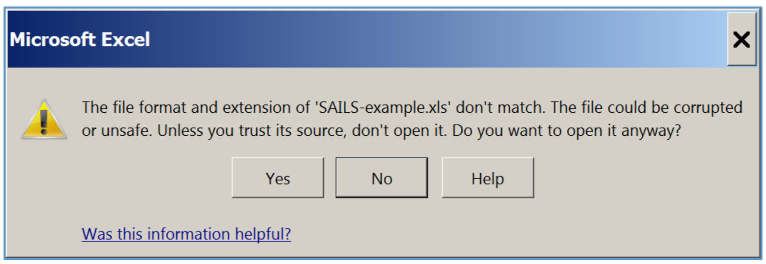
project SAILS®
Standardized Assessment of Information Literacy Skills

A Basic Guide to Analyzing Individual Scores Data with SPSS

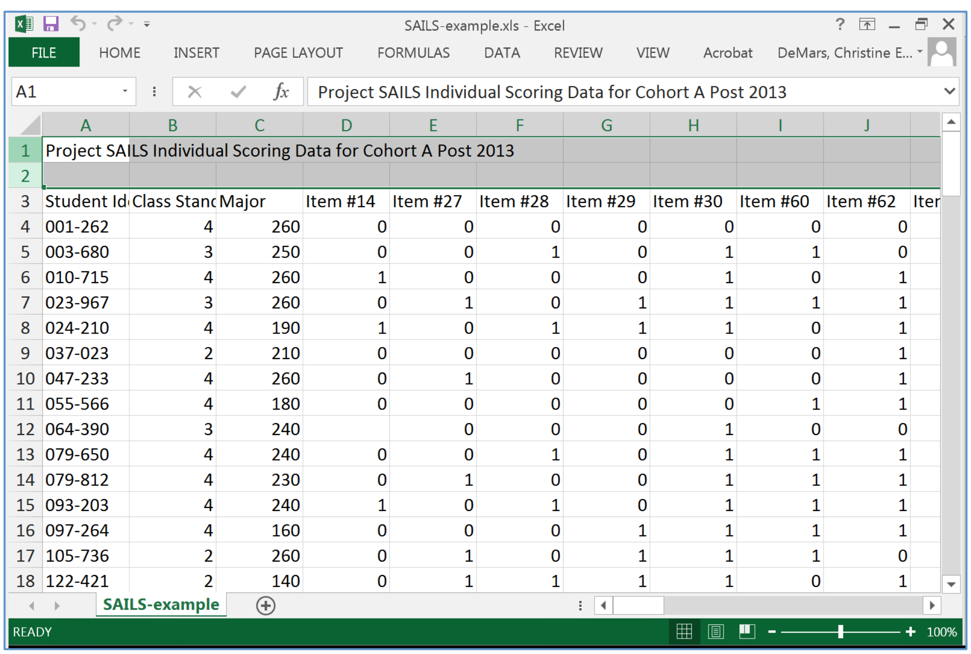
Step 1. Clean the data file

Open the Excel file with your data. You may get the following message:



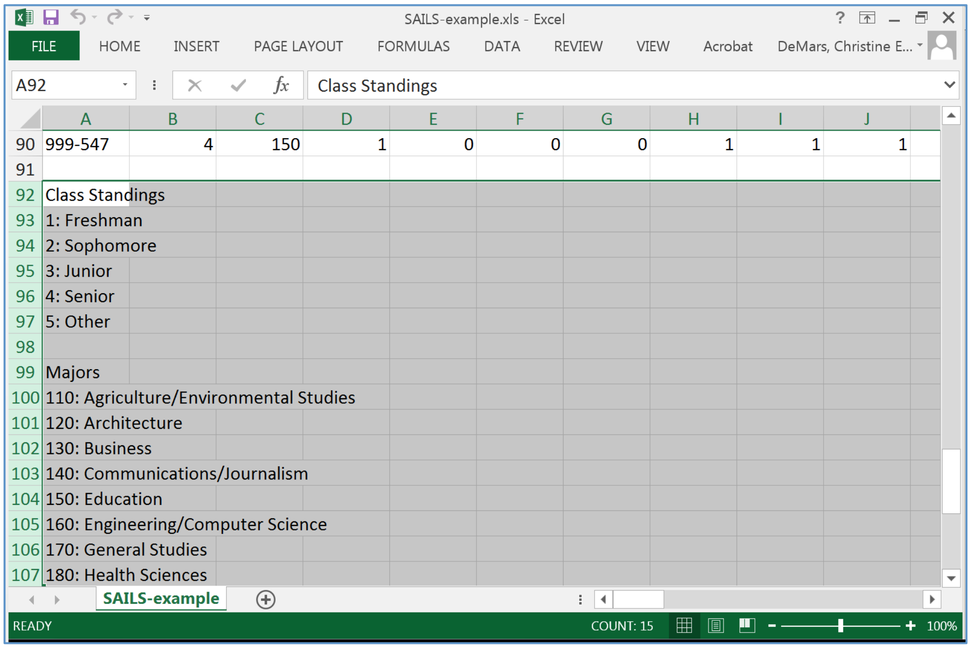
If you get this message, click yes.

Delete the first 2 lines of the data file so that the first row contains the names of the variables.



	A	B	C	D	E	F	G	H	I	J	
1	Project SAILS Individual Scoring Data for Cohort A Post 2013										
2											
3	Student Id	Class Stanc	Major	Item #14	Item #27	Item #28	Item #29	Item #30	Item #60	Item #62	Iter
4	001-262	4	260	0	0	0	0	0	0	0	
5	003-680	3	250	0	0	1	0	1	1	0	
6	010-715	4	260	1	0	0	0	1	0	1	
7	023-967	3	260	0	1	0	1	1	1	1	
8	024-210	4	190	1	0	1	1	1	0	1	
9	037-023	2	210	0	0	0	0	0	0	1	
10	047-233	4	260	0	1	0	0	0	0	1	
11	055-566	4	180	0	0	0	0	0	1	1	
12	064-390	3	240		0	0	0	1	0	0	
13	079-650	4	240	0	0	1	0	1	1	1	
14	079-812	4	230	0	1	0	0	1	1	1	
15	093-203	4	240	1	0	1	0	1	1	1	
16	097-264	4	160	0	0	0	1	1	1	1	
17	105-736	2	260	0	1	0	1	1	1	0	
18	122-421	2	140	0	1	1	1	1	0	1	

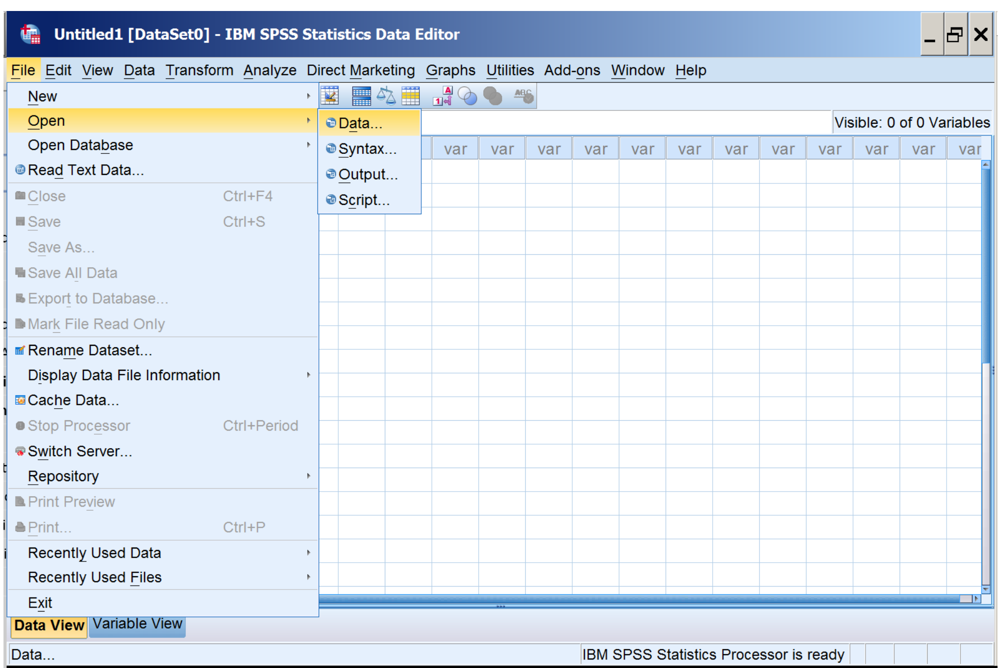
Then delete all lines after the data. (Note that you will need the information in these lines later so save it in another worksheet or file. Then you will be able to check and see that students in Major 130 were business majors, using the listing below as an example.)



Save the file with a new name.

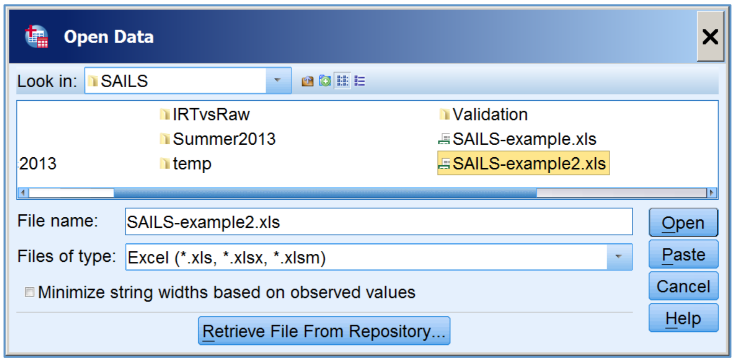
Step 2. Import the data file into SPSS

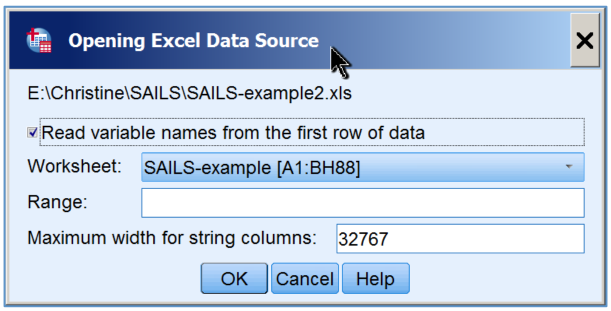
Open SPSS, and go to File -> Open -> Data



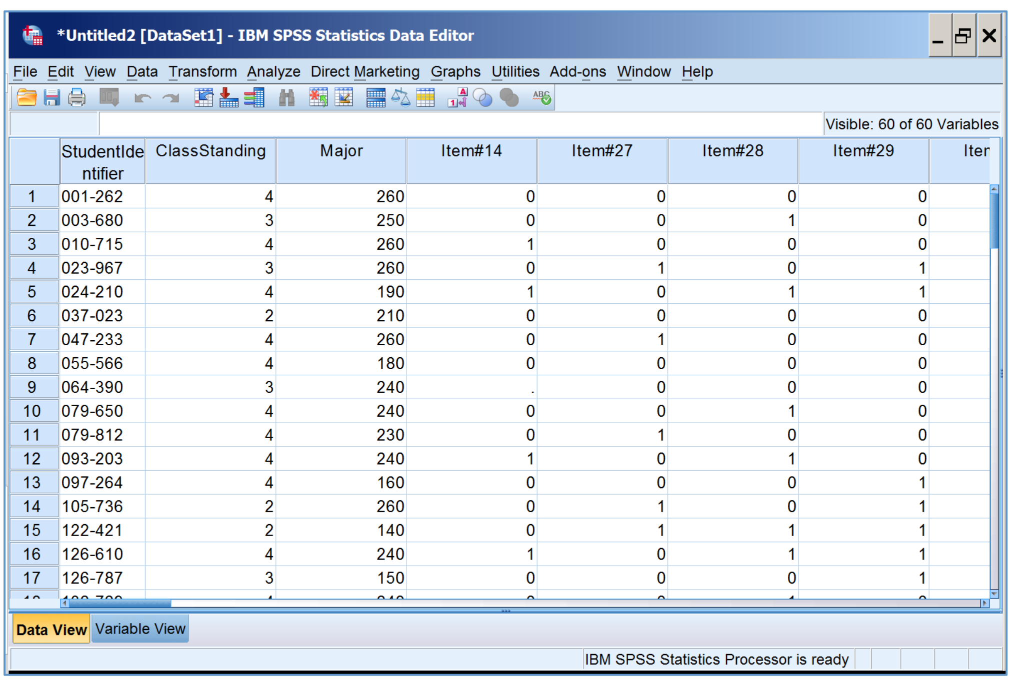
In the dialog box, choose "Excel" for "Files of type:"

Navigate to the folder where you saved the Excel file and click on your file so that the name shows up in the File name box. Click "Open" and then "OK" in the next box (make sure "Read variable names from the first row of data" is checked).





Your data should now look something like this:



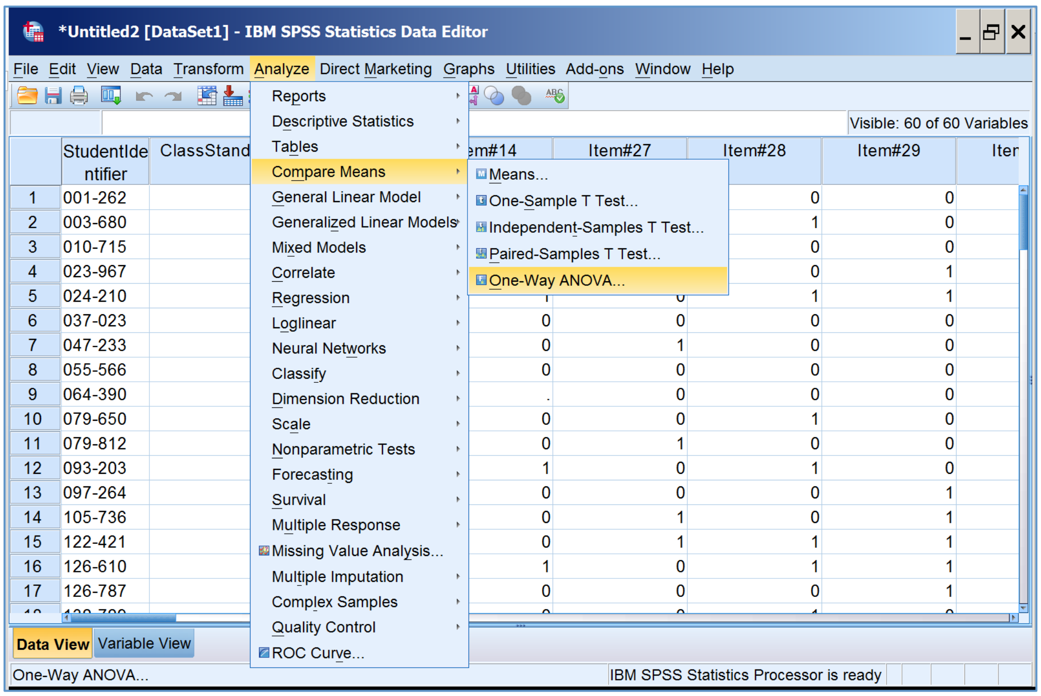
Step 3. Performing an Analysis

Analysis 1. Comparisons of Groups with ANOVA and Tukey Post Hoc

You might want to know if students from different classes or majors scored statistically significantly differently on the test. Statistical significance means that the differences are unlikely to be this large or larger due to chance; it does not always mean the differences are large enough to be meaningful (that requires personal judgment). You can use an ANOVA (analysis of variance) to estimate this.

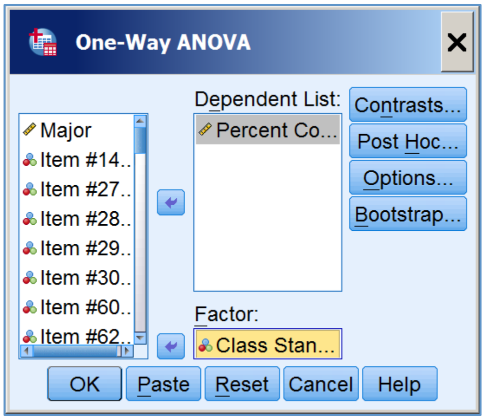
There are many ways to run an ANOVA in SPSS. Here is an example using the one-way ANOVA procedure.

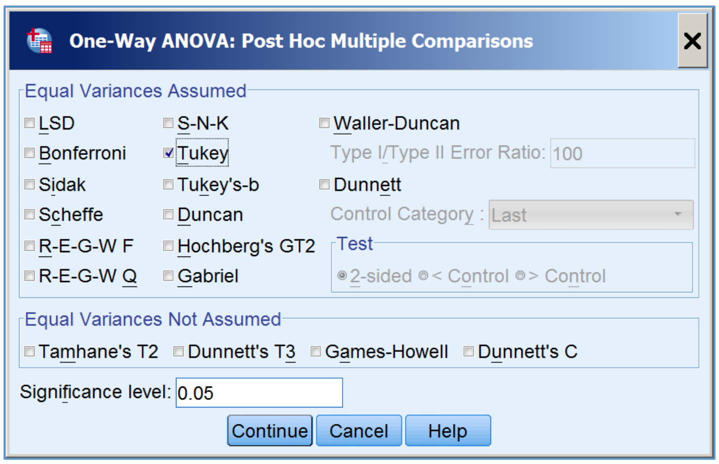
Analyze -> Compare Means -> One-Way ANOVA



Choose the dependent variable (the variable on which you want to compare groups – either Percent Correct or Number Correct) and use the arrow key to click it into the Dependent List. Choose Class Standing or Major as the Factor.

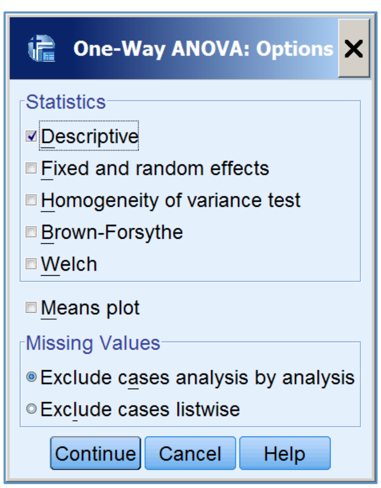
If you have more than two groups (for example, freshmen, sophomores, juniors, and seniors), you will likely want to compare differences between pairs of groups. Click the "Post Hoc" button.





These post-hoc tests (except for LSD) are methods for keeping the TOTAL Type I error rate, across all comparisons, below the level specified in the Significance Level box. Tukey is one of the most common and is recommended.

Once Tukey (or other post-hoc test) is selected, click Continue and return to the One-Way ANOVA box. At this point, if you haven't run the means (averages) for each group, click Options. Here you can check “Descriptive.”



Click Continue, then once you are back at the One-Way ANOVA box, click OK. You will get output similar to this:

Descriptives

Percent Correct

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
2	7	39.98571%	12.057975%	4.557486%	28.83395%	51.13748%	27.300%	61.800%
3	15	37.10000%	6.902484%	1.782214%	33.27753%	40.92247%	27.300%	47.300%
4	65	47.83692%	12.549683%	1.556597%	44.72726%	50.94658%	23.600%	76.400%
Total	87	45.35402%	12.409400%	1.330427%	42.70922%	47.99883%	23.600%	76.400%

My dataset had 7 sophomores, 15 juniors, and 65 seniors (column headed "N"). With this small sample, it would be a good idea to combine sophomores and juniors into a single group. However, to illustrate multiple group comparisons, they are separate for this example.

The juniors scored approximately 3 percentage points below the sophomores, but the seniors scored nearly 8 percentage points above the sophomores (column headed “Mean").

The column labelled "Std. Deviation" indicates how much the scores varied within groups. The juniors' scores varied less than the sophomores' or seniors' scores.

The column labelled "Std. Error" is the standard error of the group mean—the standard deviation of the scores divided by the square root of the sample size. The sophomores and seniors have roughly similar standard deviations, but the seniors have a smaller standard error (more stable mean) because the mean is based on more students.

The “95% Confidence Interval” gives a plausible range[1] for the mean—the interval will be narrower when the standard error is smaller. Finally, “Minimum” and “Maximum” are the lowest and highest scores in each group.

The ANOVA table is also part of the output. It answers the question, “Overall, do the groups differ significantly in their mean scores?”

ANOVA

Percent Correct

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1624.376	2	812.188	5.872	.004
Within Groups	11619.040	84	138.322		
Total	13243.416	86			

[1] If we drew repeated samples from a population and constructed 95% confidence intervals for each sample, 95% of the ranges would include the population mean.

Conventionally, if the probability (the value in the Sig column) of observing group means that differ this much or more is < .05, the groups are deemed to be significantly different. Differences this large or larger would occur by chance less than 5% of the time if the group means were equal. In this case, we would report: "The groups were statistically significantly different ($F_{2,84}$ = 5.872, *p* = .004)."

Post Hoc Tests
Multiple Comparisons
Dependent Variable: Percent Correct
Tukey HSD

(I) Class Standing	(J) Class Standing	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2	3	2.885714%	5.383468%	.854	-9.95901%	15.73044%
	4	-7.851209%	4.678493%	.220	-19.01390%	3.31148%
3	2	-2.885714%	5.383468%	.854	-15.73044%	9.95901%
	4	-10.736923%*	3.368900%	.006	-18.77498%	-2.69887%
4	2	7.851209%	4.678493%	.220	-3.31148%	19.01390%
	3	10.736923%*	3.368900%	.006	2.69887%	18.77498%

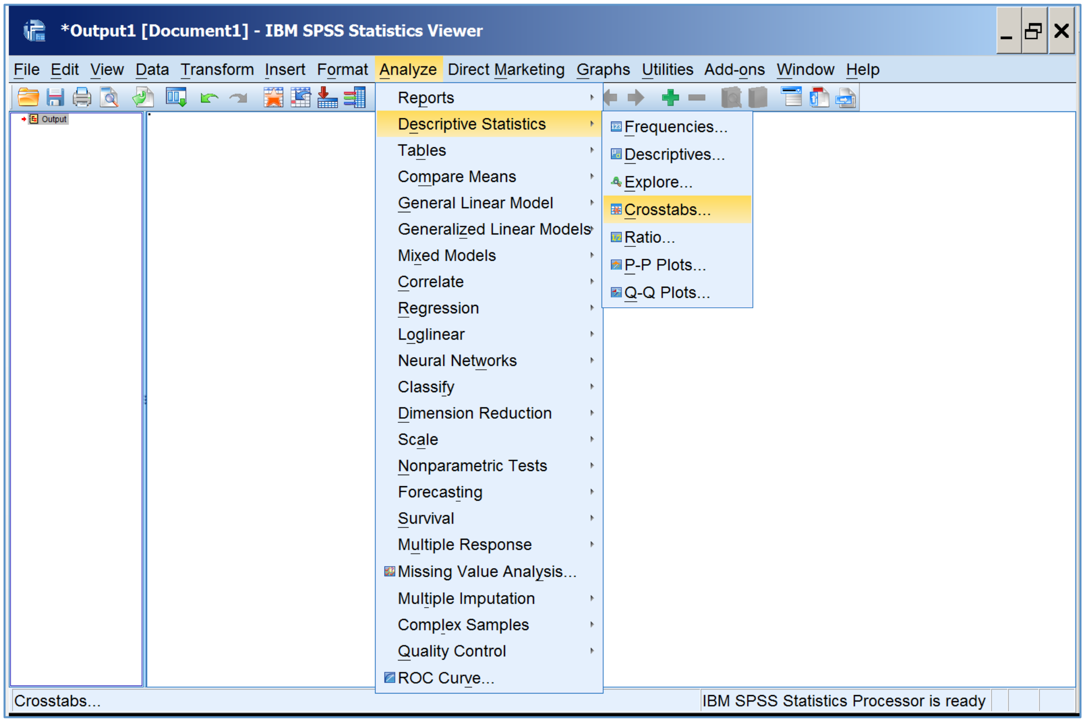
*. The mean difference is significant at the 0.05 level.

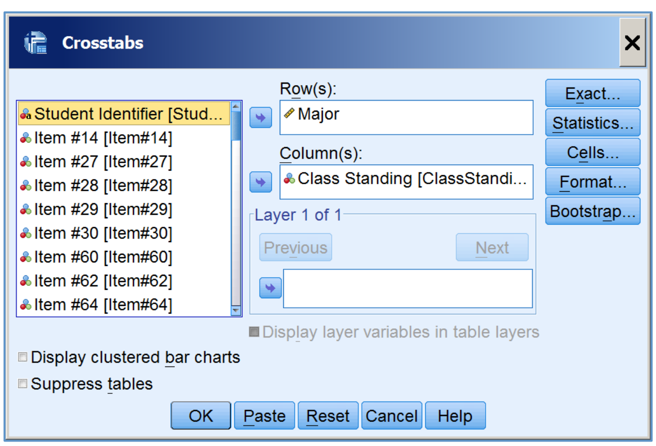
The post-hoc table shows which of the groups are different from each other. The Tukey test takes into account that we are computing three differences and adjusts the reported significance levels accordingly.

The "Sig" column indicates the probability of a difference in means this large or larger if there were no difference. The difference between sophomores (group 2) and juniors (group 3) is not statistically significant (*p* = .854), nor is the difference between sophomores and seniors (group 4) (*p* = .220). Juniors and seniors are significantly different (*p* = .006). Seniors scored 10.7 percentage-points higher. A plausible range for the difference between seniors and juniors is 2.7 to 18.8 percentage points. The range is large because the sample size is small. The difference of 10.7 percentage-points is an unstandardized effect size.

Analysis 2. Selecting Groups to Analyze

Comparisons between students in different majors might also be of interest. Unless your sample is very large, you will want to select the most common majors or combine related majors. To see how many students you have in each major within each class, run Crosstabs (Analyze -> Descriptive Statistics -> Crosstabs)



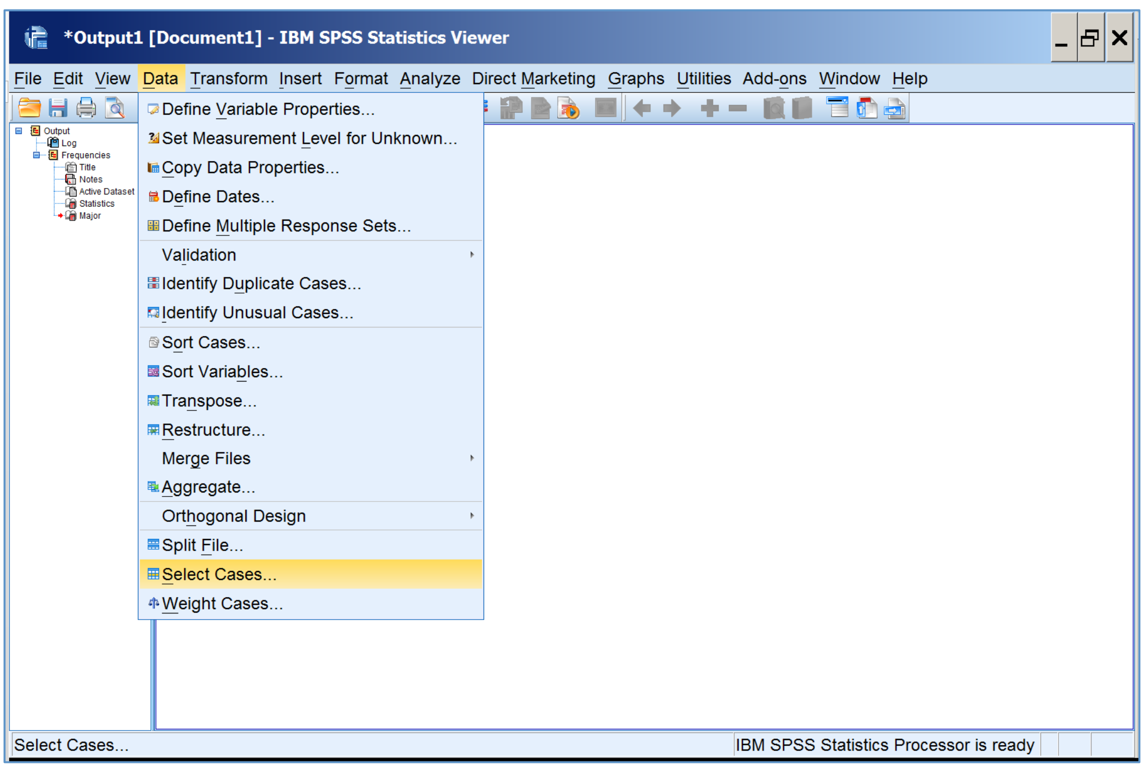


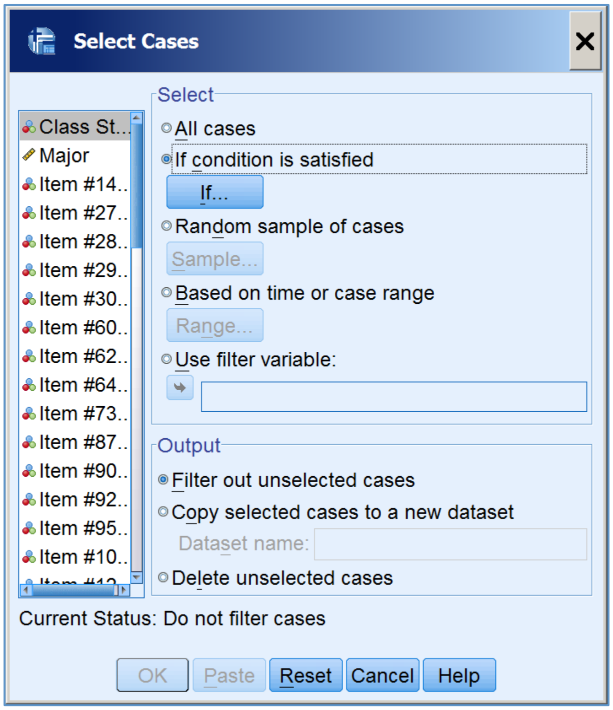
The output will show how many students in each major are in each class (for example, there are 4 sophomores (class standing 2) in major 260).

Major * Class Standing Crosstabulation
Count

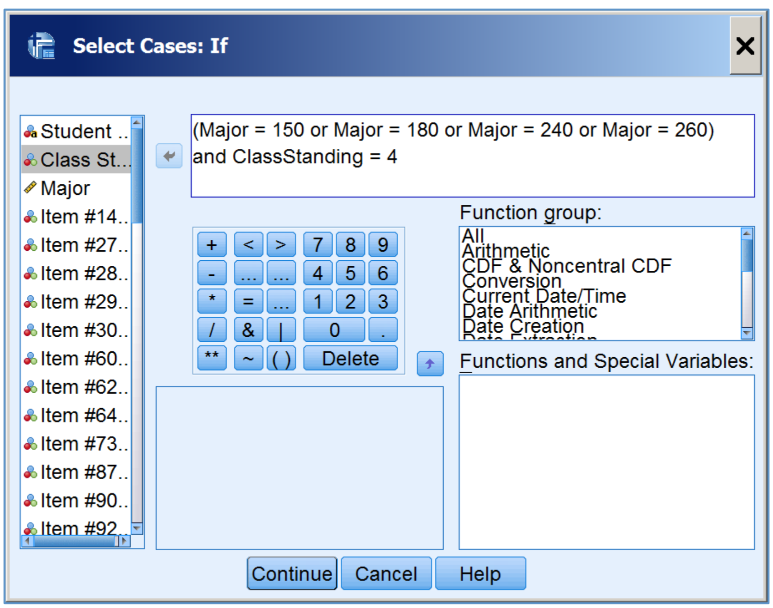
		Class Standing			Total
		2	3	4	
Major	130	0	0	1	1
	140	1	0	3	4
	150	0	4	14	18
	160	0	0	2	2
	180	1	4	8	13
	190	0	0	1	1
	210	1	1	5	7
	230	0	0	3	3
	240	0	3	10	13
	250	0	2	5	7
	260	4	1	13	18
Total		7	15	65	87

In my data, I did not have enough sophomores or juniors to compare majors. For the seniors, I decided to compare majors with 10 or more students. Do this through Data -> Select Cases.





Click on “If” and type the conditions you want into the box (you can enter variable names by selecting from the box on the left and clicking the arrow, but you will have to type the ANDs and ORs and ='s and values). Example:



Run the ANOVA as for ClassStanding (see examples above), but using the variable Major in the "Factor" box. Here is the output from my sample data.

Descriptives

Percent Correct

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
150	14	46.75714%	13.645834%	3.647003%	38.87827%	54.63601%	27.300%	72.700%
240	10	56.36000%	12.213763%	3.862331%	47.62280%	65.09720%	34.500%	74.500%
260	13	41.26154%	10.109776%	2.803947%	35.15226%	47.37081%	23.600%	60.000%
Total	37	47.42162%	13.216386%	2.172760%	43.01506%	51.82818%	23.600%	74.500%

ANOVA

Percent Correct

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1298.434	2	649.217	4.424	.020
Within Groups	4989.789	34	146.759		
Total	6288.223	36			

My groups differed significantly ($F_{2,34}$ = 4.424, *p* =.020).

To see how the groups differed from each other examine the Tukey results:

Post Hoc Tests
Multiple Comparisons

Dependent Variable: Percent Correct
Tukey HSD

(I) Major	(J) Major	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
150	240	-9.602857%	5.015835%	.150	-21.89384%	2.68812%
	260	5.495604%	4.666033%	.474	-5.93821%	16.92942%
240	150	9.602857%	5.015835%	.150	-2.68812%	21.89384%
	260	15.098462%*	5.095583%	.015	2.61206%	27.58486%
260	150	-5.495604%	4.666033%	.474	-16.92942%	5.93821%
	240	-15.098462%*	5.095583%	.015	-27.58486%	-2.61206%

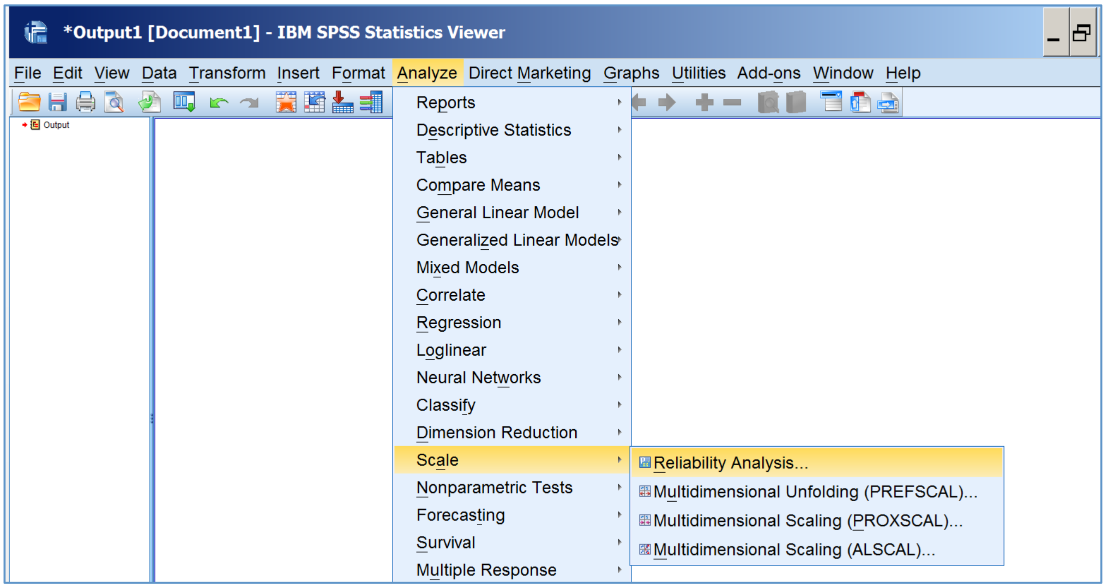
*. The mean difference is significant at the 0.05 level.

Major 260 (Other) scored 15 percentage-points lower than Major 240 (Science/Math). This difference was statistically significant (*p* =. 015) using the Tukey procedure to control the familywise Type I error rate. (see comments above about post hoc comparisons for more information).

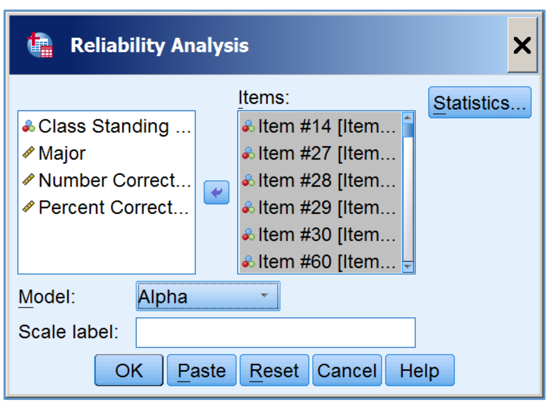
Analysis 3. Reliability

You might want to know the reliability of the scores for your students. Reliability is an index of how well students can be separated by their scores. If your students are more alike than the SAILS norming sample, they will not be separated as reliably. If your students are more variable than the norming sample, they will be separated more reliably.

On the Analyze menu, go to Scale -> Reliability Analysis



Click the items included in the score (generally, all items). You can select multiple items by clicking on the first one, holding the shift key, and selecting the last item. Click the arrow to move the selection to the "Items" box.



The most common reliability estimate is coefficient alpha (α), also called Cronbach's α. For items scored right/wrong (such as on the SAILS test), this index is also called KR-20. This is an estimate of the correlation between scores on the current test form and another hypothetical test form selected randomly from the same hypothetical pool of items. Choose Alpha in the selection box next to Model. An example of the resulting output:

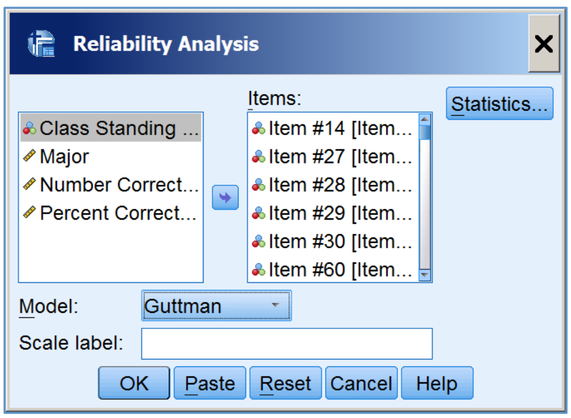
Reliability Statistics

Cronbach's Alpha	N of Items
.791	55

This indicates that the correlation between two randomly selected forms is estimated to be .791 for this group of students. Another way of interpreting this is that about 79% of the variance in scores is due to true differences and 21% is due to measurement error.

Another reliability estimate that is increasingly used is Guttman's lambda-2 (λ_2). $\lambda_2 \geq \alpha$. While α is an estimate of the correlation between scores on randomly-equivalent test forms, λ_2 is intended as an estimate of the correlation between scores on parallel test forms and thus will generally be higher than α (although it is still a lower-bound to the correlation between parallel scores).

Choose Guttman in the selection box next to Model.



Reliability Statistics

Lambda	1	.776
	2	.812
	3	.791
	4	.657
	5	.791
	6	.
N of Items		55

Of these Guttman indices, lambda 2 ($\lambda_{2)}$ or lambda 3 ($\lambda_{3)}$ is the estimate generally used. Guttman proposed λ_1 only as a starting point in his derivations, not as an index to actually be reported. λ_3 is identical to α. Here, the estimated correlation between parallel forms, based on lambda 2 ($\lambda_{2)}$, is .812. Equivalently, 81% of the variance is due to true scores and 19% is due to error (error is defined slightly differently for λ_2 than it was for α).

Interpreting reliability estimates

As stated above, reliability is an index of how well students can be differentiated by their scores. If the examinees don't really differ very much in their abilities, most of the differences in their scores will be due to error (low reliability). But assuming error variance remains constant, the same test might be able to reliably differentiate between examinees in a group that has a wider range of differences in ability. For example, the SAT is probably not very reliable for differentiating among Harvard students but it is very reliable in the population of college applicants as a whole.

There is no single guideline for interpreting reliability estimates but there are recommendations from the literature:

At least 0.70 for group-level interpretations, such as program evaluation, or research;

At least 0.80 for low-stakes individual decisions, such as instructional feedback or a small part of a course grade;

At least 0.90 if the scores are to be used for high-stakes decisions about individual students, such as admissions testing, promotion, or classifying into remedial and advanced courses.

Buckendahl, C. W., Impara, J. C., & Plake, B. S. (2002). District accountability without a state assessment: A proposed model. *Educational measurement: Issues and Practice,* 21 (4), 6-16.

Loo, R. (2001). Motivational orientations toward work: An evaluation of the Work Preference Inventory (Student form). *Measurement and Evaluation in Counseling and Development,* 34, 222-233.

Roid, G. H. (2006). Designing ability tests. In S. M. Downing & T. M. Haladyna, *Handbook of Test Development* (pp. 527-542). Mahwah, NJ: LEA.

Skorupski, W. P. (2008, August). A review and empirical comparison of approaches for improving the reliability of objective level scores. Available at http://www.ccsso.org/Documents/2008/Approaches_for_Improving_the_Reliability_2008.pdf