

When Good Looks Aren't Enough



Lisa Eckler

When Good Looks Aren't Enough



Lisa Eckler CONSULTING INC.



When Good Looks Aren't Enough

Introduction

- ❑ Holistic, common-sense approach
- ❑ Allocate most effort to what's most important
- ❑ Basic coding techniques:
 - ❑ PROC CONTENTS
 - ❑ PROC COMPARE

When Good Looks Aren't Enough

Define terms (1)

Verification is the act of reviewing, inspecting, testing, etc. to establish and document that a product, service, or system meets the regulatory, standard, or specification requirements.

When Good Looks Aren't Enough

Define terms (2)

Validation refers to meeting the needs of the intended end-user or customer.

When Good Looks Aren't Enough

Define terms (3)

Data **quality assurance** is the process of profiling the data to discover inconsistencies, and other anomalies in the data and performing data cleansing activities to improve the data quality.

– Wikipedia



When Good Looks Aren't Enough

Mantra

- ❑ Check your assumptions
- ❑ Confirm similarities
- ❑ Focus on differences

When Good Looks Aren't Enough

Is QA a programming task?

- Yes... mostly
- The routine parts can and should be automated and repeatable

When Good Looks Aren't Enough

**“Computers are useless.
They can only give you
answers.”**

– Pablo Picasso



When Good Looks Aren't Enough

Questions

- Is it packaged properly?
- Have I seen this before?
- Is it – or some part of it – unlike anything I've seen before?



When Good Looks Aren't Enough

What parts can be programmed?

	Relationship to existing dataset		
	Entirely new	Similar to existing	Replacing existing
What to check:			
All rows present?	H	A/H	A
All columns present?	H	A/H	A
Reasonable values?	H	A (mostly)	n/a
Does display reflect data?	H	H (mostly)	H
All of the above?	H	A(/H)	A(/H)

H = human task

A = automated, programmable task

When Good Looks Aren't Enough

PROC CONTENTS (1)

```
proc contents data = SASHELP.MON1001;  
run;
```

The CONTENTS Procedure

Data Set Name	SASHELP.MON1001	Observations	132
Member Type	DATA	Variables	407
Engine	V9	Indexes	0
Created	Wednesday, May 12, 2004 11:38:23 PM	Observation Length	2036
Last Modified	Wednesday, May 12, 2004 11:38:23 PM	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label	M-competition 1001 Series, Monthly		
Data Representation	WINDOWS_32		
Encoding	us-ascii ASCII (ANSI)		

Engine/Host Dependent Information

Data Set Page Size	16384
Number of Data Set Pages	19
First Data Page	3
Max Obs per Page	8
Obs in First Data Page	4
Number of Data Set Repairs	0
File Name	C:\Program Files\SAS\SAS 9.1\ets\sashelp\mon1001.sas7bdat
Release Created	9.0101M3
Host Created	XP_PRO

When Good Looks Aren't Enough

PROC CONTENTS (2)

Alphabetic List of Variables and Attributes

#	Variable	Type	Len
2	S0387	Num	5
3	S0388	Num	5
4	S0389	Num	5
5	S0390	Num	5
6	S0391	Num	5
7	S0393	Num	5
8	S0394	Num	5
9	S0395	Num	5
10	S0396	Num	5
11	S0397	Num	5
12	S0398	Num	5
13	S0401	Num	5
14	S0403	Num	5
15	S0404	Num	5
16	S0405	Num	5
17	S0406	Num	5
18	S0407	Num	5
19	S0409	Num	5
20	S0410	Num	5
21	S0412	Num	5

... etc. (listing all 407 variables,
spanning ~ 10 pages)

Lisa Eckler CONSULTING INC.



When Good Looks Aren't Enough

Storing CONTENTS listing in dataset

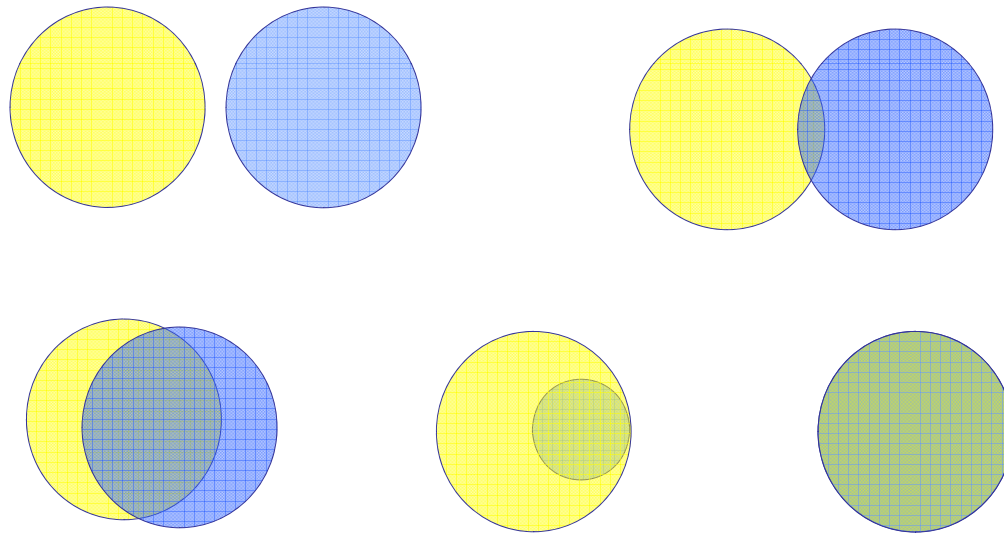
```
proc contents data = SASHELP.MON1001  
              noprint  
              out = CONTENTS_MON1001;  
run;
```

- ❑ produces a dataset with 407 obs
- ❑ no printed output
- ❑ can print, export, etc. as needed

When Good Looks Aren't Enough

Have I seen this before?

How is this result *expected* to compare to what you have seen before?



When Good Looks Aren't Enough

When the new dataset should be identical

record_id	a	b	c
1	*	*	*
2	*	*	*
3	*	*	*

record_id	a	b	c
1	*	*	*
2	*	*	*
3	*	*	*

```
proc compare compare = NEW  
              base    = ORIG;  
run;
```

When Good Looks Aren't Enough

When the new dataset should be identical

• • •

Number of Observations in Common: 3.

Total Number of Observations Read from WORK.ORIG: 3.

Total Number of Observations Read from WORK.NEW: 3.

Number of Observations with Some Compared Variables Unequal:
0.

Number of Observations with All Compared Variables Equal: 3.

NOTE: No unequal values were found. All values compared are
exactly equal. ←



When Good Looks Aren't Enough

When some columns should be the same

record_id	a	b	c
1	*	*	
2	*	*	
3	*	*	

record_id	a	b	d
1	*	*	
2	*	*	
3	*	*	

```
proc compare compare = NEW (keep = record_id a b)
                base   = ORIG(keep = record_id a b);
run;
```

When Good Looks Aren't Enough

When some "cells" should be the same

record_id	a	b	c
1	*	*	
2			
3	*	*	

	a	b	d
1	*	*	
3	*	*	
4			

```
proc compare compare = NEW (keep = record_id a b)
    base = ORIG(keep = record_id a b);
    where record_id in (1,3);
run;
```

When Good Looks Aren't Enough

Capturing the structure of each table

```
proc contents data = ORIG_105  
    out = CONTENTS_ORIG(keep = name type);  
run;
```

```
proc contents data = NEW_106  
    out = CONTENTS_NEW(keep = name type);  
run;
```

When Good Looks Aren't Enough

Joining the descriptions of table structures

```
proc sql;  
  create table CONTENTS_EITHER as  
  select CONTENTS_ORIG.name,  
         CONTENTS_NEW.name as name2,  
         CONTENTS_ORIG.type,  
         CONTENTS_NEW.type as type2  
  from CONTENTS_ORIG full join CONTENTS_NEW  
  on CONTENTS_ORIG.name = CONTENTS_NEW.name;  
quit;
```

When Good Looks Aren't Enough

Show only differences in table structure

```
proc print data = CONTENTS_EITHER noobs n  
  split = '*';  
  where (name <> name2) or (type <> type2);  
  title1 'List DIFFERENCES ONLY';  
run;
```

When Good Looks Aren't Enough

List DIFFERENCES ONLY

Variable Name	Variable Name	Variable Type	Variable Type
var_a100	var_a100	1	2
	var_a106	.	1
	var_a107	.	1
var_a99		1	.

N = 4

When Good Looks Aren't Enough

Summary

- ❑ Simple, automated steps to confirm that the parts that are supposed to be familiar are
- ❑ Check your assumptions
- ❑ Confirm similarities
- ❑ Focus on differences

When Good Looks Aren't Enough

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.

® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.

When Good Looks Aren't Enough

Questions?

lisa.eckler@sympatico.ca

Thank you!

Lisa Eckler CONSULTING INC.



When Good Looks Aren't Enough

Removing stored formats from data

```
data COPY_OF_NEW;  
    set NEW;  
run;  
  
proc datasets lib = work;  
    modify COPY_OF_NEW;  
    attrib _all_ format=;    ** ← **;  
quit;
```

When Good Looks Aren't Enough

See you in Baltimore!



Lisa Eckler CONSULTING INC.