## Exploring the Data

In this demonstration, we explore the relationship between several predictor variables and a response variable.

This PROC SGSCATTER statement specifies the school data set. The COMPARE statement generates a panel of scatter plots with shared axes for variables in the school data set: the response **Reading3** (to be plotted on the Y axis) and the predictors **Words1**, **Letters1**, and **Phonics1** (which will appear as consecutive plots on the X axis). As you'll see, the three scatter plots will share a common Y axis. A TITLE statement specifies a title for the panel of scatter plots.

We run the code.

```
proc sgscatter data=mydata.school;
   compare y=reading3 x=(words1 letters1 phonics1);
   title 'Scatter Plots of READING3 by WORDS1 LETTERS1 and PHONICS1';
run;
```

It is always a good idea to check the log to verify that the code ran as expected. In this log, everything looks good.

Now, let's look at the results. Notice the common Y axis for **Reading3**, which is shared by scatter plots for the three predictors.

If we drew a straight line through each of the three scatter plots, all would have a positive slope. So, we can see that all three predictor variables appear to have a positive relationship with the response variable. As **Words1** increases, **Reading3** also increases, and so on. Which predictor seems to have the strongest relationship with the response? In the **Words1** plot, notice that the points are closer to a regression line that could be fit to the data. Because this is clearly a linear relationship, we can say that **Words1** has the strongest linear correlation with the response. Which predictor has the weakest correlation? In the **Phonics1** plot, the points are more widely scattered. Thus, **Phonics1** appears to have the weakest linear correlation with the response variable.

---