On Teaching Statistical Practice: From Novice to Expert

Joel B. Greenhouse and Howard J Seltman[1]

1. Introduction

Most would agree that the statistical sciences are experiencing a tremendous surge in popularity. Current excitement about statistics and data analysis is due in part to our ability to generate, manage and use massive amounts of data (Big Data) for scientific discovery and for making predictions about future events. The media have been generous with Big Data success stories, e.g., *Moneyball* (the book and movie) and Nate Silver's FiveThirtyEight website, and industries from health care to retail are recognizing the opportunities of using predictive analytics to personalize patient and consumer experiences. Students are flocking to our classes, and the demand for well-trained statisticians continues to grow. As a result, increased attention is being paid to the development of curricula and training programs that adequately prepare students for the modern work place. See, for example, the ASA's Workgroup on Master's Degrees (Bailer et al. 2013), this journal's excellent Special Issue on Statistics and the Undergraduate Curriculum (2015), and the guest editorial in that issue by Nick Horton

[1] **Authors' Affiliation and Acknowledgments**
Joel B. Greenhouse is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (Email: *joel@stat.cmu.edu*). Howard J. Seltman is Senior Research Statistician, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (Email: *hseltman@stat.cmu.edu*). Parts of this paper were presented by the first author at the inaugural Rod Little Distinguished Lecture in the Department of Biostatistics at the University of Michigan, September 2016. The authors thank Marsha C Lovett, Director of the Eberly Center for Teaching Excellence & Educational Innovation, Carnegie Mellon University for her helpful suggestions and comments on this article. We also acknowledge with gratitude the talented, dedicated and good looking students who participated in the MSP program.

and Johanna Hardin (2015). Although discussions about *what* students should be learning are thoughtful and engaging, less attention has been paid to *how* the teaching of statistical thinking and practice can be improved.

How learning can be improved has been the focus of research in cognitive science for the past 30 years. Cognitive science is concerned with identifying and understanding the basic phenomena related to learning, for example, skill acquisition and problem solving, and developing precise theories that explain and predict these phenomena. These theories specify a fixed set of general mechanisms designed to explain learning and performance across a broad range of situations. Application of these theories to education, an emerging field called the Learning Sciences, has helped guide and improve both curricular development and course instruction. An illustration of how several principles of learning were put into practice in the design of an introductory statistics class can be found in Lovett and Greenhouse (2000). Excellent descriptions and summaries of the research on learning and its applications in education can be found in the National Academy of Sciences' report *How People Learn: Brain, Mind, Experience, and School* (Bransford, Brown, Cocking, (editors) 1999), and *How Learning Works: 7 Research-Based Principles for Smart Teaching* (Ambrose, Bridges, DiPietro, Lovett, Norman 2010). Much of the material for this article is drawn from these sources and the references therein.

The purpose of this article is three-fold. First, we review principles of learning derived from cognitive science that help explain how learning works. Second, we adapt these principles to the challenge of teaching good statistical practice; and third we illustrate the use of these principles in the context of the planning and curricular design of a new

graduate program in applied statistics, emphasizing how these principles can be used not only to improve instruction at the course level but also at the program level.

At the outset, to help provide a context for this discussion, we briefly describe the design and pedagogical considerations of our applied master's program. In 2008, the Department of Statistics at Carnegie Mellon University began planning a new professional master's degree program. Our motivation was based, in part, on an awareness of a growing workforce need for statisticians who not only could analyze and manage real data but could also communicate effectively with clients and managers (See, e.g., Committee on Enhancing the Master's Degree in the Natural Sciences, 2008). In addition, a growing number of our own undergraduate students, though well trained, were reporting not feeling "ready" to enter the job market with just a bachelor's degree. Our goal was to develop a two-semester, professional master's degree program for about 20-25 students per year that emphasized statistical practice, computing, data management and practical workplace skills. To underscore the focus of the program we named it and the degree the Master's of Statistical Practice (MSP).[2] We enrolled our first class in fall 2009. As noted above, a goal of this article is to illustrate the application of research findings about how learning works to the design of and pedagogical approach to the MSP program.

[2] Readers interested in learning more about the logistical details of the MSP program are referred to the Appendix.

Table 1. Sample MSP Curriculum

Fall	**Spring**
Perspectives in Data Science I	*Perspectives in Data Science II*
Applied Linear Models	*Experimental Design & Time Series*
Stat Methods in Epidemiology	*Data Mining*
Graphics & Visualization (mini 1)	*Consulting*
Multi-level Hierarchical Models (mini 2)	

Based on feedback from alumni and employers we felt we knew what the content of the program should include (see, for example, Table 1)[3]. There would be foundational courses such as, Applied Linear Models, Experimental Design and Time Series; modern topics like Data Mining, and Graphics and Visualization; special topics, such as, Multi-level Hierarchical Models, and Statistical Methods in Epidemiology (i.e., a course in methods for risk assessment); computing including data management; and a capstone, investigator-initiated consulting project. We also heard loud and clear the importance of both written and oral communication skills. To address the need for professional and workplace skills, we developed a two-semester course called *Perspectives in Data Science* organized around rotating 3-4 week modules that included computing skills, e.g., R, SAS, SQL and an introduction to Big Data methods; communication skills, including oral and written; career counseling and interviewing skills; professional and

[3] Not surprisingly these topics align nicely with the recent recommendations of the ASA Workgroup on Master's Degrees (Bailer et al 2013).

research ethics; and assorted other topics. There would not be a separate theory course, but rather, theory would be taught as needed in context in the methods courses.

Even though it seemed clear what we needed to teach, the pedagogical challenge was how to teach all of these topics so that students would actually learn the material and be job ready. In the following sections we describe principles of learning from cognitive science that helped inform our pedagogical approach, and how these principles have helped guide us in developing a curriculum for teaching statistical practice. The paper is organized as follows. In the next section, we discuss the cognitive science perspective on how learning works and introduce evidence-based principles that have been shown to facilitate learning. Next, in section 3, we adapt and apply these principles to the teaching of statistical practice. In section 4, we report the results of a nonscientific survey of our alumni soliciting feedback on how the MSP program helped prepare them for their jobs, and conclude in section 5 with a discussion.

2. How Learning Works: An Introduction

Since we wanted our students to develop a high degree of competency in the practice of statistics and statistical thinking, it was not only necessary to identify the elements that characterize good statistical practice but also how to teach it. We turned to cognitive science research to help us develop a curriculum where “novice” students would become more like “expert” data analysts. An expert is someone who has a high degree of proficiency, skill, and knowledge in a particular subject. Research in cognitive science has shown that experts differ from novices in a number of ways, especially with

respect to their approach to problem solving (see, e.g., Chi, Feltovich, Glaser 1981; Ericsson, Krampe, Tesch-Romer 1993). For example, experts:

1) Organize knowledge around key principles and concepts
2) Pose useful questions to themselves about the information they are exploring
3) Recognize meaningful patterns and connections in information
4) Retrieve important aspects of their knowledge with little attentional effort

Recognizing that novices approach problems differently than experts, we draw on this research to provide guidance on how to train novices in the competencies that are characteristic of an expert data analyst.

To illustrate these ideas in a relatively simple context, consider a typical client-statistical consultant interaction: The client presents with a data set containing records of patient blood pressure measurements and treatment received and asks the consultant, "What is the effect of treatment on blood pressure?" A novice might immediately perform a significance test without considering the context of the problem. In contrast, experts organize knowledge around core concepts or big ideas that guide their approach to problem solving. As an illustration of points 1 and 2 from above, an expert's approach to the client's question might likely begin by asking about the study design and identifying which variable, treatment or blood pressure, is the response (**Y**) and which is the explanatory (**X**) variable. After identifying the variable types the expert data analyst would then categorize the variables as either quantitative or categorical, reasoning that if **Y** is quantitative and **X** is categorical she would boxplot **Y** vs **X**, whereas if **Y** is

categorical (e.g., high vs low blood pressure) and **X** is categorical (e.g., drug vs placebo) she would display the data in a two-way contingency table. The general point illustrated by this example is that the expert data analyst based on acquired content knowledge and a deep understanding of her subject matter plans a data analytic approach that looks first for features and patterns in the data before fitting a model or doing a test (point 3) (see also, Tukey 1977). It is also important to note that experts are able to move through this process, i.e., posing questions and flexibly retrieving aspects of their knowledge, fluently and without conscious thought (points 2 and 4).

As noted earlier, we wanted to develop a curriculum and pedagogical approach that would help novice students become more like expert data analysts. Specifically, our goal was to help our students attain a high degree of competence or mastery in the practice of statistics. For students to achieve mastery within a domain, they need to (1) develop a set of key component skills, (2) practice them to the point where they can be combined fluently and used with a fair degree of automaticity, and (3) know when and where to apply them appropriately (see Ambrose et al 2010, Chapter 4). In the following, we discuss each of these steps towards achieving mastery in the context of statistics instruction, and specifically, in the context of the MSP program.

2.1 Component Skills

To help students master complex tasks, like analyzing a data set or writing a report, it is necessary to decompose these tasks into subtasks or component parts. Once the component parts for a task are identified it is then necessary to provide students with the opportunity to practice these skills in relevant settings. For example, as we have

noted, a high priority in our program is developing skills in written communication. Most scientific papers can be broken down into component parts that correspond to the standard organization of a data analysis report, i.e., **I**ntroduction, **M**ethods, **R**esults and **D**iscussion, referred to as an IMRaD (Wolfe, Britt and Alexander 2011). From the first week of classes we introduce this format to the MSP students. Initially, the students practice writing the different components separately, e.g., just a methods section or just a results section, and as the semester progresses they get more complicated assignments that require them to integrate the IMRaD components into a full report. In addition, they are expected to use the IMRaD format in all of their courses across the curriculum. By the time we get to the capstone-consulting course in the second semester students have become quite familiar and fluent with the IMRaD format.

2.2 Deliberate Practice

Mastering complex tasks requires not only the temporary decomposition of subskills and the opportunity to practice them separately, but also their eventual recomposition and the opportunity to practice them in combination. Learning to drive a car provides a nice illustration of these principles. When first learning, a novice driver has to keep in mind a sequence of steps, e.g., adjust the mirrors, apply the brake, turn the key, put the car in reverse, check the rear view mirror etc., and a set of facts, e.g., traffic rules and laws, the function of car's controls and gauges, and a set of skills, e.g., accelerating smoothly, parallel parking. The novice driver also has to learn how to integrate all of these component skills and knowledge. As the student driver gains more experience, driving becomes effortless and automatic, requiring little conscious awareness.

With practice, students gain greater fluency in executing individual subskills and are better prepared to tackle the complexity of multiple tasks. Research has shown that learning and performance, however, are best fostered when students engage in practice that focuses on a specific goal or criterion, represents an appropriate level of challenge, is of sufficient quantity and frequency to meet the performance criteria and is coupled with targeted feedback (see, e.g., Anderson, Conrad, Corbett 1989; Lovett 2001). Consider the challenge of teaching students the habit of first looking for features and patterns in data before fitting a model or doing a significance test. From the outset of the MSP program we emphasize the importance of exploratory data analysis (EDA) and data visualization and we introduce this approach in the context of analyzing real data sets. We subscribe to the principle of "deliberate practice", that is, engage the students in highly structured activities which have clearly defined learning objectives using authentic data analysis problems (Ericsson, Krampe, Tesch-Romer 1993). For example, using parallel boxplots to compare the distribution of a quantitative response variable between two-independent groups provides practice in (i) looking for patterns in data, (ii) describing features of distributions to answer substantive questions about group differences, and (iii) describing features of distributions to investigate underlying formal assumptions of procedures like the two sample t-test.

Research has shown that practice must be coupled with feedback that explicitly communicates about some aspect(s) of students' performance relative to specific target criteria, provides information to help students progress in meeting those criteria, and is given at a time and frequency that allows it to be useful (see, e.g., Balzer, Doherty, O'Conner 1989; Ambrose et al, Chapter 5, 2010). Throughout the curriculum, we do

structured in-class data analysis exercises (students are provided with laptops) that give students practice with concepts and methods that they may have just learned in class. In-class exercises such as these allow instructors to give students immediate and targeted feedback and to correct any misunderstandings in real time.

2.3 Transfer

Achieving mastery in a subject not only requires component skills and the ability to integrate them successfully but also requires that students know when and where to use what they have learned. The application of skills learned in one context to a novel context is referred to as transfer (see, e.g., Chi, Feltovich, Glaser 1981; Barnett and Ceci 2002). To help students be able to transfer relevant skills to new settings it is necessary for them to have a robust understanding of underlying principles and deep structure, that is, they need to understand what to do and why. Students who learn to drive a Ford, for example, should be able to transfer their skills to be able to drive a Cadillac.

To help facilitate transfer the use of component skills should not be tied too closely to the context in which they were originally learned (Mason Spencer, Weisberg 1986). Beginning a data analysis, for example, by identifying the response and explanatory variables and whether they are quantitative or categorical is an approach that is not tied to a particular context and provides a generalizable framework for planning an analysis. Giving students the opportunity to apply what they know in multiple contexts fosters less context-dependence and more flexible knowledge. We do this using structured homework assignments and less structured data analysis projects.

It is also important for students to have a robust understanding of underlying principles and deep structure (see, e.g., Schwartz, Lin, Brophy and Bransford 1999). For example, we do not teach EDA just as a collection of discrete techniques but rather we motivate and emphasize the use of EDA as a crucial step in discovering patterns in data that help answer substantive questions.

3. On Teaching Statistical Practice – In Practice

In developing the MSP program a key pedagogical challenge was how to teach good statistical practice and thinking. In other words, how do we help novice students become more like expert data analysts?[4] Following the steps for developing mastery described in the previous section, we first needed to identify the component skills for doing data analysis, which we describe in the next section, and then create the opportunities for authentic practice and transfer. It is important to note that we incorporate these elements and integrate them throughout the entire program and not just in a particular course. For example, during the first semester in our *Perspectives of Data Science* course we introduce and practice the component skills for doing data analysis, and during the second semester, specifically in the capstone, consulting course, the students have the opportunity to integrate and transfer those skills in a real context. From the outset of the MSP program, we recognized the importance of

[4] We note that while this paper was in review a complementary article by Kass et al. (2016), "Ten Simple Rules for Effective Statistical Practice," concerned with communicating the elements of effective statistical practice to researchers appeared in *PLoS Computational Biology.*

providing the students with an authentic collaborative consulting experience, as similar as possible to what they would encounter in the work place.

3.1 Component Skills

Statistical practice is a form of problem solving and consists of a complex combination of component skills. Research in cognitive science has shown that experts follow four general steps in problem solving: (I) Represent the problem; (II) Determine the solution strategy; (III) Execute the strategy; and (IV) Evaluate the results (see, Bransford, Brown, and Cocking 1999). Using these general steps as a framework, we unpacked the complex tasks involved in statistical practice to identify 9 component skills that an expert statistician uses in engaging in data analysis.

We decomposed the first step, “Representing the problem,” into two component skills:

1. Understand the problem and the context
2. Understand the variables and data structure

Good statistical practice requires a nontrivial understanding of the real-world problem, the population for whom the research question is relevant, the study that generated the data, and the variables in the data set. Colin Mallows in his 1997 Fisher Memorial Lecture called these considerations the zeroth problem arguing that these issues must be formulated “before we can analyze the data, and certainly before we have a model.” He believed that “. . . statisticians must think about the real problem and must make judgments as to the relevance of the data in hand. . .” (Mallows 1998). We would add that these considerations are essential for study planning and study design as well.

Furthermore, experienced statisticians appreciate that without a keen understanding of how the data were acquired and how the data were preprocessed serious errors can arise that may affect the validity, reproducibility, and quality of the research (see, e.g., Zhu, Hernandez, Muller, Dong, and Forman, 2013).

The second step for problem solving, “Determine the solution strategy,” can be decomposed into two component skills specific to data analysis:

3. Exploratory Data Analysis (EDA)
4. Model specification: (a) Analytic Plan

Research in cognitive science (see, e.g., Chi, Feltovich, and Glaser 1981) has shown that experts recognize meaningful patterns based on previous experiences, which helps guide them in planning a problem-solving strategy or critiquing a theoretical perspective. With respect to helping the novice data analyst become more like an expert, an emphasis on EDA provides students with a perspective and set of tools that enables them to search for patterns in data (as well as exceptions to patterns) and to develop insight. Based on both subject-matter knowledge and exploratory data analysis, the student should be better informed to specify an analytic plan that considers feasible classes of models as well as to anticipate issues in the data that might compromise the validity of inferences based on those models.

For the third general step for problem solving, “Execute the strategy,” we identified three component skills, skills that are perhaps most closely associated with doing good data analysis:

5. Model specification: (b) Model selection
6. Assess model validity
7. Sensitivity analysis

Component skills 5-7 are the subject matter of methods-specific courses, like applied linear models or time series. A theme that is emphasized across our curriculum is that data analysis involves judgments, including, for example, judgments regarding model selection, validity of underlying model assumptions, the role of confounding variables, and issues of causality. We repeatedly emphasize the importance of investigating the robustness of inferences to these different judgements and data analytic decisions. This perspective helps underscore the role of data analysis in the larger framework of scientific discovery and the importance of recognizing the crucial interplay between the statistical modeling and the research context. We also note that in any real data analysis this set of subcomponent skills (#5-7) would be engaged iteratively.

Finally, we decomposed the fourth general step for problem solving, “Evaluate the results”, into two component skills:

8. Assess whether the research question was answered
9. Communicate results and conclusions effectively to the investigator in non-technical language

The importance of interpreting the statistical results of an analysis in the context of the problem cannot be emphasized enough. The students must be able to reflect on how the analysis has addressed the original research question as well as how the analysis has generated new insights. To do this effectively students must understand the research domain and the context of the problem (i.e., component skill #1). Finally, an essential component of good statistical practice is the ability to interpret and communicate the results of a statistical analysis so nonstatisticians can understand the findings.

We also include under the umbrella of statistical practice a discussion of those skills necessary for interacting with a client in a consulting setting. We introduce those component skills through readings and class discussion several of which we discuss in the next section.

3.2 Authentic Practice and Transfer

We recognize that there is a wide gap between learning the component skills for doing data analysis and working with an actual client on an open-ended consulting project. We help our students bridge this gap in a number of ways, including role modeling different aspects of interacting with a client, coaching students through the steps of doing data analysis, and supporting or scaffolding students as they engage in different data-analytic tasks. These methods, modeling, coaching and scaffolding, are examples of methods grounded in the *theory of cognitive apprenticeship*, a theory that explains how experts can more effectively promote learning in real-world situations (Collins,

Brown and Newman 1989). In this section we illustrate the role of cognitive apprenticeship in the MSP program. (See, also, Smucker and Bailer 2015.)

During the first semester of the *Perspectives of Data Science* course, we begin to prepare the students for the capstone consulting experience by introducing the soft-skills associated with consulting, including how to interact with a client (see, for example, Kirk 1991; Cabrera and McDougall, Chapter 2, 2002) and how to be a successful statistician (see, e.g., Hahn and Doganaksoy, Chapter 6, 2011). In addition, a significant component of this course is the introduction of an authentic consulting project that the students work on together as a class with faculty supervision. This project is much more actively directed by the faculty than the capstone consulting project in the second semester. The emphasis of this experience is primarily on interacting with the client, understanding the problem and the data (component skills #1 and #2), EDA, some preliminary modeling, reflecting on what has been learned and what the next steps should be (e.g., component skill #8), and communicating the results to the investigator. The faculty instructor serves as a role model, actively modeling and coaching the students. As an example, in the fall 2015 semester the class project was based on a request for analysis (RFA) from the US Census Bureau. The RFA stated:

> The Census Bureau is in need of innovative research to support the 2020 Census communications campaign . . . The agency is seeking a geographically-defined dataset that contains an audience segmentation of the US population and housing. For our purposes, audience segmentation is the process of dividing housing units into homogeneous subgroups based upon defined criterion. For social marketing purposes used during the Decennial Census,

> audience segments are defined by a wide variety of variables such as demographics, likelihood to self-respond, preferred information channels, media consumption, and knowledge, attitudes, and perceptions of the government. This segmentation will be used to help plan and implement the 2020 Census Integrated Communications Campaign.

Specifically, the Bureau wanted segments to be predictive of Census self-response and should have characteristics that are “actionable” in terms of helping develop media targeting strategies. We were fortunate to have a representative from the Census Bureau (the client) present the project to the class in person. A nice feature of this project was that students working in teams could choose a state and investigate non-response and segmentation in their state while using a common set of methods.

In the second semester of the MSP program, the capstone consulting project provides the opportunity for the students to engage in a less scaffolded consulting project with an authentic client and a real open-ended research problem. There are approximately 10-12 clients drawn from academia as well as from the public and private sectors. Some representative examples of clients and projects can be found in Table 2.

Table 2. Selected Examples of Clients and Consulting Topics

Client	Topic
Public/Government	
Pittsburgh Public Schools	An Assessment of Early Childhood Education in the Pittsburgh Public Schools
	Evaluating the Impact of School Closing on Students' Academic Performances
Pittsburgh Police Bureau	The Epidemiology of Homicides in Pittsburgh 2008-2014
Academia	
English	Analysis of Hillary Clinton's Rhetorical Style
Psychology	Classification of Aphasia Types using Behavioral Measures of Communication
	Stress and Depression in Parents of Diabetic Adolescents
	The Role of the Prefrontal Cortex in Patients with Congenital Prosopagnosia
Learning Science	Evaluation of Online Learning in Community Colleges
Private	
Market Research	Predicting Consumer Sentiment using Internet Marketing Surveys
Manufacturing	Factors Related to Aluminum Smelting Efficiency

Teams consisting of 2-3 students are assigned to each project along with one of the two faculty mentors assigned to the course. The faculty mentor is available to provide support as needed. For example, since these are real consulting projects some projects may have well-defined goals whereas others may be more open-ended. The faculty mentor is available to help the students sharpen the focus of the project as well as provide general expert advice where this is required. To help scaffold the students through the project, we set milestones that correspond to the 4 general steps for problem solving discussed in section 3.1 - (i) represent the problem; (ii) determine the solution strategy; (iii) execute the strategy; and (iv) evaluate the results. At each milestone, each team makes a power point presentation to the entire class on their progress. In addition, the students submit written progress reports using the IMRaD format which over the course of the semester culminates in a final report for their client. A nice feature of these milestone presentations is that the students get to learn about issues of statistical practice that arise in the other projects.

4. Reflections on the MSP Program

As noted earlier, a central goal of our program was to help our students become more like expert data analyst and to help get them job ready. In this section, we provide some feedback from the student perspective on our progress towards this goal. With respect to employment, our students have had no difficulty finding jobs. All have found employment within 3 months of graduation. The typical position is as an entry master's level data analyst. In the early years of the program many of the students were looking for and found positions in banking or finance (e.g., Bank of America, Capital One, Citi, TransUnion) or with consulting firms (e.g., Booze-Allan, Deloitte). More recently, our

students have been looking for and finding positions in the tech industry (e.g., Google, Zillow, Facebook) and companies doing marketing analytics (e.g., CivicScience, dunnhumby). A handful of students have taken biostatistics positions at research institutions or have gone on to PhD programs in statistics or biostatistics. Informal conversations with employers have provided positive feedback about our students' training and job preparedness.

We were interested in hearing from our graduates after they had been working for at least one year as to what features of the MSP program they found particularly helpful for their jobs. We had email addresses for approximately 60 alumni from the entering cohorts 2009-2013 (out of a total of 99 students) and asked them to respond to the following prompt:

> *What are your personal reflections on how the MSP program helped you better develop your skills as a statistical scientist and what aspects of the program helped you to learn these?*

We received 30 responses. We did not intend for this survey to be an evaluation of our master's program, but rather, we were interested in seeing if the students were able to identify themes based on the cognitive principles of learning that we have discussed in this paper. From their open ended responses we identified 6 major themes: Statistical Methods (i.e., course work), Computing Skills (e.g., R, SAS, SQL), Communication Skills (e.g., oral and written presentations), Soft Skills (i.e., professional skills such as working in groups, elevator talks, resumes), Authentic and Targeted Practice (e.g, working with real data), and Statistical/Critical Thinking. The results (number of times a

theme was identified among the 30 respondents) are presented in Table 3. A respondent could identify more than one theme. Interestingly, these themes also align with the results of the survey conducted by the ASA Workgroup on Master's Degrees (Bailer et al 2013). Clearly, the students value the opportunities they have had for authentic practice and appreciate the opportunities to engage in statistical thinking.

Table 3. Frequency of Responses to Survey Prompt

Statistical Methods	Computing	Communication	Soft Skills	Authentic Practice	Statistical/Critical Thinking
18 (60%)	19 (63%)	15 (50%)	7 (23%)	22 (73%)	21 (70%)

This is by no means a scientific evaluation of the MSP program but rather an opportunity to let the students report in their own words what aspects of their graduate training they have found useful. To provide more context, the following is a representative selection of actual student responses organized around the three themes of (i) authentic practice; (ii) transfer; and (iii) statistical/critical thinking.

(i) On Authentic Practice

- I recall that for all the courses, professors would talk about real examples in class and we did lots of practical data analysis. It is very useful to learn how to apply the methods instead of just the methods themselves. So we will be able to handle unexpected situations, e.g. what if the assumptions are not satisfied for this specific data.

- [As an] undergraduate, the datasets we get were always clean, and all we needed to do was to start our analysis on it. But in one of [our] classes, we actually got a dataset that had not been cleaned . . . I think it's really good practice for us to get use to cleaning the dataset first before analysis.

- MSP has helped me cultivate the ability to interpret and communicate results, especially through writing the IMRaD report. Writing this type of document is necessary in real work, where we need to communicate our findings with others. MSP has provided lots of opportunities and suggestions on writing IMRaD, which makes us think more concisely and be able to tell the story behind the analysis.

- It's important to learn how to introduce ourselves. In one of your classes, we all wrote our "elevator speech", and practiced over and over again in classes and to different people. Even though it got a bit boring after repeating the same thing for 50 times, it turned out to be really handy when we were asked to introduce ourselves [e.g., at interviews] . . . Since I've practiced it so many times, I didn't have to pause and organize my thoughts.

(ii) On Transfer

- One thing that I feel is worth highlighting is the consulting project, as it was a unique experience that helped a great deal in developing our statistical thinking skills. Given a research question and a data set, we thought through each aspect of the analysis, from cleaning the data, deciding what methodology to use, doing

the analysis and presenting the results. These are skills that are applicable to working in industry, as many jobs (including mine) are project-based.

- After graduating from the MSP program, I joined a national bank responsible for building marketing models. The year I spent in the MSP program helped me become a better statistical modeler. For example, methods I learned in the epidemiology class, such as, survival analysis could be applied to determine the credit default risk for the bank, or logistic regression could be used to determine the likelihood customers would accept the bank's product offer.
- I learned to first understand the problem, check assumptions and apply statistical knowledge to solve real world problems. . . . As a result, I am not only able to use statistics to solve problems in sciences but also in various kinds of fields such as education, clinical trials and so on.
- I thought the course where we each did our own statistical consulting project in teams was really excellent. Not only did we learn from our own projects, but because we presented to the group regularly, we learned from everyone else's experiences as well.

(ii) On Statistical Thinking

- I have learned how to translate a real world problem into a statistical question.

- . . . the most important statistical skill wasn't strictly a technical one. It was critical thinking – the ability to extract the signal from the noise, to ask the right questions, to gleam insights, and to foresee potential issues.
- I also developed a ‘6th sense’ for real-world datasets. Questions like “why are these values missing?”, “why are they so large?” and “If we execute the strategy the data suggests, will it do what we envision?” crop up all the time and need good answers.
- I also learned that being a data scientist incorporates a bit of guesswork, where a client comes to you with a question, but the question they need answered might not actually be the one they are asking. It also involves teaching, which I like, because you get to explain why a particular method is appropriate, what the assumptions and limitations are, and how to interpret the results.
- In the real world, data are rarely simple. The MSP program taught us to think through what exploratory data analysis is necessary to gain a better understanding of the data, which statistical methods are most appropriate to solve a problem, and what are the weaknesses to the results. This understanding allows me to think critically to learn and apply new statistical methods that I may not have been exposed to in the classroom.

Although a selection, these comments are representative of the responses we received.

5. Discussion

Perhaps in response to the emergence of the field of Data Science, many in the Statistics community have been actively promoting the central role of statistical thinking in the practice of statistics. For example, Horton and Hardin (2015) quoting Diane Lambert, propose that, "If statistics is the science of learning from data, then our students need to be able to think with data." Greenhouse (2013) argues that statistical thinking is the "bedrock of data science." Brown and Kass (2009) assert that ". . . the primary goal of statistical training at all levels should be to help students develop statistical thinking." Although this advocacy has been eloquent and strong, little has been written about how to improve the training of statisticians and others to think statistically (see also Wild and Pfannkuch 1999). In this article we have introduced principles of learning based on research in cognitive science and have illustrated how we have adapted and applied these principles to training novice students to become more like expert data analysts, including how to think statistically. Specifically, we have identified 9 component skills that we believe characterize an expert's approach to statistical practice, and have developed a curriculum where students have the opportunity to practice these skills in authentic settings. Feedback from students and employers suggests that the use of these principles of learning has been successful in helping our students develop expertise in the practice of statistics.

The principles of learning presented in this article offer guidance not only in the design and implementation of an integrated curriculum, such as the MSP program, but in designing an individual course as well. It is useful to note that the steps in helping students achieve mastery discussed in this article are also the steps one would use in

designing and implementing an individual course. That is, to achieve mastery students must acquire component skills, have opportunities for deliberate practice with targeted feedback, become fluent and proficient in the application of these skills, and have the opportunity to apply what they have learned in authentic settings. We have presented a general framework for how learning works that we hope provides a roadmap for instruction in general, and for developing mastery in statistical practice in particular.

References

Ambrose SA, Bridges MW, DiPietro M, Lovett MC, and Norman MK (2010). *How Learning Works: Seven Research-Based Principles for Smart Teaching*. San Franciso: Jossey-Bass.

Anderson JR, Conrad FG, and Corbett AT (1989). Skill acquistition and the LISP tutor. *Cognitive Science*, 13(4):467-505.

Bailer J, Hoerl R, Madigan D, Montaquila J, and Wright T (2013). Report of the ASA workgoup on master's degree.
http://magazine.amstat.org/wp-content/uploads/2013an/masterworkgroup.pdf

Balzer WK, Doherty ME, and O'Connor (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, 106:410-433.

Barnett SM, and Ceci SJ (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4):612-637.

Bransford JD, Brown AL, and Cocking RR (editors) (1999). *How People Learn*: *Brain, Mind, Experience and School.* Washington DC: National Academy Press.

Brown EN, and Kass RE (2009). What is Statistics? *The American Statistician*, 63(2):105-110.

Cabrera J and McDougall A (2002). *Statistical Consulting*. New York: Springer.

Chi MT, Feltovich PJ, and Glaser R (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2):121-152.

Collins A, Brown JS, and Newman SE (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.) *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Lawrence Erlbaum Associates.

Committee on Enhancing the Master's Degree in the Natural Sciences (2008). *Science Professionals: Master's Education for a Competitive World.* Washington, DC: National Academy Press..

Ericsson KA, Krampe RT, and Tesch-Romer C (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363-406.

Greenhouse JB (26 July, 2013). Statistical thinking: The bedrock of data science. *Huffington Post.*
http://www.huffingtonpost.com/american-statistical-association/statistical-thinking-the-bedrock-of-datascience_b_3651121.html

Hahn GJ and Doganaksoy N (2011). Characteristics of a successful statistician. Chapter 6 in *A Career in Statistics*. Hoboken, NJ: John Wiley & Sons.

Horton NJ, and Hardin JS (editors) (2015). Teaching the next generation of statistics students to “Think with data”: Special issue on statistics and the undergraduate curriculum. *The American Statistician*, 69(4): 259-265.

Kass RE, Caffo BS, Davidian M, Meng X-L, Yu B, and Reid N (2016) Ten Simple Rules for Effective Statistical Practice*. PLoS Comput Biol* 12(6): e1004961. doi:10.1371/journal.pcbi.1004961

Kirk RE (1991). Statistical consulting in a University: Dealing with people and other challenges. *The American Statistician*, 45(1):28-34.

Lovett MC (2001). A collaborative convergence on studying reasoning processes: A case study in statistics. In S Carver & D Klahr (eds.) *Cognition and instruction: Twenty-five years of progress* (pp. 347-384). Mahwah, NJ: Erlbaum.

Lovett M and Greenhouse J (2000). Applying cognitive theory to statistics instruction, *American Statistician*, 54:196-206.

Mallow C (1998). The Zeroth Problem, *The American Statistician*, 52(1): 1-9.

Mason Spencer R, and Weisberg RW (1986). Context-dependent effects on analogical transfer. *Memory and Cognition*, 14(5): 442-449.

Schwartz DL, Lin X, Brophy S and Bransford JD (1999). Toward the development of flexibly adaptive instructional designs. In C. M Reigelut (ed.), Instructional Design Theories and Models: Volume 2. Hillsdale, NJ: Erlbaum.

Smuckler BJ and Bailer AJ (2015). Beyond Normal: Preparing undergraduates for the work force in a statistical consulting capstone. *The American Statistician*, 69(4):300-306.

Special Issue on Statistics and the Undergraduate Curriculum (2015). N Horton and J Hardin (eds.). *The American Statistician*, 69(4): 259-424.

Tukey JW (1977). *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Wild CJ and Pfannkuch M (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review*, 67(3):223-265.

Wolfe J, Britt C and Alexander KP (2011). Teaching the IMRaD Genre: Sentence combing and pattern practice revisited. *Journal of Business and Technical Communication*, 25(2): 119-158.

Zhu Y, Hernandez LM, Muller P, Dong Y and Forman, MR (2013). Data acquisition and preprocessing in studies on humans: What is not taught in statistics classes?" *The American Statistician*, 67(4):235-241.

Appendix

FAQs about Carnegie Mellon University's Master's of Statistical Practice Program

1) How many students apply and how many are admitted?

   We receive over 350 applications per year. Our target enrollment is about 25 students and to reach that target we admit approximately 45 students.

2) How many women and how many US citizens are enrolled?

   A little less than 60% of our students are female and over 50% are US citizens. The numbers vary from year to year.

3) What are the course prerequisites for admission into the program?

   We expect students to have had a calculus based mathematical statistics course, and a regression/econometrics course. We also expect students to have had a course in matrix algebra.

4) What sort of background do the students have?

About 80% of the students are coming straight from an undergraduate program. Most students have a statistics, mathematics or economics undergraduate degree but we also have students with degrees in engineering, business or the social sciences.

5) Do non-native speakers need to demonstrate English language proficiency?

Yes. We have found that English language proficiency is critical in helping our students find jobs. In addition, since we hire our master's students as TAs and graders it is necessary for them to pass a University administered English language proficiency examination.

6) Is this a pay-to-participate program?

Yes. Students pay full tuition.

7) Do students receive any financial assistance?

We hire our master's students to be teaching assistants or graders. In addition to providing them with some financial assistance, we have found that the opportunity to explain statistical ideas to undergraduates is a valuable experience which helps them develop communication skills that are valued in the work place.

8) Do students receive help in finding jobs?

Yes. The MSP Director along with campus career specialists work with each student to help prepare their resume, cover letters, and to develop their elevator talk. In addition to advising students about career choices and opportunities, the faculty will help students with practice interviews.

9) Do students receive training in project management and working in groups?

Yes. In the fall semester as part of their professional training we hold several workshops led by campus experts on project management and group work. Throughout the year the students are assigned projects that require them to work in groups.

10) Do the MSP students take all of their classes together?

Yes. The program is coordinated and well integrated so that material learned in one class, e.g., writing the IMRaD report, is used in other classes. Under the leadership of the MSP Director, MSP faculty meet at least 3 times during each semester to help ensure program integration and to monitor student progress.