**Elements of Style in Data Analysis**

*Russell Almond*

0.0 Abstract

1.0 Why study style?

* Gaining understanding of the data
* Communicating to others.

1.1 The "audience."

1.2 Meta-data and data analysis.

2.0 The basic elements of style.

* Transcript documentation
* Graphic design
* Technical writing

2.1 Documentation and transcripts.

* + Use descriptive names and Variable Labels.
  + Label your cases, but protect your data.
  + [Use plenty of comments.](http://webmail2.fsu.edu/attach/labScoringGuide.html?sid=&mbox=INBOX&uid=34695&number=2&process=js&filename=labScoringGuide.html" \l "use comments)
  + Edit out the chaff.
  + Export your transcript.

2.2 Graphs and tables.

2.3 The writeup.

3.0 Grading key

* Correctness – 6 points
* Style – 4 points

**0.0 Abstract**

Data analysis is as much an art as it is a science. This document is an introduction to the elements of style in data analysis. An analysis done with good style has several advantages: it is easy to understand, reproduce and communicate. There are three important elements of style as used in these computer labs: (1) Commenting code, (2) Designing figures and graphics, and (3) Effectively communicating the results of the analysis.

**1.0 Why Study Style?**

The major factor separating the field of Statistics from the related field of Mathematics, is the vague and all-encompassing activity referred to as *data analysis*. Data analysis is the process of turning data gathered about a phenomenon into inferences---in terms of probabilistic models---about the observed behavior. Good data analysis involves many things: good scientific method, insight into the field being studied, a solid understanding of the appropriate statistical models, and a little bit of luck. Although founded on the mathematical discipline of theoretical statistics, data analysis as currently practiced is two parts art to one part science.

In this class, we study how to perform data analysis, at least for simple situations. Learning the mathematical techniques we cover in lectures and homework is comparable to learning to play the notes on a collection of musical instruments. Learning data analysis is like learning to play music tunefully and with expression; it is best learned by seeing examples and experimenting and practicing on your own. The labs provide this experience in a context which offers criticism of your technique, before you analyze data that really matter (your own data!).

Data analysis has two interrelated goals: (1) gaining your own understanding of the data, and (2) communicating the results of your analysis to others. The stylistic techniques presented here are meant to address one or both of those goals. Let us examine these goals one at a time.

1. Multiple steps to understanding, aided by comments. The task of gaining understanding of the data is a complicated process where you try one approach after another. Typically, you start with graphical techniques, like histograms, boxplots and scatterplots to get a feel for the scope and shape of the data. Then you might go on to more complex procedures like regression or analysis of variance. You might take some false steps along the way, for example, you might perform the whole analysis before realizing it would be better to analyze the logs of the data.

It is natural to borrow some of the stylistic features of computer science for our definition of style in statistics. Descriptive variable names, like MathScoreAtTime1, as opposed to cryptic abbreviations, like mst1, make your session transcripts easier to read and allow you to readily identify the variable (and recall whether it is important) a week or a year later when you are looking at your work again.

Similar remarks apply to comments. Don’t ever assume that you don’t need to make comments on a program or a data-analysis session because you will only need to do it once. Two months (or two years) later you may need to do a very similar task, and if properly commented, your old transcript will provide a good starting place for the analysis.

2. Text, figures, and tables. Equally important is communicating the results of your data analysis to somebody else. Two principle vehicles help to communicate results of analyses: written text, and figures and tables. The stylistic features here are borrowed from English composition and graphic design. Text should concisely and completely describe the analysis. Figures and tables should be clear and self explanatory; properly designed, they offer powerful tools for describing relationships and properties of data. Together the text, figures, and tables should directly lead your audience to the conclusions you reached without describing any false steps in the original analysis.

Selecting the right figures and tables to support your story is a critical part of technical writing. Well chosen figures and tables will tell your story or provide key evidence for it. Irrelevant figures and tables will distract from the story you are trying to tell. As a rule, *all figures and tables should be given a number and a descriptive caption* (for example, see Table 1), *and should be referenced in the text*. If you don’t have anything to say about it, why include it and distract from your story?

Table 1. Style points associated with figure and table choice and captions

|  |  |
| --- | --- |
| **Points** | **Problem** |
| -1 correctness point | Not including a table or figure necessary to support argument (e.g., no scatterplot showing linear relationship in a regression analysis) |
| -1 style point | Unnecessary tables or figures included, e.g., SPSS output unrelated to research question (e.g., table giving variable names and sample sizes) |
| -1/2 style point | No captions on figures or table |

**1.1 The "audience."**

In writing your report, as in any writing, you must know your audience. We will assume that the "audience" of your report will be a colleague[[1]](#footnote-2) in a related discipline. This individual is bright and busy; she is technically competent but has been out of school for about 10 years. She had a statistics course long ago and remembers most of the concepts, but has forgotten much of the jargon (so don’t use it!). She is current and up to date in the background knowledge of the general field of the project. Finally, she is really busy and will toss your paper aside if things are missing, incomplete or contradictory. She is very similar to the "consulting client" of the professional statistician, or a reviewer for a journal or conference paper. This individual is one of the two most important classes of audience for data analysis write-ups, the other being the "general public." Reports for general audiences are more difficult and are not addressed in this class.

Because your colleague is busy, the report must be brief. Similarly, most journals and conference proceedings have page limits for submissions. However, there is a fundamental tension between brevity and completeness. Obviously, your report must be long enough to be complete, but at the same time be short enough for your audience to read it quickly. It takes considerable skill to balance the two goals, so you may not get this right until the third or fourth lab.

To illustrate the desired level of detail, consider a description of how the U.S. Gross National Product changed throughout the twentieth century. We do not need to explain every little wiggle, but must describe important features, like exponential rises and the 1929 and 2008 depressions.

Another concern when writing for a "colleague in a related field" as your audience is the level of technical detail. Writing at the right technical level is to some degree a matter of experience. On the one hand, you don't want to “snow” your reader with technical terms, and on the other, you don't want to substitute wordiness for well designed technical terms. A further difficulty is that as the complexity of the labs (and your own later analyses) increases, the amount of statistical sophistication you should assume from your reader also must increase. It takes a reasonably sophisticated reader to understand the implications of an analysis of variance, while a relatively naive reader should be capable of reading the results of your exploratory data analysis.

The following terms (and similar terms at about the same level) are the limit of what you should assume your audience understands for Lab 1: *mean, median, mode, range, standard deviation, variance, outlier, skew, long (short) tails, normal distribution, probability, confidence interval,* and *significant.* Other terms will be appropriate depending on the problem. For example, if the statement of a later lab problem says to perform a *t* test, you can assume that your audience knows what a *t* test is. In some cases, your audience can learn from your write-up. For example: "The Q-Q plot in Figure 4 – a graphical test of normality – shows that the residuals are approximately normally distributed satisfying the assumptions of the *t* test."

The best advice for avoiding jargon is to write as much as possible "in your own words." Being less familiar with the jargon, you will not naturally use it unless you are deliberately trying to imitate me or the textbook. If you find yourself using too much jargon, try to write about the *implications* of what you see, rather than trying to directly *describe* what you see. If you don't see a natural way to avoid the jargon, you may be describing a statistical concept that cannot be easily paraphrased. In such cases, you may use the jargon, but make sure its meaning is clear from context. (And make sure the meaning is clear to you! Ask about it in class, in the discussion groups or in office hours if you are confused about a statistical term you need for your lab writeup.)

Incidentally, the instructors and the TA's who will help us provide feedback on your labs closely resemble the fictitious busy colleague who is reading your report. We want to be able to quickly scan your report and assess your understanding and presentation of the material while not being overwhelmed with details. We will appreciate your brevity. Generally, the written portion of the main body of your lab should not exceed four double spaced pages. Reports exceeding this limit by more than a paragraph will be marked down, unless the additional material is *really* worth reading.

**1.2 Meta-data and data analysis.**

Statisticians have lately begun talking about *Meta-data*---data about the data. For example the meta-data describe where the data[[2]](#footnote-3) come from, how they were gathered, the units of the measurements, whether variables are discrete or continuous, what was the sampling mechanism, how reliable the measurements are, etc..

A lot of important information is buried in meta-data. For example if data were gathered using a biased sampling scheme, we would not believe the results as much as if we knew they came from a randomized sampling mechanism. (For this reason, we should cite sources of the data when available.) Similarly, it is hard to draw conclusions if we don't clearly understand what the data are measuring.

For these reasons, your report should always contain a summary of the meta-data. This should be about a paragraph (unless there is something of particular note, such as a very bad sampling scheme or a problem with missing data). It should summarize the important meta-data so that your conclusions are obvious. It should also include a citation for the source of the data when that is known.

In a paper formatted according to the American Psychological Association (APA) style guidelines this information goes in one of two places. If you gathered the data yourself, you would include a Methods section that describes the procedure for collecting the data. You also normally include a Measures section that describes the individual variables. In the lab assignments for this class, the methods come from somebody else’s research, so we cite the author and briefly summarize their work. However, the variable definitions are critically important. If we don’t know what the variables are and what they mean, how can we interpret the statistics and graphs?

**2.0 The Basic Elements of Style.**

From the above discussion, we can identify three elements of style: (1) transcript documentation, (2) graphic design and (3) technical writing. Specific approaches to these elements are described in Sections 2.1, 2.2 and 2.3, respectively. The hints described here are meant to correspond to several different levels of analysis.

You will want to save this handout and refer back to it as you do each new lab. For that matter, it will be worth re-reading the text of previous lab assignments anew when you start each new lab. Sometimes the assignment handout describes techniques that are more general than are needed to solve the problem of that particular lab. This is because the labs anticipate techniques and hints that will be used later.

**2.1 Documentation and transcripts.**

Although modern versions of SPSS feature a graphical user interface for helping you to figure out all of the details of your analysis, it is built upon a command line interface that was the core of the earlier versions. Each time you press the "OK" button on a dialog, the SPSS interface generates the corresponding command and sends it to the core engine.

One consequence of this architecture is the SPSS generates a transcript of every analysis you do as you do it. Into this transcript it inserts both the commands (If you don't see the commands, check the settings of your options in the viewer tab of the options dialog) and the output. These logs are an incredibly valuable tool, and you should become familiar with them and how to use them for your advantage.

Any significant project you work on (like your thesis or dissertation) will likely span months, if not years. You will likely do initial analyses of the data and discuss the results among the research group. You will come back and revisit those materials several weeks later when you write a report or a paper. Some of the analyses from that time can be used directly, but you will often think about new analyses and displays you should do while you are writing up the results section. If you submit the paper to a journal, the reviewers will likely call for additional analyses. At that point, six months will have elapsed since you last looked at the data. Unless you took *really* good notes, you will have to “scramble” (work hard) to recreate your earlier work from scratch.

Here are some specific hints that will both help you when asking advice from colleagues and will help you jog your own memory when returning to your analysis. (These hints were written with SPSS in mind; if you are using some other stat package the details will be different but the principles will be the same.)

1. *Use descriptive variable names and Variable Labels.* The reason for this is obvious. If the variable name is reasonable, you can remember it, even after you leave the project and return to it. Use standard words, or reasonable abbreviations. To separate words you can use a separating character like the period (.), underscore (\_), or you can use capitalization (TotalScore). Thus math.year1, math.year2, rdg.year1, and rdg.year2 are two math-test scores and two reading-test scores, taken in years 1 and 2 of a study.

SPSS variable names must correspond to the rules of computer variable names: They must be unique, and certain characters (like the space) are prohibited. While a variable name like Doll may be readily understood to the analyst, a more casual reader of the report may not know what it means.

To get around the restrictions on variable names, SPSS provides a "Variable Label" which allows you to provide a longer more descriptive name for the variable. (You can see the variable labels by selecting View > Variables.) Here you can type in a longer description, say Dollars spent (in thousands) for the cryptic variable name Doll. When SPSS has variable labels, it will use them in place of the variable names in tables and graph (see the next two hints). *Nota bene! Failing to use Variable Labels is a very common way to lose style points in a lab!*

1. *Use descriptive names for the rows of data matrices.* SPSS does not provide built-in support for row names, so what you need to do is to define a column containing a string variable with the row names or case IDs. You can then select this variable as the label for a number of different plotting functions. This helps when you want to select certain points (e.g., outliers) to label on a graph.

Even if you are simply using numbers as labels, it is still good practice to create a string variable for storing IDs. If you later create a derived data set by dropping some cases, or sorting the data to merge it back with the original, you could create a big mess that may be impossible to clean. (This is also a good reason to save copies of your files with new names if you plan to attempt this kind of operation.) Also, even if the IDs are numbers, it is good practice to add a letter to the front of the ID to make it a string, e.g., S0043 for student 43. This is especially important as many programs (beware of Microsoft Excel®) will strip leading zeros from ID numbers, making it harder to match them up with files from other programs. Also, it is hard to mistake a string ID number (which is really a label) for a number during analysis.

Do **NOT** have the names of your participants as part of your working files, nor other identifiers which would reveal the identity of the subjects, such as Social Security Number, addresses, phone numbers, or FSU IDs. This is especially true if the data file contains sensitive information such as grades. In general, contact information for the subjects should be kept in a separate administrative file and protected with a password. Remember, you don't want to write hundreds of letters of apology because you forgot your laptop at Starbucks and all of your participants’ personal information was available to potential identity thieves.

1. *Use plenty of comments.* SPSS allows you to insert titles into your transcript. It also allows you to import a big block of text (say the body of an email from a colleague explaining the coding scheme used in the data). It is difficult to "over-comment" your code (it is mostly a problem of your time, not style). Also leave plenty of white space (blank lines). A well written program or session transcript should be easily read by anyone, even someone with little knowledge of the specific computer language. The following list is a number of places you *should consider using comments.*
2. *New task or sub-task.* Whenever you start analyzing a new data set in an analysis comment it, usually with lots of white space. This might also be a good place to insert a page break.
3. *Similarly you should* comment landmarks in your analysis, such as "Exploratory Phase," "Model fitting" and "diagnostic checks." Any of those steps might have many parts, for example the latter might consist of "normality tests" and ["heteroscedasicity check"](http://webmail2.fsu.edu/attach/labScoringGuide.html?sid=&mbox=INBOX&uid=34695&number=2&process=js&filename=labScoringGuide.html" \l "jargon.fn)[[3]](#footnote-4). Successive layers of sub-task get smaller comments, fewer flags and less white space. In a well written transcript, if you want to check a particular formula, your landmark comments should quickly guide you to the formula you want to check.
4. *Interesting observations.* When you detect an "interesting" feature of the data, comment on it! These comments have two purposes: (1) A well placed series of observational comments in your transcript will aid you when you write your report; you can change the comments into a more narrative form and the rest will follow. In particular, it is helpful to comment the commands you used to create figures for your report. This will enable you to find them quickly (and later go back and clean them up; see the next section). (2) They remind you of analyses you may want to do later. For example, if you see an outlier, you may want to check its influence on model conclusions after you have fit the model. Observational comments help remind you about those plans. Furthermore, keep in mind the context of the problem, and note any features that support the goals of the study, or might challenge the intended interpretation.
5. *New variables.* Unless exceedingly obvious, any new variable deserves a comment. Why did you create or recode a variable (e.g., why did you subtract 15.4 from all of your values? Is it the mean, and you have centered?) ? Furthermore, if the new variable represents data, this is a good place to include the "meta-data" for that variable, *e.g.,* units, source, or any other similar information of note.
6. *Cryptic code.* If you write cryptic code (something to avoid), then a comment or two will help you recall why you did what you did. A better solution would be to simply avoid cryptic code. However, for instance, you may create a special variable to select a particular subset for some data exploration. This deserves a comment, as it will be difficult to recall and interpret later.
7. *Edit out the chaff*. A typical data-analysis session generates a lot of stupid mistakes: typos, places where you forgot the name of an argument to a function, or to select an option you want, and sometimes really stupid analyses (e.g., making a frequency table of all of your participant IDs). These are the sorts of mistakes *my* sessions are full of, and if you are a typical data analyst, yours will be as well. You should remove these mistakes and false trails, making your session easier to read, preferably, while you are computing, or as soon as you are finished the session, while it is still fresh in your mind. SPSS allows you to delete tables and even entire analyses from your transcript.
8. *Export your transcript as a Word or RTF file*. This has two important advantages. One is, you can edit the transcript from a computer that doesn't have SPSS installed (e.g., your home computer). Second, you will often want to copy tables and graphs from the transcript into the report. Having a Word or RTF version will allow you to do this from a computer without the proper version of SPSS.

Previously, we have required that the actual SPSS lab transcripts be turned in for credit. This will not be required this year (it produced many problems). However, we suggest that if you have questions about whether or not you have done the computing part of the lab correctly, that you take your transcript to the TAs for feedback.

**2.2 Graphs and tables.**

Your use of graphs and tables will “make or break” your report. The field of graphic design, and the subfield of statistical graphic design, is quite complex, and we won't go into all of the issues in this class. We will try and cover some of the basics in class and in the lectures.

Often the results of a complicated analysis (like an analysis of variance) can be clearly shown by a simple picture (e.g., a number of side by side boxplots or error bars). In fact, statisticians serving as expert witnesses may do a complicated analysis to feel confident that they can see exactly what is going on, but then look for a series of simple pictures, that can easily be understood by judge and jury, to express what they have discovered. You don't need to go quite that far for these exercises, but your tables and graphs are a critically important part of your presentation. The purpose of the text is almost simply to show the reader what to look for in the graphs; the graphs will make your argument for you.

One important feature of SPSS is that you can double click on tables and graphs and then edit them. In particular, you can change the terse default labels (often just the name of a variable or case) and replace it with a more complete name. If you take the time to define labels for your variable, SPSS will use them instead of the short and cryptic variable names. Also, check out all of the options when making a plot or table. Often one or more of them will help the data tell the story you want to tell.

Two important resources can help you make the plot you want, rather than the plot that SPSS wants to make for you! We will cover some things in class, but you may find you need to change plots in ways we did not discuss. First, look at the online help. That will explain the options available for any given plot type. Second, search the Internet. Googling "SPSS histogram with density" may get you to somebody's tutorial on how to do exactly what you want to do.

*Rules for Titles on Plots*

1. The *main title* should briefly describe the contents and should contain the figure number (you could also write that in by hand, or type it on later).
2. The *sub title*, if necessary, should contain extra information, like a data subset or the values of parameters used in a simulation.
3. The *axis labels* should contain a plain English description of the contents of the axis, including units (if appropriate). If you set variable labels in SPSS, these will often appear in the graphs.

If you are using different plotting symbols or line types for different groups, you will also need a legend. Again, the labels here should be understandable by your colleague who has never seen your data collection sheets, and it can’t be just a list of your variable names.

Often you need to put multiple plots together to make a figure. This is especially true if you are comparing multiple conditions. If the point is to compare the distribution of the variable on the natural and log scale, you need to shrink the two histogram so they can be seen on the same page to they can easily be compared. SPSS plots can be copied and pasted into a word processor, and then resized to make a natural layout. Be careful, if the plot becomes too small, the titles, labels and legends will be illegible. You will need to edit the plot to increase the font size.

Ideally, plots should be integrated into the text. This, however, can be difficult (e.g., a plot may cross a page, so not be legible) and dependent on the system you use to type your report, therefore you won't be required to do it. As an alternative, I suggest a section of figures at the end (many journals still require this). If you do this, each figure should be clearly numbered so that it can be referenced in the text. Each figure should convey useful and new information; don't overwhelm your reader with too many repetitive (or worse yet, useless) graphs. And definitely do not let your graphs be broken across two pages (or the graph get separated from its caption)!!

*Tables*

If you find yourself putting a lot of numbers into the text, you may find that they are better organized in a table. A table makes the numbers easy to find and convey. There are some situations, like analysis of variance, where tables are absolutely necessary. Tables, like figures, can either be integrated into the body of the text, or put in a clearly labeled appendix. A rule of thumb for display: for one to three numbers, use prose, for four to twenty numbers, use a table, for more than twenty numbers, use a graph. Whatever you do, do not repeat all of the numbers you have put into a table in your text.

*Significant digits.* When building tables it is important to think about how many decimal places of accuracy you really need. If you are measuring money spent in thousands of dollars, are differences of pennies really important? Reporting the amount to six decimal places will obscure differences that are really important. Generally speaking, your numbers will have only three or four meaningful digits. Round the numbers after that. SPSS will NOT do this for you!!

SPSS tables often have far too many digits and need to be edited. SPSS will allow you to change the number of digits (right click on the number), but it is quite laborious. You can also edit the table in your word processor, but again you are working one entry at a time. Two possible solutions: (1) Paste the table into a spreadsheet program (e.g., Excel) and edit it there. Spreadsheets have much better controls for formatting tables, including adjusting the number of digits on a column-by-column basis. (2) Retype the table into your word processor. While this seems like a lot of work, often much of the information in the SPSS table is just not important and you can do a much better job of formatting if you start from scratch with just the information you want.

Unless the table is very long, it should all fit on the same page. The headers of the rows and columns provide context for the information in the table. It is a problem if that context is lost on the previous page. Beware; word processers (in particular, Microsoft Word®) break tables in the middle, and separate tables from their captions. You may need to move text around or insert a page break to force the table onto a single page.

*Nota bene!* Pasting SPSS tables into your final report without thinking about appropriate columns or number of digits or format is a very good way to lose style points. SPSS tables do NOT follow APA rules (e.g., they have vertical lines which APA tables do not have). So if you do use the SPSS table content, modify the format to follow the rules.

*Figure and Table Numbering*

Most journal styles (in particular, American Psychological Association, or APA, style) use consecutively numbered tables and figures. Tables and figures are given separate sequences of numbers (Tables 1, 2, 3,... and Figures 1, 2, 3, ...). The usual names for these sequences are “Figure X” and “Table X”, respectively. Also, for an article, simple integer numbers are used (occasionally with 'a' and 'b' letters for sub-figures), although books sometimes use two part numbers (e.g., Figure 3.4 is the fourth figure in Chapter 3). When "Figure" or "Table" is followed by a number, the first letter should be capitalized. The space between the name and the number should be a non-breaking space (this is produced by pressing Option-space in Word and many other word processors). This prevents the name from being separated from the number when the paragraph is broken into lines.

The caption for the figure or table should consist of the type of object (Figure or Table), the number, and a brief phrase describing what it there. This is important; many readers will skim the figures and tables and they shouldn't need to search long to find the descriptions. The caption for a table is usually placed above the table and the caption for a figure is usually placed below the figure (don't ask me why), but some publishers may have different rules.

In addition to having a number and a clear caption, each figure or table should be referenced somewhere in the text. If it is not important, don't waste your reader’s time with it. If it is important, tell your reader what they should be looking for. *Nota bene! Including useless tables and graphs from SPSS output is a common way to lose style points in a lab.*

**2.3 The writeup.**

Just as the principles governing style in analysis and transcripts are borrowed from computer science, so the principles of style in writing are borrowed from English composition. All the normal rules apply. Technical writing is not easy; these labs will provide a good chance to practice.

Here are several things to bear in mind when writing your final report:

1. *Type the report.* You *do* have access to computer accounts which can be used to type the write up, or you can do it any other way that is convenient for you.
2. *Keep it concise.* You shouldn't need to exceed four pages of double spaced written text, plus more pages for tables and figures. (Your total page count will be higher if you intersperse tables and figures).
3. *Organize your information.* It may help to write a paragraph outline. Then make a list of all observations you need to note in the write-up and find the best home for them. The specific questions of the lab write-up may help you organize your thoughts.
4. *One paragraph per idea; one idea per paragraph.* One problem that we frequently see is that the introduction, hypothesis, and data description are all jumbled up in one paragraph. Not only does that make it harder to read, but it makes the paper harder to grade. We may miss the fact that you actually performed a critical step that was part of the work, because it was mixed up with other stuff.
5. *Avoid passive voice, first and second person when possible.* APA style prefers first person to “the researcher”, but allows first person only when the object of the pronoun is clear. Thus, “we” is okay when in means researchers, but not when it refers to the writer and the reader (some of your reports will have only one author, then you should use “I”). Consider the following sentence, “We [or I] performed an analysis of variance.” Rewriting it into passive voice, “An analysis of variance was performed.” is no better. The key is to make the ANOVA the subject; thus, “An analysis of variance found that ….” Similarly, avoid referring to the reader in the second person (as we do in this paper). For example, instead of “You can clearly see the relationship between the variables in Figure X”, write “Figure X shows the relationship between the variables.” Note that we dropped the “clearly”; this word is unnecessary and insults the reader if they don’t see the relationship you claimed was clear.
6. *Proofread your work.* This will help you spot typos, editing mistakes, and awkward language. Good writing usually takes many drafts. If you have difficulty writing in English (I'm not just talking about international students here; some of the worst offenders were nominally native English speakers); get a friend to proof read your work for you. Also pay attention to your word processor’s red underlining for misspelled words. It does not recognize some of our technical terms, but if you run two words together, or misspell something it recognizes, it will underline that text so it can be fixed.

We are strongly suggesting that you use American Psychological Association (APA) style when writing your report. Our TA from last year Jiwon Nam kindly prepared a brief guide to APA style for use in the lab writeups (posted on the blackboard site) and Dr. Becker has prepared a general set of writing hints. APA style is reasonable similar to other common styles, such as that of the American Statistical Association (ASA) or the American Medical Association (AMA). All of these styles have a similar purpose: organizing the information in such a way that authors can be sure they have included all of the necessary parts and that readers can readily find the information.

Here is a sample outline to help you with your planning. This outline is for a general data analysis, not all parts will be applicable to all labs:

1. *Introduction.* What is the purpose of the lab? What will the reader see in the report?
2. *Background.* A typical research paper would have a background section reviewing related research. This is not necessary for the labs you can skip this part.
3. *Research question and hypothesis.* Although this is often two sections in APA papers, the information is somewhat redundant and can be collapsed in the labs. What is the scientific question we are trying to answer? What do we expect to see? The hypothesis is the scientific hypothesis and should draw upon your background knowledge of the things in the experiment or study.
4. *Data description & Measures.* Briefly describe enough of the meta-data for the reader to be able to follow the analysis and judge the soundness of the data and methods of analysis. If the source of the data is a scientific paper, or web site, you should cite it here.
5. *Results and analysis.* This is the part of the general research paper outline that is the most important for the labs. It should address the following topics in roughly the order shown (although not all of these topics are important in each lab):
   1. *Exploratory data analysis*. Describe the main features of the data, and any of special note (e.g.*,* outliers, skewness, heavy tails, bimodality and other peculiarities).
   2. *Model fitting*. If you fit the data to a probabilistic model (e.g., regression, analysis of variance), give the estimated values of the parameters and their standard errors. If you did a formal hypothesis test you will need to report, the test statistic (e.g., *t* or *F*) the number of degrees of freedom (two for *F*-tests) and the *p*-value. You should also report some measure of the size of the effect. This is where you first mention whether or not the difference is statistically significant.[[4]](#footnote-5)
   3. *Diagnostics*. If you fit a model, you should check the assumptions of the model fitting process and report on how well the data fit the model. Report about the distribution shape of the residuals here (if that is appropriate for your lab).
   4. *Alternate models*. Sometimes you may want to re-run the analysis with a variation, *e.g.,* after removing an outlier or taking logs. In those situations, report any differences between the original analysis and the modified one (this will be a very brief report if there is not much difference).
   5. *Predictions.* The regression lab will call for you to make predictions about certain individuals; those also must be reported in the results and analysis section.

6.  *Conclusions*. What do you conclude? What are the implications of what you saw? How does what you observed relate to the goals of the study? These questions and a host of similar ones should be answered. The goal of the conclusions section is to relate the statistical results to the goal of the study. The facts that the *t*-statistic reveals that the difference is significant and that the *p*-value is less than .01 are not conclusions; they are evidence to support your conclusions. A conclusion is something like "The results suggest that students using the experimental curriculum learned significantly (p <.01) more about fossils than students in the business as usual control.” The conclusion should not introduce new results, but instead should restate the important results from the results section. It should also tie those results back to how they relate to the broader research questions.

1. *Follow-ups.* Does the analysis suggest further experiments? If so, what are they? Is a larger sample size needed to tell what is going on? Do limitations of the sampling procedure limit the generalizability of the results?
2. *Bibliography.* If you have formal citations related to the source of the data (or anything else), put them in a separate section at the end.

9. *Formatting*. You do not need to use formal section headings (although that can help you organize your report), but each of these topics (if applicable) deserves a separate paragraph.

**3.0 Grading Key.**

Labs are graded with 6 points out of 10 for correctness and 4 out of 10 for style. This means that even if you get everything perfectly correct, you’re only doing 60% of the work if it is not written up well. In practice, these aspects of writing are tightly correlated: good style will help us easily realize that you have all parts of the analysis done correctly; poor style may cause us to overlook some of your hard work, or cause a journal referee to refuse to even read your manuscript!. The stylistic hints contained here should help you get more value out of your labs, and your later work. Also be aware: 10 out of 10 is a rare occurrence.

1. The 6 points for correctness vary from lab to lab. They generally correspond to steps in the analysis you need to have done correctly, and to key observations needed in your writeup. These are generally additive: you earn more points by placing needed things in your lab, although don’t just write everything you can think of – if you say something that is clearly wrong you may get points taken away.
2. The 4 points for style work the other way. Roughly 2 points are given for the text and organization and 2 points are given for the figures and tables. You start with the four points and the graders will take away points (or half-points in less severe cases) based on the following kinds of problems:
3. Minor organizational issues (-1 pt). The document is mostly organized coherently but with a few problems (e.g., introducing new results in the conclusion section, or burying the problem statement in the introductory paragraph).
4. Major organizational issues (-2 pts). Document is missing key elements, like introduction, research question, or conclusions, or elements are so jumbled together that the graders cannot find them. (This is generally cumulative with the previous issue.)
5. Problems with tables and figures (-1 pt). Tables and figures lack captions or captions are unclear. Graphs use raw SPSS variable names as axis labels. Too many significant digits in tables. Generally one of these problems will cost a half point, and multiple problems will cost a full point.
6. Unclear wording (-1 pt). This is not just for poor grammar or minor usage problems, but wording that is confusing to the point that the reader can’t understand what you are trying to convey (or where you are obviously stating something different than intended). Minor problems in grammar, usage and mechanics will not result in points being deducted, although the graders will note problems when giving you feedback. If, however, the number of mechanical problems is so great that it becomes difficult to follow your paper, you will get points taken off.
7. Mismatch between text and figures (-2 pts). Lots of tables or figures have no explanation or are not clearly tied to the story the paper is trying to tell. Information in tables or graphs is misinterpreted. *Remember, every table or graph should be referenced in the text.* If the table or graph doesn't have anything important to say, why are you wasting the reader’s time with it?

Your style score cannot go below zero no matter how bad the report. However, if you are earning zero style points, it is quite likely that you are losing correctness points because we cannot see whether or not you have done the lab properly.

**4.0 Collaborative Writing.**

As collaborative projects are the norm in science, learning how to write as part of a team is a critical skill. This is important as different team members have different strengths. Maybe one of you is good at working with SPSS, another understands the statistical terms really well and another is good at producing text. However throughout the term we want each of you to develop all of these skills. If the collaboration is done properly you can draw on all of those strengths to do good work. This section talks about some important features of collaborative writing.

**4.1 Defining clear roles and expectations.**

A critical part of your team collaboration will be establishing the roles for the team members. Who will do what? When are things due? When are you meeting and where (including online and telephone meetings)? Two things which can cause problems are misunderstandings about who is doing what (“I thought you were doing the graphs.” “I thought you were!”) and about when it is needed (“I needed that yesterday!”). Also, if things come up during your work (as they inevitably will) make sure that you *inform your teammates* about the problems so that they can make adjustments.

*If you work on the projects collaboratively, you will be asked to fill out a questionnaire about how your teammates contributed to the process.* Experience has taught us that students deal harshly with other students who do not meet their commitments. So make sure that everybody is clear in the expectations and that you meet your commitments.

**4.2 Role of the lead author.**

The lead or corresponding author is almost always the first author in the paper. (For the 5400 labs, we are stating that the Lead Author is the first one; journals have other was of marking the corresponding author.) This is the author who communicates with the editor of the journal, or in our case, submits the paper on Blackboard. It is also the author who is responsible for gathering the feedback from the reviewers and passing it back to the team members for deciding how to deal with the issues that are raised.

The lead author is almost always the primary “editor.” This is the person who assembles the contributions from the other team members into the final document, does the final edits, and chooses between alternatives suggested by the other authors. It is always worthwhile for the lead author to do a final reading of the paper before turning it in, making sure that there are no obvious mistakes and everything is included that should be.

Because this role is so important, *we are requiring that students be lead author on at least one lab.* Make sure you plan your teams accordingly.

**4.3 Role of the contributing authors.**

The contributing authors, the authors other than the lead author, also have important roles to play. If you are a contributing author, you are signing your name to the work. Therefore it behooves you to check to make sure that it is correct and corresponds to all of the style rules.

Lead authors should always circulate the penultimate draft to all of the contributing authors, who should then “sign off” on it (say it is acceptable) or note minor changes. If major changes are required, it is probably best to insist on another cycle of internal reviews before the final version is turned in. This all takes time, so make sure you build it into your workplan.

Last, in the end *all* contributing authors and the lead author are responsible for the lab writeup (or any other paper). In one of Dr. Becker’s courses a “lead author” on a team project cut-and-pasted text from a journal without citing it. This was identified as plagiarism. Dr. Becker called in all of the paper’s authors to interview them about the text. It came to light that one member had been responsible for the text and the other members had not examined the text carefully. So they were all, technically, guilty of plagiarism. Similarly, if a team member takes on the task of preparing the tables, all members must check the tables for format and accuracy. This is what being a “good collaborator” involves.

**5.0 Final Thoughts.**

Grading for style is an evolving concept, and this grading key is subject to change as the course progresses, and the instructors learn more about grading on style. I would also be happy to hear your thoughts on the subject of grading on style (although not the reason why *you* should have gotten 4 pts instead of 3!). Send me email at ralmond@fsu.edu or drop by my office and let me know your thoughts.

Writing effectively, especially when data analyses are involved, will be a critical skill no matter your future path in life. Many of you have a dissertation or thesis in your immediate future. If you choose the path of a researcher, you will need this skill when you write articles for journals. If you choose the life of a practitioner, you will find yourself writing reports for managers, clients, or sponsors. The lab writeup gives you a chance to practice this critical skill.

**6.0 Acknowledgements.**

Dr. Becker kindly translated an earlier versionof this report done in HTML into Word. In the process she corrected a number of errors and clarified a number of places where Dr. Almond used jargon. The moral is that having a colleague read your paper can really help improve it!

1. In a previous course I taught, I used a fictitious "manager" as the audience. Actually, it wouldn't be too bad to think of explaining things to Dilbert's pointy-haired boss, or to a well intentioned but naïve school-board member. Your advisor and reviewers aren't usually aren't quite that stupid, but if the report isn't crystal clear, they usually don't have time to figure out the details. [↑](#footnote-ref-2)
2. "Data" is a plural noun; its singular is "datum." Thus, one properly writes "The data *are* ...". The automatic e-rater® grammar check supplied with TurnItIn® gets this wrong. [↑](#footnote-ref-3)
3. The principle of avoiding jargon applies only to the write-up, not the transcript. The transcript will only be read by you, or somebody else who has or desires a good working knowledge of statistics. [↑](#footnote-ref-4)
4. The word significant has a special meaning in statistics which will be discussed ad nauseum in class. Unless you are using it in the context of a formal statistical test, you should avoid it. [↑](#footnote-ref-5)