WHERE, Oh, WHERE Art Thou? A Cautionary Tale for Using WHERE Statements and WHERE= Options
Britney D. Gilbert, Juniper Tree Consulting, Porter, Oklahoma

ABSTRACT
Using WHERE statements in your SAS programming can make processing data more efficient. However, the programmer needs to have a full understanding and be aware of how SAS processes multiple WHERE statements or combinations of WHERE statements and WHERE= dataset options in DATA steps and PROCEDURES. This paper explores examples of the combinations of WHERE statements and WHERE= dataset options in both DATA steps and PROCEDURES and the resulting logs and output.

INTRODUCTION
Whether we need to process large datasets or produce repeat tables, listings, or figures, the WHERE statement and/or WHERE= dataset option is one of the fundamental tools a programmer has in their toolbox. As such, it is important to understand how they work independently and in combination in both DATA steps and PROCEDURES.

To review, WHERE statements increase efficiency by only processing records that meet the conditions specified in the WHERE clause. When using multiple input datasets, a WHERE= dataset option can be used to select records for processing for that specific input dataset. But, what happens when we use a WHERE statement and WHERE= in combination or if we program multiple WHERE statements in a DATA step or PROCEDURE?

To demonstrate the behaviors of WHERE statements and WHERE= dataset option, a sample dataset and SAS® 9.3 was used. In this sample ADSL dataset, there are 100 subjects from three sites (001, 002, 003) in three regions (US, Europe, Other). Sex and age variables are included and an age grouping (AGEGR1) and a population flag (FASFL) has been derived.

WHERE STATEMENTS AND OPTIONS IN A DATA STEP
Using a WHERE reduces the number of records read into the DATA step as opposed to IF statements where all records are read first and then filtered by the conditions of the statement. However, unlike IF statements, multiple WHERE statements cannot be used in a single DATA step unless you use a WHERE ALSO or a WHERE SAME AND statement.

The WHERE ALSO and WHERE SAME ALSO statement will execute without a proceeding WHERE statement, but the log will indicate that the “WHERE clause has been augmented.” This is exemplified in Display 2. Has the WHERE clause been augmented? One condition was specified and one condition was processed.

If a log check program is used for review, then the programmer needs to be cautious when only looking for key phrases without a full log review of the subsequent log NOTES and record counts to verify expected results.
USING MULTIPLE WHERE STATEMENTS IN A DATA STEP

As previously mentioned, multiple WHERE statements cannot be used in a DATA step in the same manner as IF statements. When SAS® encounters a second WHERE statement in a DATA step, the first is replaced by the second.

In Display 3, the programming examples want to subset our sample ADSL dataset for the Full Analysis Set (FASFL = "Y") and Male Subjects (SEXN = 1).

At first, the program uses a DATA step with two WHERE statements: WHERE SEXN = 1 and WHERE FASFL = "Y". Reviewing the log NOTES, we see that the WHERE statement has been replaced and the resulting dataset EXAMPLE1A is only subsetted by FASFL = "Y" and has 99 observations. This is not our expected result.

Next, the program uses a DATA step with a WHERE statement followed by a WHERE ALSO statement: WHERE SEXN = 1 and WHERE ALSO FASFL = "Y". With this program, the resulting dataset EXAMPLE1B has 51 records and the log confirms that the records with both FASFL = "Y" and SEXN = 1 are read and processed. This is our expected result!

Finally to confirm the behavior of a WHERE SAME AND statement, a similar program is constructed using a WHERE statement followed by a WHERE SAME AND statement. The resulting dataset EXAMPLE1C has the expected 51 observations and the subsetting is confirmed in the log.

USING A WHERE STATEMENTS AND A WHERE= OPTION IN A DATA STEP

Using the combination of a WHERE statement and a WHERE= dataset option in a DATA step, is not a successful programming strategy. When a WHERE statement and WHERE= dataset option apply to the same dataset, SAS® uses the DATA step option and ignores the WHERE statement.

In Display 4, the programming example attempts to subset the ADSL for the Full Analysis Set (FASFL = "Y") and Male Subjects (SEXN = 1) using a WHERE= option, WHERE = (FASFL = "Y"), and a WHERE statement, WHERE SEXN = 1. In these cases, only the WHERE= option will be processed and the resulting dataset EXAMPLE2 has 99 observations. Again, this is not our expected result.

Reviewing the log NOTES, confirm that records with FASFL = "Y" were processed and not FASFL = "Y" AND SEXN = 1. Further, a WARNING appears in the log indicating that the WHERE statement cannot be applied.
WHERE, Oh, WHERE Art Thou? A Cautionary Tale for Using WHERE Statements and WHERE= Options, continued

WHERE STATEMENTS AND OPTIONS IN A PROCEDURE

Because IF statements are only valid in DATA steps, WHERE statements and WHERE= dataset options are used to subset observations in SAS® PROCEDURES. However, using multiple WHERE statements and combinations of WHERE statements and WHERE= dataset options behave differently.

USING MULTIPLE WHERE STATEMENTS IN A PROCEDURE

Multiple WHERE statements in a SAS® PROCEDURE behave the same as when they are used in a DATA step. When SAS® encounters a second WHERE statement in a PROCEDURE, the first is replaced by the second.

In Display 5, the programming examples want to sort a subset our sample ADSL dataset for the Full Analysis Set (FASFL = “Y”) and Male Subjects (SEXN = 1).

At first, the program uses a PROC SORT with two WHERE statements: WHERE SEXN = 1 and WHERE FASFL = “Y”. Reviewing the log NOTES, we see that the WHERE statement has been replaced and the resulting dataset EXAMPLE3A is only subsetted by FASFL = “Y” before sorting and has 99 observations. Again, this is not our expected result.

Next, the program submits a PROC SORT with a WHERE statement followed by a WHERE ALSO statement: WHERE SEXN = 1 and WHERE ALSO FASFL = “Y”. With this program, the resulting dataset EXAMPLE3B has 51 records and the log confirms that the records with both FASFL = “Y” and SEXN = 1 are read and sorted. Here is our expected result!

Finally to confirm the behavior of a WHERE SAME AND statement in a PROCEDURE, a similar program is constructed using a WHERE statement followed by a WHERE SAME AND statement. The resulting dataset EXAMPLE3C has the expected 51 observations and the subsetting is confirmed in the log.

USING A WHERE STATEMENTS AND A WHERE= OPTION IN A PROCEDURE

Unlike the DATA step, in a PROCEDURE a WHERE= dataset option and a WHERE statement can be used in combination successfully.

In Display 5, the programming examples want to sort a subset our sample ADSL dataset for the Full Analysis Set (FASFL = “Y”) and Male Subjects (SEXN = 1) and then sort a subset of subjects in the Full Analysis Set (FASFL = “Y”), Males (SEXN = 1), and with Age less than 65 years (AGEGR1N = 1).

The first programming example shows a PROC SORT with a WHERE= dataset option, WHERE = (FASFL = “Y”), and a WHERE Statement, WHERE SEXN = 1. Reviewing the log shows that this is a successful programming strategy in a PROCEDURE. The resulting dataset EXAMPLE4A has the expected 51 observations.
and the log NOTES confirm that records with both FASFL = “Y” AND SEXN = 1 were processed.

The second programming example uses multiple WHERE statements with a WHERE= dataset option. Recall that previously we showed that a second WHERE statement is replaced in a PROC SORT. However, when used in combination with a WHERE= dataset option, the log returns a different log message around the WHERE statements. When reviewing the log around EXAMPLE4B, we see the WHERE clause was augmented. But, was it really? At further examination, we see 95 observations rather than the expected 48 observations. Moreover, the later log NOTES contradict the previous ones and describes the subsetting as FASFL = "Y" and AGEGR1N = 1. The SEXN = 1 was not used in the subsetting. In this case, the SAS® log NOTES are false. The WHERE clause was not augmented it was replaced!

The last two programming examples show that by using a WHERE ALSO or a WHERE SAME AND statement will execute and correctly return our expected results with 48 observations that have FASFL = “Y” and SEXN = 1 and AGEGR1N = 1.

CONCLUSION

In this paper, the different behaviors of combining WHERE statements and WHERE= dataset options has been demonstrated. Further, misleading SAS® log messages of NOTE: Where clause has been augmented has been exposed. Although, WHERE statements and WHERE= dataset options are a staple programming, the simplicity of the syntax should not be overlooked in log reviews. The programmer needs to fully understand how these statements and options in combination are being processed in DATA steps and PROCEDURES. It is critical to review logs for expected results including confirming any WHERE clause augmentation or replacement and number of observations in the resulting output dataset.

ACKNOWLEDGMENTS

First, I would like to thank my Lord, Jesus Christ. It is through Him that I find my strength, patience, and resolve.

Next, I would like to thank my family: my encouraging husband, Justin, and my kids (Hope, Faith, Justin, Danny, Charity, and Paul) who are my never-ending source of happiness.

RECOMMENDED READING

• Base SAS® Language Reference
• Base SAS® Procedures Guide

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Britney Gilbert
Juniper Tree Consulting, LLC
Britney.Gilbert@JuniperTreeConsulting.com
www.JuniperTreeConsulting.com
@JuniperTree19

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 trademarks of their respective companies.