 [0.29 to 1.51]	2.44 [1.34 to 4.51]		
			DE vs. PRA*	1.75 [1.19 to 2.33]	5.77 [3.29 to 10.29]		
			DE vs. VC	0.47 [− 0.16 to 1.11]	1.60 [0.85 to 3.03]		
	SRA, PRA, VC	1	Exp*	− 0.58 [− 0.07 to 0.38]	0.56 [0.34 to 0.94]	.02	χ ² (1) = 4.75, <i>p</i> = .029
		2	Exp	− 0.48 [− 0.98 to 0.02]	0.62 [0.38 to 1.02]	.09	χ ² (2) = 28.39, <i>p</i> <.001
			SRA vs. PRA*	0.86 [0.41 to 1.32]	2.37 [1.51 to 3.73]		
			SRA vs. VC	− 0.40 [− 0.93 to 0.13]	0.67 [0.39 to 1.14]		

None of the interactions between data-entry experience and data-checking method were significant for any of the models, all $ps > .05$. Therefore, the interactions were eliminated from the models; Model 2 represents a main effects model only. Exp = data-entry experience, DE = double entry, VC = visual checking, PRA = partner read aloud, SRA = solo read aloud. * $p < .05$ for this predictor

and 7). Double entry had significantly fewer errors than most of the other data-checking methods for four of the types of data. Solo read aloud resulted in significantly fewer errors than partner read aloud for three of the types of data and resulted in significantly fewer errors than visual checking for one type of data. The regression model predicting errors when checking the sex data did not converge, because participants rarely

made errors when checking this variable (five of the 412 participants made one error; the rest made zero errors); however, we noted descriptively that double-entry and solo read-aloud participants made no errors when checking the sex data, whereas partner read aloud and visual checking each had participants who made errors. All together, these results reinforce our previous conclusion that solo read aloud was less accurate

Table 10 Means [95% confidence intervals] for subjective evaluations

Rating	Experience	Double entry	Solo read aloud	Partner read aloud	Visual checking
Overall evaluation	No	3.25 [3.12–3.38]	3.03 ^a [2.90–3.16]	3.12 [2.99–3.24]	3.23 ^b [3.09–3.37]
	Yes	3.31 [3.10–3.50]	2.88 ^a [2.58–3.18]	3.21 [2.94–3.48]	3.38 ^b [3.11–3.66]
Accuracy	No	3.60 [3.37–3.81]	3.52 [3.31–3.73]	3.31 [3.11–3.51]	3.45 [3.23–3.68]
	Yes	3.37 [3.05–3.84]	3.20 [3.00–3.80]	3.12 [3.25–3.92]	3.22 [3.07–3.81]
Reliability	No	3.57 [3.36–3.77]	3.53 [3.36–3.70]	3.29 [3.07–3.49]	3.57 [3.36–3.77]
	Yes	3.37 [2.95–3.79]	3.20 [2.70–3.65]	3.12 [2.75–3.50]	3.22 [2.89–3.59]

1 = *strongly disagree* and 5 = *strongly agree*. Data-entry experience had no significant effect ($p > .05$) on the subjective evaluations of the data-checking methods. See Table 11 for the significance tests. None of the interactions between data-entry experience and data-checking method were significant for any of the models, $ps > .05$. ^a As compared to double entry, this data-checking method had a significantly different ($p < .05$) mean on this variable, after we controlled for data-entry experience. See Table 11 for the significance tests. ^b As compared to solo read aloud, this data-checking method had a significantly different ($p < .05$) mean on this variable, after we controlled for data-entry experience, $p < .05$. See Table 11 for the significance tests

than double entry, but more accurate than partner read aloud. In addition, solo read aloud was sometimes more accurate than visual checking.

Excluding extreme cases To ensure that these results were not driven by a few extreme scores, we repeated these analyses after excluding the worst 1% of participants on

Table 11 Hierarchical multiple linear regressions for subjective evaluations

Dependent variable	Participants	Model	Predictor	<i>b</i> [95% CI]	β [95% CI]	ΔR^2	<i>F</i> test for ΔR^2
Overall evaluation	All	1	Exp	.06 [–.09 to .21]	.04 [–.06 to .14]	.00	$F(1, 410) = 0.62, p = .430$
		2	Exp	.05 [–.10 to .19]	.03 [–.06 to .13]	.03	$F(3, 407) = 4.28, p = .005$
			DE vs. SRA*	–.27 [–.44 to –.09]	–.18 [–.30 to –.06]		
			DE vs. PRA	–.13 [–.30 to .04]	–.10 [–.22 to .03]		
			DE vs. VC	.01 [–.17 to .18]	.00 [–.12 to .12]		
	SRA, PRA, VC	1	Exp	.06 [–.11 to .23]	.04 [–.07 to .15]	.00	$F(1, 316) = 0.52, p = .470$
		2	Exp	.04 [–.13 to .21]	.03 [–.08 to .14]	.03	$F(2, 314) = 4.52, p = .012$
			SRA vs. PRA	.13 [–.04 to .30]	.10 [–.03 to .23]		
			SRA vs. VC*	.27 [.09 to .45]	.20 [.07 to .32]		
Accuracy	All	1	Exp	.01 [–.22 to .23]	.00 [–.09 to .10]	.00	$F(1, 410) = 0.00, p = .952$
		2	Exp	.01 [–.22 to .24]	.00 [–.09 to .10]	.01	$F(3, 407) = 0.82, p = .486$
			De vs. SRA	–.07 [–.34 to .20]	–.03 [–.15 to .09]		
			DE vs. PRA	–.20 [–.47 to .06]	–.09 [–.22 to .03]		
			DE vs. VC	–.12 [–.39 to .16]	–.05 [–.17 to .07]		
	SRA, PRA, VC	1	Exp	.06 [–.20 to .32]	.03 [–.08 to .14]	.00	$F(1, 316) = 0.22, p = .636$
		2	Exp	.06 [–.20 to .32]	.03 [–.08 to .14]	.00	$F(2, 314) = 0.54, p = .586$
			SRA vs. PRA	–.13 [–.39 to .12]	–.07 [–.20 to .06]		
			SRA vs. VC	–.05 [–.32 to .22]	–.02 [–.15 to .11]		
Reliability	All	1	Exp	–.21 [–.44 to .01]	–.09 [–.19 to .00]	.01	$F(1, 410) = 3.52, p = .062$
		2	Exp	–.21 [–.43 to .01]	–.09 [–.19 to .01]	.01	$F(3, 407) = 1.80, p = .146$
			DE vs. SRA	–.07 [–.33 to .20]	–.03 [–.15 to .09]		
			DE vs. PRA	–.28 [–.54 to –.02]	–.13 [–.26 to –.01]		
			DE vs. VC	–.18 [–.45 to .09]	–.08 [–.20 to .04]		
	SRA, PRA, VC	1	Exp	–.21 [–.46 to .04]	–.09 [–.20 to .02]	.01	$F(1, 316) = 2.65, p = .105$
		2	Exp	–.21 [–.46 to .05]	–.09 [–.20 to .02]	.01	$F(2, 314) = 1.35, p = .261$
			SRA vs. PRA	–.21 [–.47 to .04]	–.11 [–.24 to .02]		
			SRA vs. VC	–.11 [–.38 to .16]	–.05 [–.18 to .07]		

None of the interactions between data-entry experience and data-checking method were significant for any of the models, all $ps > .05$. Therefore, the interactions were eliminated from the models; Model 2 represents a main effects model only. Exp = data-entry experience, DE = double entry, SRA = solo read aloud, PRA = partner read aloud, VC = visual checking. * $p < .05$ for this predictor

Fig. 3 Proportions of participants with perfect data checking (with the standard error for each proportion). DE = double entry. VC = visual checking. PRA = partner read aloud. SRA = solo read aloud

total number of errors. These five participants (one from double entry, one from visual checking, and three from partner read aloud, all of whom had no previous data-entry experience) each had 20 or more errors. See the supplementary materials, Tables A and B. For total number of errors and errors on checking the math test, prior data-entry experience was no longer significant, but the effect was still in the same direction. Additionally, solo read-aloud participants now made significantly fewer errors than visual-checking participants for the study habits test, but the direction of the difference was the same as before. For the remaining analyses, removing these five participants had no effect on either the direction or the significance of coefficients. We concluded that the five error-prone participants did not have a substantial or

meaningful effect on the above comparisons of the data-checking methods.

Types of errors Double entry and solo read aloud both substantially reduced the number of easy-to-find errors, as compared to partner read aloud and visual checking (see Table 8). After we controlled for previous data-entry experience (see Table 9), partner read aloud and visual checking resulted in 22.58 and 23.98 times as many easy-to-find errors as double entry, both $ps < .012$, and 24.30 and 25.77 times as many as solo read aloud, both $ps < .010$. Solo read aloud was not significantly different from double entry in the number of easy-to-find errors that were left in the dataset, $p = .962$. In addition, partner read aloud introduced 11.22 times as many easy-to-find errors into the dataset as solo read aloud, $p = .028$. The

Fig. 4 Mean numbers of errors (with the standard error for each mean). DE = double entry. VC = visual checking. PRA = partner read aloud. SRA = solo read aloud

model comparing double entry to the other methods did not converge, but we noted that double entry resulted in even fewer easy-to-find errors being introduced into the dataset than solo read aloud.

Double entry and solo read aloud both also reduced the number of hard-to-find errors, as compared to partner read aloud and visual checking. After controlling for previous data-entry experience (see Table 9), partner read-aloud and visual-checking participants left 3.72 and 2.96 times as many hard-to-find errors in the dataset as double-entry participants, both $ps < .001$, and 2.19 and 1.76 times as many as solo read-aloud participants, both $ps < .019$. In addition, partner read-aloud participants introduced 5.77 and 2.37 times as many hard-to-find errors as double-entry and solo read-aloud participants, both $ps < .004$. However, solo read aloud was not as good as double entry. Solo read-aloud participants introduced 2.44 times as many hard-to-find errors into the dataset as double-entry participants, $p < .004$. Thus, double entry and solo read aloud are both better than partner read aloud and visual checking, but solo read aloud is not as good as double entry at avoiding hard-to-find errors. Double entry remains the gold standard in terms of reducing the number of errors and shows the greatest advantages in eliminating hard-to-find errors.

Double-entry errors Although double-entry participants had fewer than half as many errors as those using the next best method, they still made some errors: 60 errors, to be exact. We examined these errors carefully to glean insights for improving double-entry systems. We found that double-entry participants always made the two computer entries match each other, but sometimes did not make the entries match the original paper data sheet. This happened in two circumstances. Most often (32 of the 60 errors), double-entry participants failed to find an existing error. When the two entries disagreed, they changed their entry to match the original (incorrect) entry. Sometimes (28 of the errors), double-entry participants entered something incorrectly themselves and then changed the original entry to match. Thus, they introduced an error that had not existed before.

As we described above, previous data-entry experience reduced the number of errors. Consistent with that overall finding, double-entry participants with previous data-entry experience never made their entries match incorrect original entries. However, they did sometimes change correct entries to match their incorrect entries. See Table 8.

Subjective opinions

For all four data-checking methods, subjective evaluations were near the midpoint of the five-point scale. However, participants liked solo read aloud the least (see Table 10). After we controlled for previous data-entry experience (see

Table 11), participants reported significantly lower subjective evaluations of solo read aloud than double entry ($p = .003$) and visual checking ($p = .003$) and slightly (but not significantly) lower evaluations than partner read aloud ($p = .121$). In contrast, ratings of double entry were not significantly different from the ratings of either partner read aloud ($p = .124$) or visual checking ($p = .947$).

Comparing actual accuracy with perceived accuracy and reliability

Participants were poor judges of the quality of their data-checking efforts. First, a participant's subjective opinions on accuracy and reliability were not significantly related to their actual accuracy [accuracy judgment: $r(410) = .03$, 95% CI $[-.07$ to $.13]$, $p = .538$; reliability judgment: $r(410) = .08$, 95% CI $[-.01$ to $.18]$, $p = .091$]. Second, participants with data-entry experience did not differ from participants with no data-entry experience in terms of their accuracy and reliability judgments, all $ps > .06$ (see Table 11), even though their actual accuracy was substantially and significantly higher (see Tables 6 and 7). Third, participants did not give higher ratings of accuracy and reliability to those data-checking methods that had the highest actual accuracy, both $ps > .05$ (see Table 11). In contrast, there were four statistically significant differences between the data-checking methods in terms of actual accuracy (see Table 7, total errors). Thus, subjective opinions of accuracy and reliability are poor substitutes for measuring actual accuracy.

Discussion

Every data-checking method corrected the vast majority of errors. Even the worst data-checking method (partner read aloud completed by people with no previous data-entry experience) left only 3.54 errors in the Excel file, on average. Because the Excel file only contained 5% errors to start, 99.48% of the entries were correct when the data checking was complete. The best data-checking method (double entry completed by people with previous data-entry experience) resulted in 99.91% accuracy. In absolute terms, the difference between 99.48% accuracy and 99.91% accuracy is small, which may explain why most researchers think that their particular data-checking method is excellent. This may also explain why there was no correlation between the number of errors and perceived accuracy. With all methods having accuracy rates greater than 99%, it is difficult for researchers to discern the differences between data-checking methods without doing systematic research like the present study. The reader is reminded, however, that an accuracy rate of 99.48% is far from optimal. Such errors can reverse the sign of a correlation coefficient or make a significant t test nonsignificant

(Barchard & Verenikina, 2013). Moreover, the vast majority of data-entry errors in this study and in Barchard and Verenikina's were within the range for the variables, and thus could not be detected using holistic methods such as frequency tables and histograms. Therefore, it is essential that researchers use high-quality item-by-item data-checking methods.

Previous research has shown that double entry results in significantly fewer errors than single entry (Barchard & Pace, 2011), visual checking (Barchard & Pace, 2011; Barchard & Verenikina, 2013), and partner read aloud (Barchard & Verenikina, 2013; Kawado et al., 2003). This study replicated those findings by showing that double entry had the fewest errors of the four data-checking methods we examined. It was more likely than other methods to result in perfect data (more than five times as likely as the next best method) and had fewer errors (less than half as many as the next best method). As compared to the other methods, it was particularly good at finding the hard-to-find errors: those errors that would be impossible to detect using holistic methods such as histograms and frequency tables.

Double entry might result in the lowest error rates because it does not rely upon maintaining attention. If people have a small lapse in concentration while entering the data, they might stop typing or type the wrong thing. If they stop typing, then the Excel sheet will show them exactly where they need to start typing again. If they type the wrong thing, then the Excel sheet will highlight that error, making it easy to correct. Similarly, if they have a short lapse of attention while they are looking for mismatches and out-of-range values, and therefore overlook a highlighted cell, the highlighting remains in place so they can easily spot the problem later. Because lapses of attention seem less likely to result in data-checking errors for double entry, double entry may maintain its low error rate if someone checks data for several hours in a row (as might happen if this were a paid professional or a student working toward a thesis deadline), even though the checker might become increasingly tired.

In contrast, participants using visual checking and partner read aloud sometimes overlooked what should have been obvious errors, such as cells that contained entirely wrong words. This might have occurred because both methods are sensitive to lapses of attention. When using visual checking, data checkers have to look back and forth between the paper data sheet and the computer monitor, and thus constantly have to find the place where they left off. If they have a short lapse of attention, they might accidentally skip a few items, and nothing in the data-entry system would alert them to that mistake. When using partner read aloud, data checkers have to maintain strict attention on the computer screen. If they have a short lapse of attention or if they get behind in reading the screen and mentally comparing it to the data being read out

loud, they might do only a cursory comparison of some items and thus overlook some errors. Fundamentally, both methods allow errors to go undetected because short lapses of attention can allow data checkers to skip some items.

In addition to replicating the superiority of double entry over visual checking and partner read aloud, this study also examined a data-checking method that has never been examined empirically: solo read aloud. Solo read aloud was more accurate than partner read aloud. It was more than twice as likely to result in a perfect dataset and had fewer than half as many errors. Solo read aloud might be more accurate than partner read aloud because the data checker is able to control the speed at which the data are read. The person reading the data out loud reads at an even pace, which might sometimes leave the data checker rushing to catch up. This can result in some entries being checked hastily and possibly inaccurately.

Solo read aloud was significantly more accurate than visual checking for one of the six types of data (the spelling test). It might have been more accurate because solo read aloud requires participants to compare a sound to a visual stimulus, whereas visual checking requires participants to compare two visual stimuli. Previous research has shown that cross-modal comparisons are more accurate than within-modal comparisons (Ueda & Saiki, 2012). On the other hand, solo read aloud was only sometimes more accurate than visual checking. Indeed, solo read aloud was slightly (though not significantly) worse than visual checking for the five-point numeric and alphanumeric scales (the learning style and study habits scales), two types of data that are very common in psychology. Therefore, future research should determine under which circumstances each of these methods works best.

Solo read aloud was not as accurate as double entry. Solo read aloud resulted in twice as many errors and was only about one-sixth as likely to result in a perfect dataset. Therefore, double entry remains the gold standard data-checking method. However, double entry is not always possible. For example, when an instructor is entering course grades into the university's student information system or when a researcher is using a colleague's data-entry system for a joint project, double entry might not be an option. In these situations, we recommend solo read aloud or visual checking.

In our study, participants liked solo read aloud the least. This might be due to the fact that participants were checking data while sitting next to the study administrator: Participants might have felt self-conscious about talking to themselves. This discomfort is likely to be much reduced or eliminated entirely if participants are able to use solo read aloud in a private location.

Regardless of which data-checking method was used, previous experience reduced error rates. Across the four data-checking methods, the average number of errors went from 1.89 for participants with no previous data-entry experience to

1.14 for people with experience, a 40% decrease. This suggests that junior researchers should be given experience with data entry. This experience could occur during undergraduate courses (e.g., research methods, statistics, labs, and honors theses). Recall, however, that many of our participants had a lot of data-entry experience: 61 of them had over 100 h. Thus, data entry completed during coursework might not be sufficient. Junior researchers are likely to start data entry for potentially publishable research (e.g., theses and dissertations) while their error rates are still relatively high. Moreover, undergraduate research assistants with little to no previous data-entry experience are often asked to enter data for publications. This raises the question of how junior researchers and research assistants can get data-entry experience, without erroneous results being published. We recommend they use double-entry systems. This doubles the amount of data-entry experience that research assistants can obtain and is also the best method of preventing data-entry errors from ruining study results.

Although double-entry participants had fewer than half as many errors as those using the next best method, and although data-entry experience improves data-checking accuracy, experienced double-entry participants still made some errors. Therefore, there is still room for improvement. By examining the errors that double-entry participants made, this study provides two clues that might help us design better double-entry systems. First, double-entry participants always made the two computer entries match each other, but sometimes they did not make the entries match the original paper data sheet. This suggests they were more focused on making the two entries match each other than on making them match the paper data sheet. Second, double-entry participants with previous data-entry experience never left original errors in the Excel file, but they sometimes changed correct entries to match their incorrect entries. This suggests they were biased to assume that their entries were correct.

Double-entry participants may have introduced errors into the Excel sheet because they preferred their own entries to the original entries; if the entries mismatched, they assumed theirs was correct. This preference could have come about in three ways. First, participants likely noticed that the original entries had more errors than they had. When we entered the data into the Excel sheet originally, we introduced 32 errors, which is a 5% error rate. Participants entering psychological data typically make about 1% errors (Barchard & Verenikina, 2013). Thus, when the present participants identified discrepancies, they probably found, over and over again, that the error was in the original entries. They might have therefore started assuming that the first entry was wrong. However, this would probably not occur in real data entry: Likely, the error rates of the first and second enterers would be more comparable. Second, our double-entry participants might have preferred their own entries because they were explicitly told that they were to

check the original entries (they were not told that they should check their own entries). Therefore, when mismatches occurred, they might have thought that their job was to fix the original entries. Finally, people might naturally prefer their own work to someone else's. Therefore, we might be able to design a better double-entry system by ensuring that data checkers do not have stronger affiliations with one of the entries than the other, because of either the instructions they are given or a natural affinity for their own work.

In this study, double entry involved one person (the researcher) who entered the data and a second person (the participant) who entered the data a second time, compared the entries, and fixed the errors by referring to the original paper data sheet. There are three ways this system could be modified to prevent the data checker from having a greater affinity for one of the sets of entries. System 1 would be for one person to enter the data twice, compare the two sets of entries using the computer, and fix errors by referring to the original paper data sheet. System 2 would be for one person to enter the data twice and a second person to compare the entries and fix errors. System 3 would be for two separate people to enter the data and a third person to compare the entries and fix errors.

We believe System 3 would be more accurate than System 1 or 2, because two different people would be entering the data: If one of them made a data-entry error, the other person would be unlikely to make the same error. Only one study has compared double-entry systems (Kawado et al., 2003), and it did find that System 3 resulted in fewer errors than System 2. However, that study used only two data enterers, making it difficult to generalize the results to other data enterers. Moreover, both of their data enterers had previous experience, making it difficult to generalize to the types of data enterers typically used in psychology, and they were entering medical data, which are quite different from the types of data usually used in psychology. Therefore, future research should determine which double-entry system is the most accurate for the data and data enterers typically used in psychological studies. At this point, we recommend any of these double-entry systems, but particularly System 3. Double-entry modules are available through commercial statistical programs such as SPSS and SAS, and several free programs are also available, including web-based systems (Harris et al., 2009), standalone programs (Gao et al., 2008; Lauritsen, 2000–2018), and add-ons for Excel (Barchard, Bedoy, Verenikina, & Pace, 2016). All of these programs can implement double-entry System 3.

Final words

Psychology is gravely concerned about avoiding errors due to random sampling of participants, violation of statistical assumptions, and response biases. These are all important concerns. However, psychologists need to be equally concerned

that their analyses are based on the actual data they collected. Therefore, we should use double-entry systems when possible, particularly ones in which two different people enter the data and a third person compares them. When double entry is not possible, we should use solo read aloud or visual checking. Finally, we should increase the data-entry experience of junior researchers, by including double entry in undergraduate and graduate courses and by asking research assistants to double enter our own research data.

Author note Some parts of this article were presented at the 2017 conventions of the Western Psychological Association and the Association for Psychological Science.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Barchard, K. A., Bedoy, E. H., Verenikina, Y., & Pace, L. A. (2016). Poka-Yoke Double Entry System Version 3.0.76 (Excel 2013 file that allows double entry, checking for mismatches, and checking for out of range values). Available at <http://faculty.unlv.edu/barchard/doubleentry/>, or from Kimberly A. Barchard, Department of Psychology, University of Nevada, Las Vegas, NV, 89154–5030, kim.barchard@unlv.edu
- Barchard, K. A., & Pace, L. A. (2008). Meeting the challenge of high quality data entry: A free double-entry system. *International Journal of Services and Standards*, 4, 359–376. <https://doi.org/10.1504/IJSS.2008.020053>
- 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. <https://doi.org/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. <https://doi.org/10.1016/j.chb.2013.02.021>
- Bateman, H. L., Lindquist, T. E., Whitehouse, R., & Gonzalez, M. M. (2013). Mobile application for wildlife capture-mark-recapture data collection and query. *Wildlife Society Bulletin*, 37, 838–845. <https://doi.org/10.1002/wsb.322>
- Buchehe, G., Och, B., Bolte, G., & Weiland, S. K. (2005). Single vs. double data entry. *Epidemiology*, 6, 130–131. <https://doi.org/10.1097/01.ede.0000147166.24478.f4>
- Burchinal, M., & Neebe, E. (2006). Data management: Recommended practices. *Monographs of the Society for Research in Child Development*, 71, 9–23. <https://doi.org/10.1111/j.1540-5834.2006.00402.x>
- Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. New York, NY: Cambridge University Press.
- Chen, K., Chen, H., Conway, N., Hellerstein, J. M., & Parikh, T. S. (2011). Usher: Improving data quality with dynamic forms. *IEEE Transactions on Knowledge and Data Engineering*, 23, 1138–1153. <https://doi.org/10.1109/TKDE.2011.31>
- Chen, K., Hellerstein, J. M., & Parikh, T. (2010). Designing adaptive feedback for improving data entry accuracy. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology* (pp. 239–248). New York, NY: ACM Press. <https://doi.org/10.1145/1866029.1866068>
- Cummings, J., & Masten, J. (1994). Customized dual data entry for computerized data analysis. *Quality Assurance*, 3, 300–303.
- Curtin, J. (2017). lmSupport: Support for linear models (R package version 2.9.11). Retrieved from <https://CRAN.R-project.org/package=lmSupport>
- DuChene, A. G., Jultgren, D. H., Neaton, J. D., Grambsch, P. V., Broste, S. K., Aus, B. M., & Rasmussen, W. L. (1986). Forms control and error detection procedures used at the coordinating center of the Multiple Risk Factor Intervention Trial (MRFIT). *Controlled Clinical Trials*, 7(Suppl.), 34–45. [https://doi.org/10.1016/0197-2456\(86\)90158-3](https://doi.org/10.1016/0197-2456(86)90158-3)
- 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. <https://doi.org/10.1016/j.combiomed.2008.07.008>
- Gibson, D., Harvey, A., Everett V., Parmar, M. K. B., & on behalf of the CHART Steering Committee. (1994). Is double data entry necessary? The chart trials. *Controlled Clinical Trials*, 15, 482–488. [https://doi.org/10.1016/0197-2456\(94\)90005-1](https://doi.org/10.1016/0197-2456(94)90005-1)
- Goldberg, S. I., Niemierko, A., & Turchin, A. (2008). Analysis of data errors in clinical research databases. *AMIA Annual Symposium Proceedings*, 6, 242–246. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656002>
- 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. <https://doi.org/10.1016/j.jbi.2008.08.010>
- Hoaglin, D. C., & Velleman, P. F. (1995). A critical look at some analyses of major league baseball salaries. *American Statistician*, 49, 277–285. <https://doi.org/10.1080/00031305.1995.10476165>
- 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. [https://doi.org/10.1016/S0197-2456\(03\)00089-8](https://doi.org/10.1016/S0197-2456(03)00089-8)
- Kozak, M., Krzanowski, W., Cichocka, I., & Hartley, J. (2015). The effects of data input errors on subsequent statistical inference. *Journal of Applied Statistics*, 42, 2030–2037. <https://doi.org/10.1080/02664763.2015.1016410>
- Kruskal, W. H. (1960). Some remarks on wild observations. *Technometrics*, 2. Retrieved from <https://doi.org/10.1080/00401706.1960.10489875>
- Lauritsen, J. M. (Ed.). (2000–2018). *EpiData data entry, data management and basic statistical analysis system*. Odense, Denmark: EpiData Association. Retrieved May 21, 2018, from <http://www.epidata.dk>
- McFadden, E. (1998). *Management of data in clinical trials*. New York, NY: Wiley.
- Ohmann C., Kuchinke W., Canham S., Lauritsen J., Salas N., Schade-Brittinger, C., ... Torres, F. (2011). Standard requirements for GCP-compliant data management in multinational clinical trials. *Trials*, 12, 85. <https://doi.org/10.1186/1745-6215-12-85>
- Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, 9(6), 1–8. Retrieved from <http://pareonline.net/getvn.asp?v=9&n=6>
- Paulsen, A., Overgaard, S., & Lauritsen, J. M. (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, e35087. <https://doi.org/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. [https://doi.org/10.1016/0197-2456\(92\)90205-E](https://doi.org/10.1016/0197-2456(92)90205-E)
- Schneider, J. K., & Deenan, A. (2004). Reducing quantitative data errors: Tips for clinical researchers. *Applied Nursing Research*, 17, 125–129. <https://doi.org/10.1016/j.apnr.2004.02.001>

- Stellman, S. (1989). The case of the missing eights: An object lesson in data quality assurance. *American Journal of Epidemiology*, 129, 857–860. <https://doi.org/10.1093/oxfordjournals.aje.a115200>
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.
- Tu, H., Oladimeji, P., Wiseman, S., Thimbleby, H., Cairns, P., & Niezen, G. (2016). Employing number-based graphical representations to enhance the effects of visual check on entry error detection. *Interacting with Computers*, 28, 194–207. <https://doi.org/10.1093/iwc/iwv020>
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Ueda, Y., & Saiki, J. (2012). Characteristics of eye movements in 3-D object learning: Comparison between within-modal and cross-modal object recognition. *Perception*, 41, 1289–1298. <https://doi.org/10.1068/p7257>
- Walther, B., Hossin, S., Townend, J., Abernethy, N., Parker, D., & Jeffries, D. (2011). Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data. *PLoS ONE*, 6, e25348. <https://doi.org/10.1371/journal.pone.0025348>
- Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods. *American Psychologist*, 53, 300–314. <https://doi.org/10.1037/0003-066X.53.3.300>