Checking for Data-Conversion Errors

by M. E. Kabay, PhD, CISSP-ISSMP
Professor of Computer Information Systems
School of Business & Management
Norwich University, Northfield VT

In both the classic triad of information assurance (IA) and in the Parkerian Hexad< http://www.mekabay.com/overviews/hexad_ppt.zip >, integrity is a fundamental attribute of information that must be protected. Data integrity refers to the correctness of information; for example, integrity can refer to consistency with data's original and intended state.

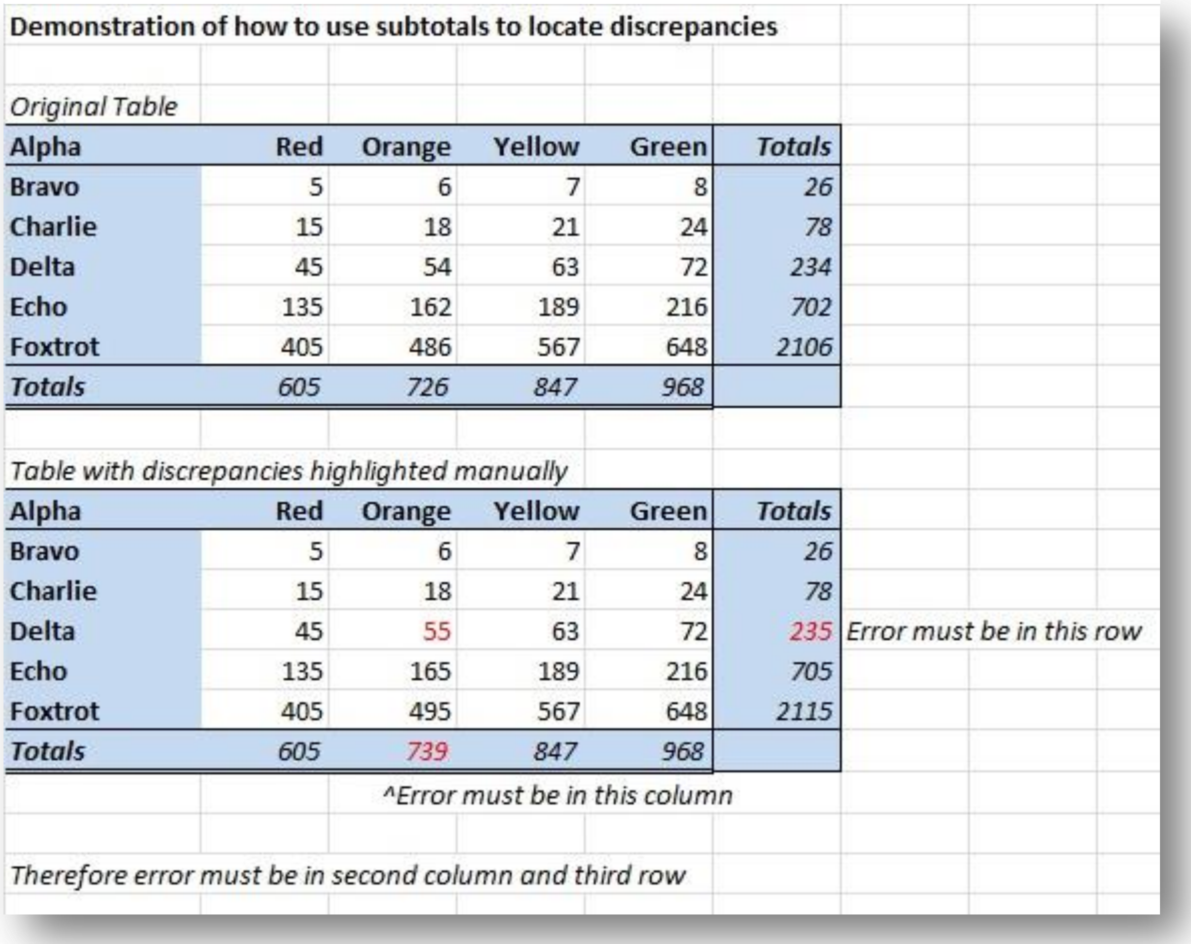
Recently, a colleague and several of his research students ran into a problem when they tried to import data from a comma-delimited file (CSV< http://creativyst.com/Doc/Articles/CSV/CSV01.htm >) into their version (2007< http://office.microsoft.com/en-us/excel-help/up-to-speed-with-excel-2007-RZ010062103.aspx?CTT=1 >) of MS-Excel, the widely used spreadsheet< http://www.webopedia.com/TERM/S/spreadsheet.html > program. They found unrecognized characters in the CSV file that showed up as squares with a question-mark inside. They asked me for help, and I loaded the CSV into MS-Word 2007, where it was obvious that the characters were TABs, even though they should not have been there given that all of the data were separated by commas.

After deleting the tabs using the global replace function (CTL-H) to locate every ^*t* character and replace it by nothing, the question arose of how to check the converted data against the original version that had contained the TAB characters. There was no point in applying the supposed correction if it caused discrepancies between the intended version of the data and the modified data.

Sure enough, we immediately located some places where additional fixes would be required to make the data conform to the intended arrangement of rows and columns. After we found the discrepancies, it became clear that none of the students had ever thought about how to locate differences between two versions of their data.

There are several ways of checking for alterations of data in an Excel spreadsheet or in those that provide similar functions.

Figure 1. Using row and column totals for error-checking.



Demonstration of how to use subtotals to locate discrepancies

Original Table

Alpha	Red	Orange	Yellow	Green	*Totals*
Bravo	5	6	7	8	*26*
Charlie	15	18	21	24	*78*
Delta	45	54	63	72	*234*
Echo	135	162	189	216	*702*
Foxtrot	405	486	567	648	*2106*
Totals	*605*	*726*	*847*	*968*	

Table with discrepancies highlighted manually

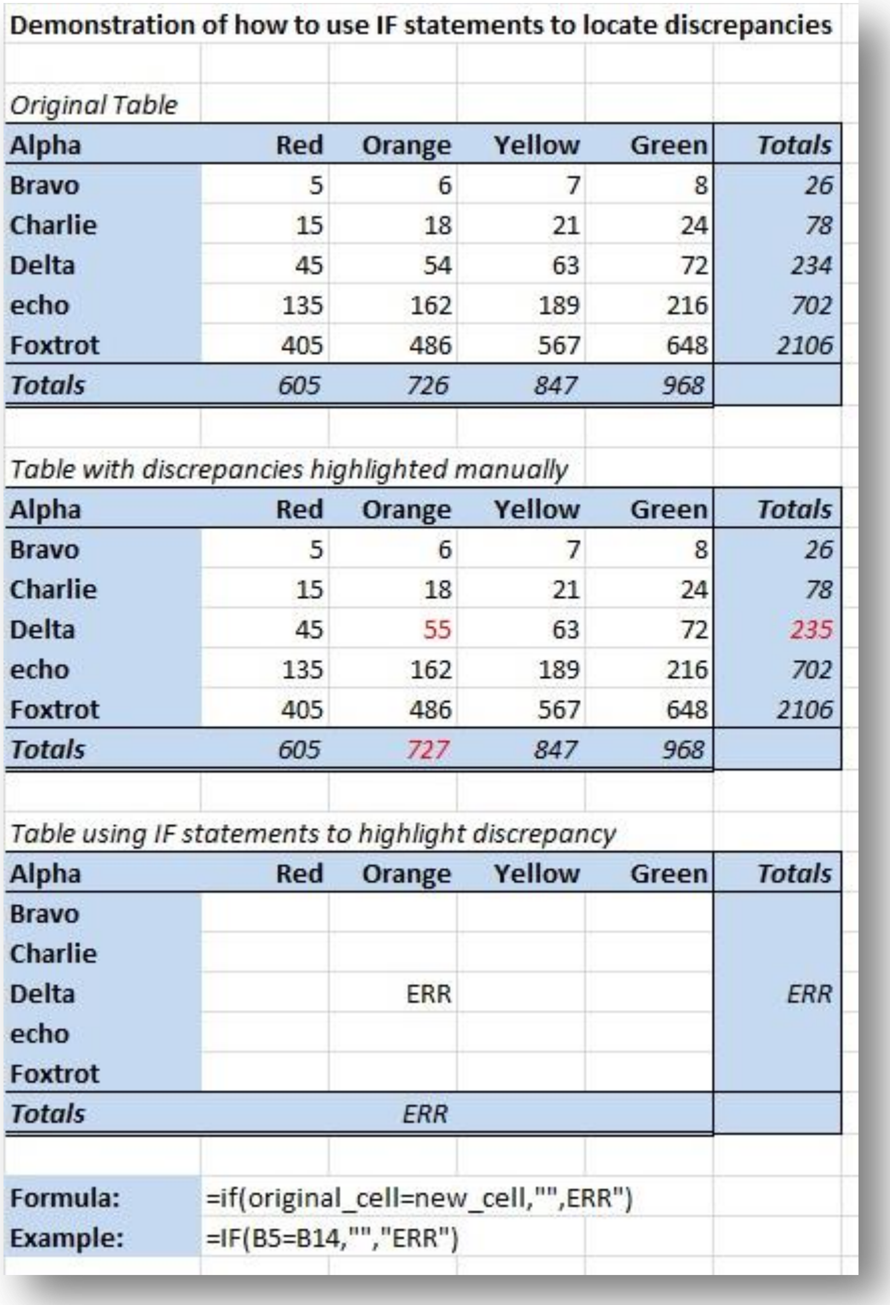
Alpha	Red	Orange	Yellow	Green	*Totals*	
Bravo	5	6	7	8	*26*	
Charlie	15	18	21	24	*78*	
Delta	45	55	63	72	*235*	*Error must be in this row*
Echo	135	165	189	216	*705*	
Foxtrot	405	495	567	648	*2115*	
Totals	*605*	*739*	*847*	*968*		

^Error must be in this column

Therefore error must be in second column and third row

- One of the oldest methods for locating changes in tabular data is to compute totals for each row and for each column and look for differences in those totals. The row and the column where the totals differ from the originals pinpoint the difference in the cell contents. This method was something I used routinely back in the days of manual calculations, before spreadsheets were so refined that they became a kind of programming language. Figure 1 shows what this simple method looks like.
- Today, a simple and quick method is to use an IF statement to put an error indicator into a cell. As shown in Figure 2, one can simply define a function that sets a cell to something like “ERR” if the original cell value doesn’t match the converted cell value.
- If for some reason you are not satisfied with simply printing an error message showing where a discrepancy lies, you can also use the conditional formatting options to colour a cell background as you see fit; in the example shown in Figure 3, all the correct cells are in green and discrepancies are flagged in red. The figure includes a screenshot of the conditional-formatting rules.

Figure 2. Using IF statements.



Demonstration of how to use IF statements to locate discrepancies

Original Table

Alpha	**Red**	**Orange**	**Yellow**	**Green**	***Totals***
Bravo	5	6	7	8	*26*
Charlie	15	18	21	24	*78*
Delta	45	54	63	72	*234*
echo	135	162	189	216	*702*
Foxtrot	405	486	567	648	*2106*
Totals	*605*	*726*	*847*	*968*	

Table with discrepancies highlighted manually

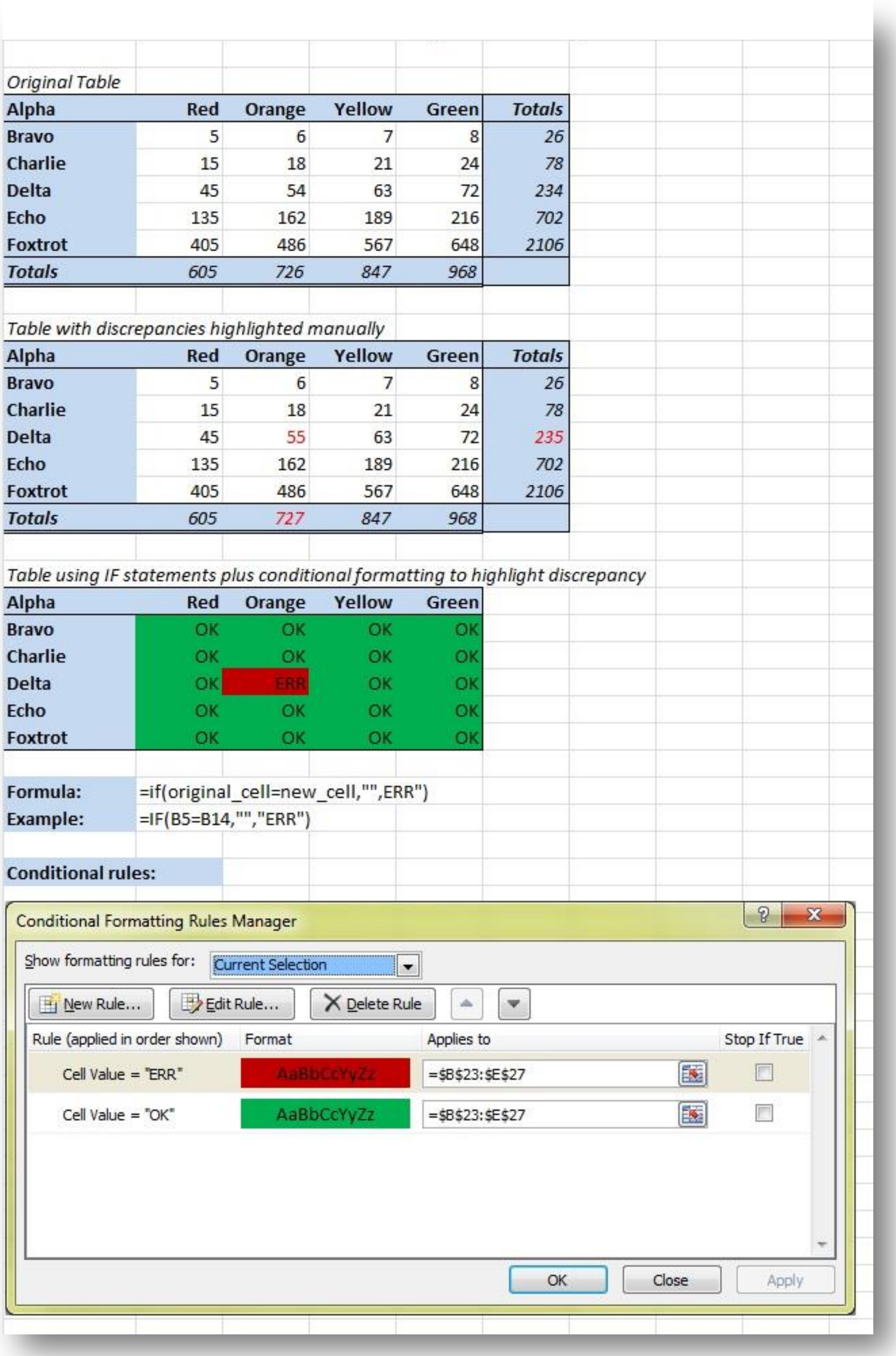
Alpha	**Red**	**Orange**	**Yellow**	**Green**	***Totals***
Bravo	5	6	7	8	*26*
Charlie	15	18	21	24	*78*
Delta	45	55	63	72	*235*
echo	135	162	189	216	*702*
Foxtrot	405	486	567	648	*2106*
Totals	*605*	*727*	*847*	*968*	

Table using IF statements to highlight discrepancy

Alpha	**Red**	**Orange**	**Yellow**	**Green**	***Totals***
Bravo					
Charlie					
Delta		ERR			*ERR*
echo					
Foxtrot					
Totals		*ERR*			

Formula:	=if(original_cell=new_cell,"",ERR")
Example:	=IF(B5=B14,"","ERR")

Figure 3. Using conditional formatting.



Original Table

Alpha	Red	Orange	Yellow	Green	*Totals*
Bravo	5	6	7	8	*26*
Charlie	15	18	21	24	*78*
Delta	45	54	63	72	*234*
Echo	135	162	189	216	*702*
Foxtrot	405	486	567	648	*2106*
Totals	*605*	*726*	*847*	*968*	

Table with discrepancies highlighted manually

Alpha	Red	Orange	Yellow	Green	*Totals*
Bravo	5	6	7	8	*26*
Charlie	15	18	21	24	*78*
Delta	45	55	63	72	*235*
Echo	135	162	189	216	*702*
Foxtrot	405	486	567	648	*2106*
Totals	*605*	*727*	*847*	*968*	

Table using IF statements plus conditional formatting to highlight discrepancy

Alpha	Red	Orange	Yellow	Green
Bravo	OK	OK	OK	OK
Charlie	OK	OK	OK	OK
Delta	OK	ERR	OK	OK
Echo	OK	OK	OK	OK
Foxtrot	OK	OK	OK	OK

Formula: =if(original_cell=new_cell,"",ERR")

Example: =IF(B5=B14,"","ERR")

Conditional rules:

Conditional Formatting Rules Manager

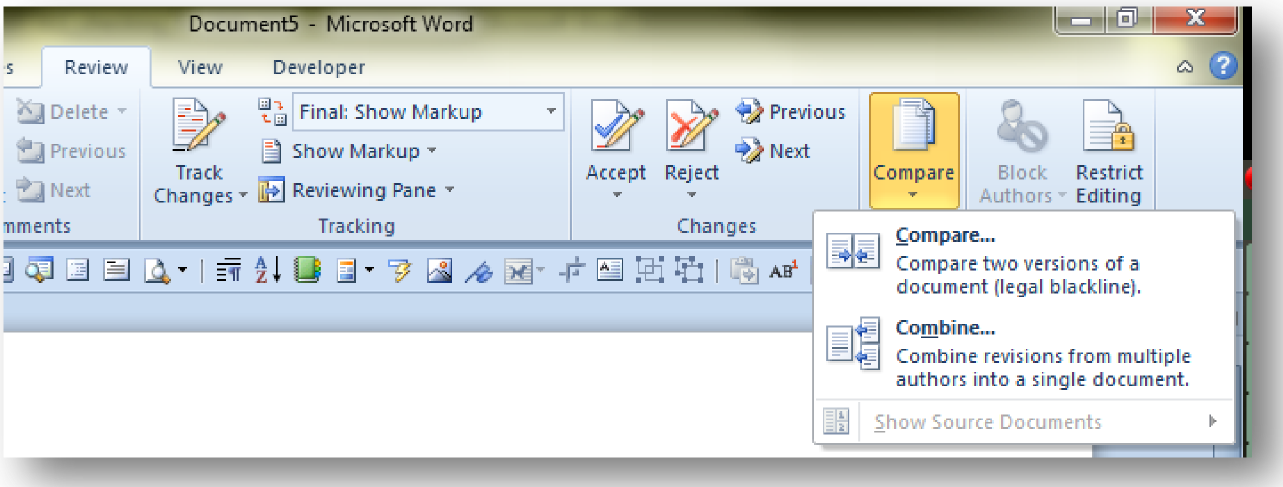
Show formatting rules for: Current Selection

Rule (applied in order shown)	Format	Applies to	Stop If True
Cell Value = "ERR"	AaBbCcYyZz	=$B$23:$E$27	☐
Cell Value = "OK"	AaBbCcYyZz	=$B$23:$E$27	☐

What if you have to compare other data files such as TXT plain-ASCII? A trick you can use is to arrange the files to be exactly the same in font, point size and position on the screen (e.g., filling an entire screen). In Windows, you can press Alt-Tab repeatedly to switch between the files. Any difference between the two will show up as a moving or changing element as you flash back and forth between the files. However, this method does depend on the viewer's attention for its effectiveness; it's also difficult to manage for large files that take more than one screen to visualize. For rows or columns in the thousands, it's impractical.

Another approach is to use WORD's file comparison feature. In Word 2007 and Word 2010, the **Review** tab has a **Compare** function that provides the option to "**Compare…** Compare two versions of a document (legal blackline)." Figure 4 shows the dialog to initiate a comparison. Everything that differs between the two documents will be highlighted in colour.

Figure 4. Using WORD comparison function.



If anyone wants to see the Excel 2010 spreadsheet used to create Figures 1 through 3, it is available here < http://www.mekabay.com/perception/011_checking_for_data-conversion_errors.xlsx > .

* * *

M. E. Kabay,< mailto:mekabay@gmail.com > PhD, CISSP-ISSMP, specializes in security and operations management consulting services and teaching. He Professor of Computer Information Systems in the School of Business and Management at Norwich University. Visit his Website for white papers and course materials.< http://www.mekabay.com/ >