Modeling First: Applying Learning Science to the Teaching of Introductory Statistics

A final peer-reviewed version of this article will appear in:
Journal of Statistics Education

Ji Y. Son,[1] Adam B. Blake,[2] Laura Fries, James W. Stigler

Abstract

Students learn many concepts in the introductory statistics course, but even our most successful students end up with rigid, ritualized knowledge that does not transfer easily to new situations. In this article we describe our attempt to apply theories and findings from learning science to the design of a statistics course that aims to help students build a coherent and interconnected representation of the domain. The resulting *practicing connections* approach provides students with repeated opportunities to practice connections between core concepts (especially the concepts of statistical model, distribution, and randomness), key representations (R programming language and computational techniques such as simulation and bootstrapping), and real-world situations statisticians face as they explore variation, model variation, and evaluate and compare statistical models. We provide a guided tour through our curriculum implemented in an interactive online textbook (CourseKata.org) and then provide some evidence that students who complete the course are able to transfer what they have learned to the learning of new statistical techniques.

[1] California State University, Los Angeles
[2] University of California, Los Angeles (all authors except as noted)

The authors gratefully acknowledge the support of the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (DRL-1229004) and the California Governor's Office of Planning and Research (contract OPR18115).

Corresponding Author:
Ji Y. Son, Psychology Department, Cal State LA. Email: jiyunson@gmail.com.

In the typical introductory statistics course, students are introduced to a great number of concepts: variance, z-score, normal distribution, p-values, ANOVA, t-test, and many others. At the end of the course, however, many students have difficulty appropriately transferring what they have learned to new situations. Gigerenzer (2018) has argued that students and researchers alike end up practicing statistics as a set of rituals, a series of actions repeatedly performed in a prescribed order without judgment. Triggered by a certain set of conditions, our students happily "run the analyses" and report the results, all without a deep understanding of why they did what they did, or what the results mean.

As learning scientists who also teach introductory statistics, we embarked several years ago on a project to apply what we know from research on learning to the teaching of statistics, and in particular on how to support students' development of deep understanding and *transferable* knowledge. Many of our students score well on the test but are not able to apply what we have taught them to new situations. We know from research on expertise that connecting problems and procedures with the core concepts of a domain makes knowledge more coherent, and therefore more flexible and transferable (e.g., Bransford, Brown, & Cocking, 1999; Lachner, Furlitt, & Nuckles, 2012; Thagard, 2007). Our plan was to first figure out what the core concepts of the domain should be (i.e., those that would be most fruitful for novices), and then to develop a pedagogical approach that would help students use these concepts to interconnect their knowledge of statistics.

Complicating our plan was the fact that statistics itself is undergoing a massive transformation. Modern statistics has moved beyond the assumptions and mathematical approximations of the last century; it has become a computational science (e.g., Nolan & Lang, 2012), gradually replacing tools such as the normal distribution with techniques such as simulation, randomization, and resampling (e.g., Hesterberg, 2015). We also have seen a shift from an emphasis on Null Hypothesis Significance Testing (NHST) to the construction and evaluation of statistical models (Rodgers, 2010). Some textbook authors have started to rethink the curriculum in light of these developments (e.g., Judd, McClelland, & Ryan, 2017; Kaplan, 2017; Tintle et al., 2015; Lock et al, 2016). The rise of data science, a more multidisciplinary approach to data, along with the proliferation of readily accessible data has also pushed the introductory statistics class to change.

Many statisticians argue that we should stop teaching NHST and/or mathematical approximations altogether (e.g., Cobb, 2007; Cohen, 1994; Cumming, 2014; Gigerenzer, 2018). The critique is not just leveled at the teaching of statistics, but also to statistics as practiced by researchers. *Nature* published a comment with 800 signatories calling for an end to these statistical practices *by scientists* (Amrhein, Greenland, & McShane, 2019).

These recent developments are important and have shifted our emphasis from the traditional topics of introductory statistics to a new focus on modeling and computational methods. At the same time, however, we worry that statistics educators will reform the content of the introductory course, but not the pedagogy. As psychologists, we see a very real possibility that ritualized learning of NHST (Gigerenzer, 2018) will simply be replaced by a new set of rituals, ones that turn the concept of bootstrapping, for example, into a flow chart that tells students what to do, step-by-step, but that does not result in deep understanding. If we are going to

redesign the introductory course to emphasize modeling and computational approaches, we also need a theory of pedagogy that leads to understanding rather than ritualization.

In this paper, we describe the project that has resulted from our efforts. We base our approach on a clear theory of pedagogy, adapted from current thinking in the learning sciences, that we call the *practicing connections framework* (Fries, Givvin, Son, & Stigler, 2020; Son, Ramos, DeWolf, Loftus, & Stigler, 2018). The practicing connections framework posits that to produce coherent transferable learning, students must practice making connections between core concepts, representations, and the world (i.e., contexts and practices involved in applying those concepts; cf. National Academies of Sciences, Engineering, and Medicine, 2018). In the first part of this paper, we propose what the core concepts and representations should be. If not NHST, the normal distribution, and ritualized procedures, which core concepts, representations, and connections to the world should students practice making in an introductory course?

The second part of this paper is a guided tour through our introductory statistics curriculum in which we demonstrate what our practicing connections pedagogy looks like as we put it into action. Our project is embodied in a free interactive online textbook called *Introductory Statistics: A Modeling Approach* (Son & Stigler, 2017-19), available for preview at CourseKata.org. This interactivity, in which we integrate modeling concepts with embedded coding exercises and questions with immediate feedback, differentiates our book from other textbooks that have been developed around modeling (e.g., Judd, McClelland, & Ryan, 2017; Kaplan, 2017). This feature also provides researchers and developers with a constant flow of data, based on more than 1200 embedded assessments, which we can use to improve the effectiveness of our book. This project is part of a larger effort, called the *Better Book Project*, to modernize research and development in education (see Stigler et al., in press). In brief, we are trying to develop a new approach to the development of curriculum materials in which researchers, designers / developers, and instructors work together to produce continuous incremental improvements to an online book.

In the third part of the paper, we present some preliminary data, which we find encouraging. In particular, we ask, can we see any evidence of the transferable knowledge we are trying to produce in our students? Although the data we present are from a sample of UCLA pre-psychology majors, our ultimate interest is in all students, not just advanced students or those pursuing majors in statistics or STEM fields. Thus, we are currently implementing our online book in a variety of settings, mostly in California, ranging from the University of California to California State University to community colleges. We especially want to find ways of facilitating deep learning and flexible, transferable knowledge among students who have been deemed "underprepared" for quantitative courses, and to prepare all students for future courses in advanced statistics should they decide to take them.

Practicing Connections: The What

Transfer can be broadly defined as applying old knowledge to new situations (e.g., new contexts, new concepts, new representations; see Barnett & Ceci, 2002 for a taxonomy of transfer). Research shows, however, that transferable learning is often difficult to achieve. Many statistics educators would agree, based on their own experience, that students who can easily solve problems that are similar to what they were exposed to during learning have difficulty

solving problems that deviate significantly from the ones they were taught to solve (e.g., Bassok, Wu, & Olseth, 1995; Son et al., 2018).

Research in the learning sciences has demonstrated that flexible transfer is best supported when knowledge is coherent (for a review see Fries et al., 2020). If the goal of learning and teaching is to create coherent knowledge structures in novices, then students should not spend their time accruing bits of disconnected information (Let's learn about the median! Make a boxplot! Calculate a z-score!). Instead, the pedagogy should be focused on helping students make appropriate connections that can help organize the domain.

According to our practicing connections hypothesis, students' knowledge will become more coherent to the extent that they practice connecting core concepts, key representations, and situations in the world. But which core concepts should they connect? Which representations? What is the range of situations and contexts we want students to include in their knowledge of the domain? These are the questions we start with.

Core Concepts

A well-known finding from research on expertise is that experts "see the structure" of a domain. Expert knowledge comprises fewer concepts that are highly interrelated compared to student knowledge (Lachner, Furlitt, & Knuckles, 2012). This allows experts to look at problems that appear to novices as unrelated and see them as embodying a common underlying structure (Chi, Feltovich, & Glaser, 1981). Although we can get a lot of help from domain experts in defining the core concepts for a domain, the core concepts that are accessible and helpful to novices may not overlap completely with those used by experts. Our goal is to choose concepts that are useful to novices as they build coherent knowledge structures that won't need to be "unlearned" later. (For example, see McGowan & Tall, 2010, for a discussion of how students' initial interpretation of the negative sign to mean "take away" interferes later with their understanding of negative numbers, and must, therefore, be "unlearned.")

The three core concepts we selected for our introductory statistics course are *modeling*, *distributions*, and *randomness*. Modeling, of course, is foundational for statistics and data analysis, yet often it is thought of as an advanced topic, not suitable for beginners. We introduce the concept of statistical model at the very beginning of our course, and continually help students connect the tools, methods, and ideas they are learning to the overall framework, DATA = MODEL + ERROR. Our second core concept is distribution, which we subdivide into three types: samples, populations (what we also call the Data Generating Processes, or DGPs), and sampling distributions. Our third core concept is randomness. Cognitive scientists have long pointed out that people intuitively consider causality and ignore information that does not lend itself to causal interpretation (e.g., Tversky & Kahneman, 1977). We attempt to train students to think of randomness as a data generating process that can *cause* distributions of data.

Representations

In order to support the making of connections, the core concepts must be embodied in a set of representations that are highly generative. We use a variety of representations in our curriculum, but we focus throughout on three that specifically support the connections that

create coherent knowledge of statistics. Research on comparison (Namy & Gentner, 2002), symbolic reasoning (Son, Smith, & Goldstone, 2011), and math manipulatives (Uttal, Scudder, & DeLoache, 1997) shows that tying disparate instances to one stable, repeating, relatively abstract instance leads to flexible generalization.

The first representations we emphasize are graphical displays of data such as histograms, box plots, scatter plots, and so on. We want students to be able to spot patterns in data, to connect raw data in tables to specific points on graphs, and to connect features of graphs to core concepts such as models and randomness.

The second representational system we use is R (taking advantage of the many innovations of *Project MOSAIC*; Kaplan & Pruim, 2019; Pruim, Kaplan, & Horton, 2017) to integrate modern computational methods such as simulation, randomization, and bootstrapping into our course. Our goal is not to teach programming, but instead to use simple R functions as a notational system for representing abstract statistical concepts, manipulating data, and also as a way of creating publication-ready graphs to represent statistical ideas. Our students start out thinking of R as just a set of commands that do things, and in fact this full set of commands is given to them as a one-page "cheatsheet".

Upon hearing about our integration of R throughout the textbook, many people have expressed concern that the additional cognitive load imposed by learning R will have a negative impact on students' learning and engagement. In fact, we have reason to believe the opposite. Rather than *extraneous* (or unnecessary) to the learning task, we argue our use of R is *germane* load, actively facilitating the learning of complex and often abstract concepts (Sweller, Van Merriënboer, & Paas, 1998). Using R to simulate a population or to shuffle data, for instance, allows students to gain firsthand experience of these abstract statistical techniques as they become concrete through the manipulation of code. And while many students do indeed begin wary about learning R (as reported on pre-course surveys), their feelings about R end up significantly more positive at the end of the course (Tucker, Shaw, Son, & Stigler, under review).

Finally, though we make minimal use of mathematical notation, we do rely heavily on the notation of the General Linear Model (GLM) as a means of connecting together topics, such as ANOVA and regression, that students typically see as unrelated. We ease into the notation by starting with word equations. For example, we first teach students to represent a relationship by writing:

[Outcome variable] = [Explanatory variable] + other stuff

We then transition them to writing the following GLM equation, all the while connecting both word equations and GLM notation to the concept of DATA = MODEL + ERROR:

$$Y_i = b_0 + b_1 X_i + e_i$$

It is worth pointing out that we do not solely rely on algebraic expressions and equations to represent statistical ideas, despite the fact that most traditional courses do. Our reason for this is twofold. First, many ideas expressed algebraically are more convincingly communicated to

novices as computational ideas instantiated in R code. Second, many of our students — and this seems true across a wide diversity of students — do not "own" algebra enough for it to help them understand statistical ideas. For these students, just the presentation of mathematical derivations may induce them to engage in calculations or activate negative emotions and thus interfere with thinking and understanding (Geller, Son, & Stigler, 2017; Givvin, Stigler, & Thompson, 2011).

The World

It isn't enough to have a set of core concepts and representations connected to each other. These, in turn, must be connected to different types of contexts in the world, which are embedded within a goal structure that defines what it means to "do statistics."

Goal structure. We organize and frame our course as a narrative about the practice of doing statistics. The narrative starts with the idea that statistics is all about making sense of variation. We divide our course into three main parts, corresponding to the main goals of statisticians when analyzing data: exploring variation, modeling variation, and evaluating models.

In the first part of the course, students use R to explore variation visually. They apply the concept of *distribution* as a lens for looking at variation (Wild, 2006), and are encouraged to think of distributions as a new kind of entity, similar to the way we might shift our attention from looking at individual birds to see the behavior of flocks. Students also are encouraged to look for patterns in data that might provide clues about the data generating process (DGP) that gave rise to variation in data.

In the second part of the course, we introduce the concept of statistical model and begin developing the idea that we can use mathematical functions as models that summarize distributions and allow us to make predictions of future cases. No matter how complex the model, it always generates a predicted score on an outcome variable based on one or more explanatory variables. In the simplest model, the function generates the same score (e.g., the mean) regardless of the value of an explanatory variable. Importantly, constructing a statistical model gives birth to error, the part of variation that has not been explained by our model. The goal of modeling is to reduce error, and students learn to fit more complex models as a means of reducing error.

Finally, in the third part of the course, we tackle the problem of sampling variation and introduce students to tools for comparing and evaluating alternative models of the DGP. We practice generating data in different ways (randomization, bootstrapping, and simulation) and examine the sampling variation that results from these computational methods. We use the resulting sampling distributions to estimate confidence intervals and provide evidence for/against different models of the DGP.

Contexts. Although in the section above we explained these goals and practices abstractly, in the student-facing instructional materials, these practices are always embedded in different contexts in which these goals are pursued. Context in statistics is often thought of as just the content of the problem — students' finger lengths, GDP of countries, mercury levels in lakes, and so on. In our practicing connections framework, we distinguish two ways that context can

work to facilitate connections. The same context can be used to connect two different kinds of concepts (e.g., using the same data set about student's finger lengths to demonstrate both group and regression models), and the same concept can be instantiated in different contexts (e.g., fitting a regression model to data from students versus countries). We call the first type *connecting contexts*, because these contexts serve as the connecting glue for different concepts, and the second type *expanding contexts*, because these contexts help students expand their application of a concept.

In our online book, we focus on just a few connecting contexts so that students can use superficially similar instances to help them make structural distinctions (Sagi, Gentner, & Lovett, 2012). For example, early on in the textbook, we explore variation in students' thumb length. Later we refer back to that same context, modeling that familiar variation and evaluating those models. We continually refer to a model where we use a student's sex to predict their thumb length (a group model) to build up connections between different measures of model fit such as Sum of Squares Model and F ratio, and to distinguish those concepts from parameter estimates (e.g., b_1). We also use the same thumb length scenario to introduce regression models by using students' heights to predict their thumb lengths.

To the surprise of many first time instructors, the idea of a categorical variable is often hard for students to keep in mind. When the context changes, they often confuse the variable (e.g., diamond cut quality) with the levels of a variable (e.g., fair, good, ideal). This is perhaps reasonable given there are also situations where the name of the variable (e.g., treatment) is the same as one of the levels (e.g., treatment vs. control). Repeated use of a familiar context helps to ground students' understanding of the role of a categorical variable in a group model. Thus, students have multiple strategies when faced with the question, "Is this a categorical variable?" Not only can they ask themselves, "Is it a variable or a value?", they can also ask, "Is it more like sex or more like male/female?"

At the same time, we also identified key features of *expanding contexts* that students should be exposed to in order to prevent limiting their understanding to just a few familiar concrete examples. In our book, and especially in class, we seek to expand the range of contexts to which students' can connect their developing knowledge of statistics.

For example, especially in majors such as psychology, students are used to thinking about cases as individual people and have more difficulty conceptualizing other entities as cases. We made sure to include specific expanding contexts such as non-people (e.g., lakes, movies, companies, countries) and to address mistaken interpretations that arise from assuming that the cases are individual people. After creating a scatterplot that shows a positive relationship between consumption of alcoholic spirits in a country and their happiness index (combining data from the World Health Organization; Kim, Ismay, & Chunn, 2018, and the Happy Planet Index; Lock, 2017), students may mistakenly believe that individual people who drink more are happier. We use these expanding contexts to address these types of misconceptions.

In summary, we can think of students' growing understanding like a rubber band stretched around a group of nails on a wooden board. With each nail representing a distinct context, we aim to stretch the rubber band, over time, to encompass as many contexts as possible. *Expanding contexts* serve this purpose and are the engine that drives flexibility and transfer. On

the other hand, we must at the same time strengthen the rubber band itself so that it doesn't break as we stretch it to encompass more varied contexts. *Connecting contexts* play this role, strengthening understanding and preparing it to stretch. Both ideas are critical to our design.

Practicing Connections: How Concepts are Developed

Having laid out the components we want students working to connect—the core concepts, key representations, and contexts—we can now turn to the *how* of practicing connections. What does the pedagogy look like that we hope will support students' developing understanding of statistics? How do we implement this pedagogy in Version 1.0 of our book? To answer these questions, we will take a single core concept (DATA = MODEL + ERROR), and show how it is developed throughout the book.

Exploring Variation

We start the book by moving from variation in the world to variation in data. This part of the book (Chapters 2-4) generally takes 2-3 weeks in a semester-long course, about a chapter per week. Although our book does not focus on measurement or research design, we do try in Chapter 2 to make sure students understand where data come from, and how data get organized into data frames (where rows are cases and columns are variables). The rest of the book is about understanding variation in data and considering different sources of variation (including real characteristics of the DGP as well as variation induced by the data collection process). Starting in Chapter 3, we use R to introduce various kinds of graphs and the five-number summary as tools that help us see patterns of variation. Distribution, a core concept in our approach, is the primary conceptual lens through which we view variation (Wild, 2006). It is a way to see the forest for the trees, the flock instead of birds, the traffic instead of cars.

We can learn a lot by examining distributions of sample data. Usually, however, our interest goes beyond the variation in the data; we want to *explain* that variation. Why is there variation? What caused it? How can we predict it? These questions reveal that the goal of statistics is to know something about the Data Generating Process (or DGP) or the population (the long run result of a DGP). When we examine distributions of data, we do so to help us understand the DGP that caused that variation (see Figure 1 for how the sample and DGP connect). These two kinds of distributions (data and the DGP) make up two-thirds of what we refer to as the *distribution triad*. (Later, in the evaluating models part of the course, we will formally bring in the third kind of distribution, the distributions of statistics, or sampling distributions).

In Chapter 4 we introduce an intuitive definition of what it means to "explain variation." We can say that one variable *explains* some variation in another variable if knowing a case's value on the first variable helps us to make a slightly better guess about what that same case's value on the second variable might be. For example, sex explains variation in thumb length because knowing a student's sex helps us make a slightly better prediction of their thumb length. Students learn to "see" this idea by looking, for example, at the faceted histogram in the lower left of Figure 1. Just by visualizing the data, students have the intuition that our predicted value for a male student's thumb length should be a little bit bigger than our prediction for a female student.

Even before we introduce the formal idea of a statistical model, we begin to build on our intuitive definition of "explain," teaching students to write word equations to express informal models. Students see that knowing someone's sex can help to predict their thumb length but they also see that there is a lot of variation still unexplained by sex. We teach students to represent these ideas by writing word equations such as **Thumb = Sex + other stuff**. (Later we will turn "other stuff" into the more formal concept of "error.")

Similarly, we help students construct an informal idea of what sampling variation looks like, and why it's important, in the beginning of the book, well before we get to the formal concept of sampling distributions. Students see the relationship between Sex and Thumb length in sample data (see the faceted histogram on the left of Figure 1). We then ask them, if we took another sample of students, do you think the graph would look the same? Most agree that it would be similar, but not exactly the same. This naturally leads to the insight that a relationship we observe in our sample may not actually exist in the Data Generating Process (represented at the top of Figure 1); it may just be a result of sampling variation.

At this point, we bring in another of our core concepts, the concept of randomness as a process. Students use the shuffle() function in R (part of the mosaic package; Pruim, Kaplan, & Horton, 2017) to break the relationship between Sex and Thumb length, generating a new data set in which values for the two variables are randomly paired. They examine the results of these randomized data sets to explore how different randomly generated samples could look from one another. Finally, they compare the actual data (e.g., histograms of male and female thumb lengths) to the randomly generated histograms (e.g., thumb lengths that have been shuffled between two groups), and try to judge whether the actual data stands out from, or fits right into, the randomly varying samples. In Figure 1, we represent these shuffled samples in the bottom right corner, the same place that sampling distributions will go when they are formally introduced later in the course. Even though students have not fit models or calculated anything at this point in the textbook, students begin to explore the role of randomness in figuring out which DGPs could have produced our sample.

Using intuitive and informal ideas as a foundation on which to later build understanding of formal statistical concepts is a feature of our pedagogy that we carry throughout the book. Even though students have not calculated any statistics or fit any models yet at this early stage in the course, we already are introducing students to core concepts such as *model* and *sampling variation.* In this way, we are helping students to prepare conceptual "slots" in which to put the more abstract and formal ideas that come later in the book. They also will have started to become aware of a problem for which formal models will ultimately provide a solution—the evaluation of one sample distribution in the context of multiple randomly generated distributions.

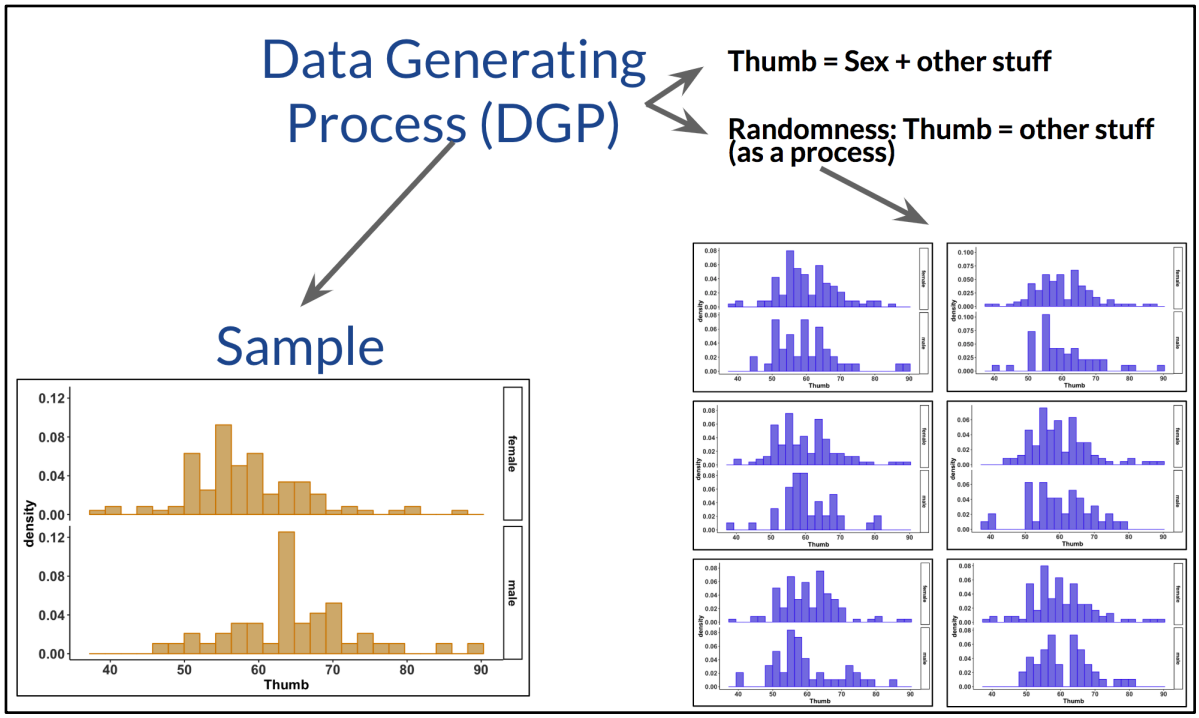

Figure 1. Sample R plots students created to explore the distribution triad. When exploring variation, students focus on connecting two types of distributions (sample and DGP), representing possible explanations with informal models instantiated as word equations, and exploring those explanations with data visualizations (e.g., faceted histograms). They further explore randomness as a DGP by actually generating data with a random process (using an R function, *shuffle*) which breaks any relationship between the explanatory variable (Sex) and the outcome (Thumb) by shuffling which Thumb length goes with which Sex group. The histograms in blue visualize six independent examples of shuffled data. Does the empirical sample look similar to these shuffled samples? Considering this question puts students on the path to understand the logic of sampling distributions, far ahead of their formal introduction.

Modeling Variation

In the second part of the book, we move from exploring variation to modeling variation. This part of the book (Chapters 5-8) takes about a week per chapter, with more time spent on Chapter 7 (1-2 weeks). In Chapter 5, we introduce the formal concept of statistical model, starting with the mean of a quantitative variable, a model often referred to as the *empty model*. Interestingly, while most books teach the mean as one of a number of descriptive statistics, we introduce it as a model. Prior to introducing the mean as a model, we do not have students calculate any statistics (apart from those required for the five-number summary).

Just as before, although our focus in this part is on constructing models from data, we always keep in the forefront the idea that what we are trying to model is the DGP that produced the data (depicted at the top of Figure 2). Thus, we introduce the terms *statistic* and *parameter*, stressing that the arithmetic mean we calculate from our data is a statistic, and that statistics

are used as imperfect estimates of parameters. Parameters must be estimated because, as we have seen, we have no way to directly measure the DGP.

We explore the properties of the mean in the context of "the mean as a model" rather than more traditionally as a measure of central tendency. We advance the idea that the mean is often the best estimate we have of the actual population mean (the mean of the DGP), and the best predictor of the value of a subsequent observation. Because the prediction is almost certainly wrong, it is natural to introduce the concept of residual as a way to calculate, for each data point, how far off the model prediction is. This emphasizes the mean's special property — that it uniquely balances the negative and positive residuals. In Chapter 6, we then introduce statistics such as sum of squares, standard deviation, and variance as ways of aggregating the residuals to indicate how much total error there is around our model.

In the context of this simple model, we begin to develop in very concrete terms the basic idea behind statistical modeling: DATA = MODEL + ERROR. In the distribution (of data), each score can be expressed as a combination of two components: the mean (as a model) and a deviation from the mean (error). In a specific context, we can present it as **Thumb = mean + error** (a developmental step beyond our previous formulation, **Thumb = other stuff**). We can represent this idea using simple mathematical notation, the notation of the General Linear Model (e.g., $Y_i = b_0 + e_i$). We connect these representations to the idea of the empty model, one with no explanatory variable. This is further supported by the R code, lm(Thumb ~ NULL, data = Fingers). We also use GLM notation to distinguish the model fit to the sample from the empty model of the DGP (e.g., $Y_i = \beta_0 + \epsilon_i$ in the DGP part of Figure 2).

The sum of squares (SS) as a measure of error is related to the larger goal of explaining variation in some outcome variable, or, in a complementary fashion, reducing error variation. The mean gives us a place to start because the error measured as SS has already been reduced as much as possible. Any variation due to explanatory variables is still part of the error term in this simplest of models. Adding explanatory variables into the model can reduce this error. Although the models we explore next are more complex than the empty model, they are still quite simple, adding in a single explanatory variable. We generally call them "complex models" to distinguish them, relatively, from the simpler, empty model.

In Chapter 7, we add a grouping variable to our statistical model (e.g., **Thumb = Sex + error**). Using the notation of the General Linear Model, we represent this new model as the mean of one group (b_0) plus the increment to be added to get the mean of the other group (b_1). Thus, students learn to connect the GLM equation $Y_i = b_0 + b_1 X_i + e_i$ to a two-group model where X_i represents whether a case (i) belongs to the second group or not (coded 1 or 0, respectively). For example, each student's thumb length is expressed as the mean of the female group plus the increment to be added to get the mean of the male group multiplied by whether the student is in the male group (1) or not (0).

The error term (e_i) is defined as the residual calculated by subtracting each individual's actual score from their predicted score (in this case as the difference between each individual's score and their own group mean). This development continues to emphasize how each data point can be concretely partitioned into two parts that can be added together: MODEL (e.g., their group mean) and ERROR (their deviation from the group mean). Similarly, SS Total (based on the sum

of the squared deviations of each data point from the empty model's predictions) can be partitioned into SS Model (based on the deviations of the complex model's predictions from the empty model's predictions) and SS Error (based on the deviations of the data from the complex model's predictions). More complex models reduce the error relative to simpler models, a reduction we can measure using PRE (Proportional Reduction in Error = (SS Total - SS Error) / SS Total or SS Model / SS Total). Error is not just random, however; it also includes variation due to additional as-yet-unmeasured explanatory variables. The development of error (i.e., SS and PRE) are similar to treatments in other textbooks with a focus on modeling (Judd, McClelland, & Ryan, 2017; Kaplan, 2017).

Although we can reduce error by making more complex models, we also sacrifice degrees of freedom. Degrees of freedom is, in a sense, the currency of statistical power. We "earn" more degrees of freedom by having a larger sample but "spend" degrees of freedom every time parameters are added to a model. We introduce the F statistic as a measure of PRE that takes number of parameters per degree of freedom into account (see Judd, McClelland & Ryan, pp. 52-56). In our simple one-parameter models, *F* can be thought of as the ratio of PRE/1 to (1 - PRE)/(n-1). In other words, it is the PRE obtained per model parameter (1 for our simple models) divided by the average PRE that could be obtained by adding all possible remaining parameters (i.e., if there were one parameter added for every remaining degree of freedom).

Once we have developed a model with a grouping variable as the explanatory variable, it is straightforward to follow the same approach in Chapter 8 to building models that have a quantitative (as opposed to categorical) explanatory variable. It is here that our connections to core concepts and representations begin to pay off. Whereas in the traditional course students are led to see ANOVA (including concepts such as F ratio) and regression as two separate topics, we show them that both types of models can be represented by the same GLM equation. All that has changed is the interpretation of the two parameter estimates, going from the mean of group one (i.e., the prediction when X_i is 0) and the increment from group one to two (i.e., the increment to add when X_i increases by 1), to the y-intercept of the best-fitting line (i.e., the prediction when X_i is 0) and the slope that defines how much is added for each unit increase in X_i.

Figure 2 depicts how the same structure built up in the "exploring variation" phase of the course is reprised once we have been able to fit models and measure error quantitatively. We return to the broader goal of attempting to understand the DGP that gave rise to our empirical sample. This time, instead of using the shuffle() function and looking at the resulting visualizations qualitatively, we can also compute, for example, the best fitting estimates of the mean differences (i.e., b_1) in each of our shuffled distributions. Now we can look at whether our sample b_1 is similar at all to our randomly generated $b_1$values. We can also reconceptualize the empty model of the DGP as one where $\beta_1 = 0$ and thus notice why $b_1$values generated from the "null" DGP will cluster around 0 (see bottom right of Figure 2).

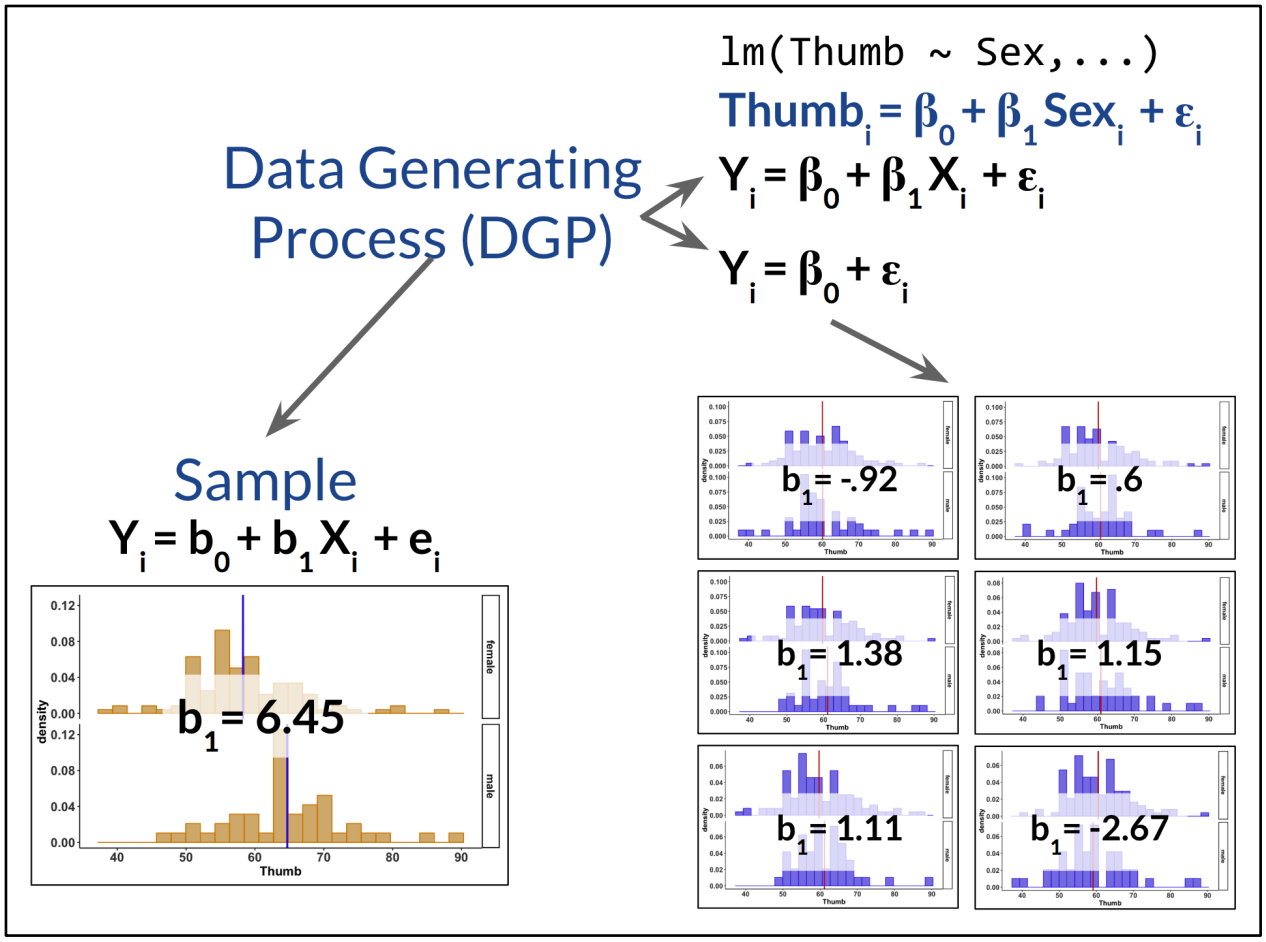

Figure 2. When "modeling variation," concepts explored previously are now formally modeled with the General Linear Model. We have overlaid the best fitting sample statistics on top of the R plots students created as they added a quantitative component to their understanding of models in the distribution triad. Students connect their word equations and R code for visualizations with new R code for model fitting, GLM notation, and the best fitting estimates.

Evaluating Models

Finally, we move from modeling variation to evaluation and comparison of statistical models. This part of the book (Chapters 9-11) usually takes more time, about 1-2 weeks per chapter. (Note: There is a Chapter 12 but it is a brief review of all the concepts of the whole course in a new example context.) By Chapter 9, students have practiced exploring informal models of variation, fitting models to data, and evaluating how well the models fit the data by looking at the proportion of error reduced by a complex model compared with the empty model (PRE). But how well do our models fit the DGP? How accurate are the parameter estimates we compute based on data? And most important, when we compare two models (such as a group model versus the empty model), how do we decide which one best represents the DGP? That

is, how do we know whether the model that uses Sex to predict Thumb length is better than the one that does not?

Answering questions such as these requires a journey into the third realm of the distribution triad, sampling distributions. Clearly, if we had studied a different sample we would have come up with slightly different parameter estimates (b_0 or b_1), and different measures of fit, such as PRE or F. Figure 3 shows that the mean difference between male and female thumb lengths in the sample was 6.56 mm, but a different sample would have a different best fitting value for b_1. Sampling distributions, which exist in our imagination, are the distributions from which our parameter estimates are drawn. Just as interpreting a single score requires us to know about the distribution from which it came, interpreting a statistic (such as a parameter estimate) requires us to know something about the distribution from which it comes.

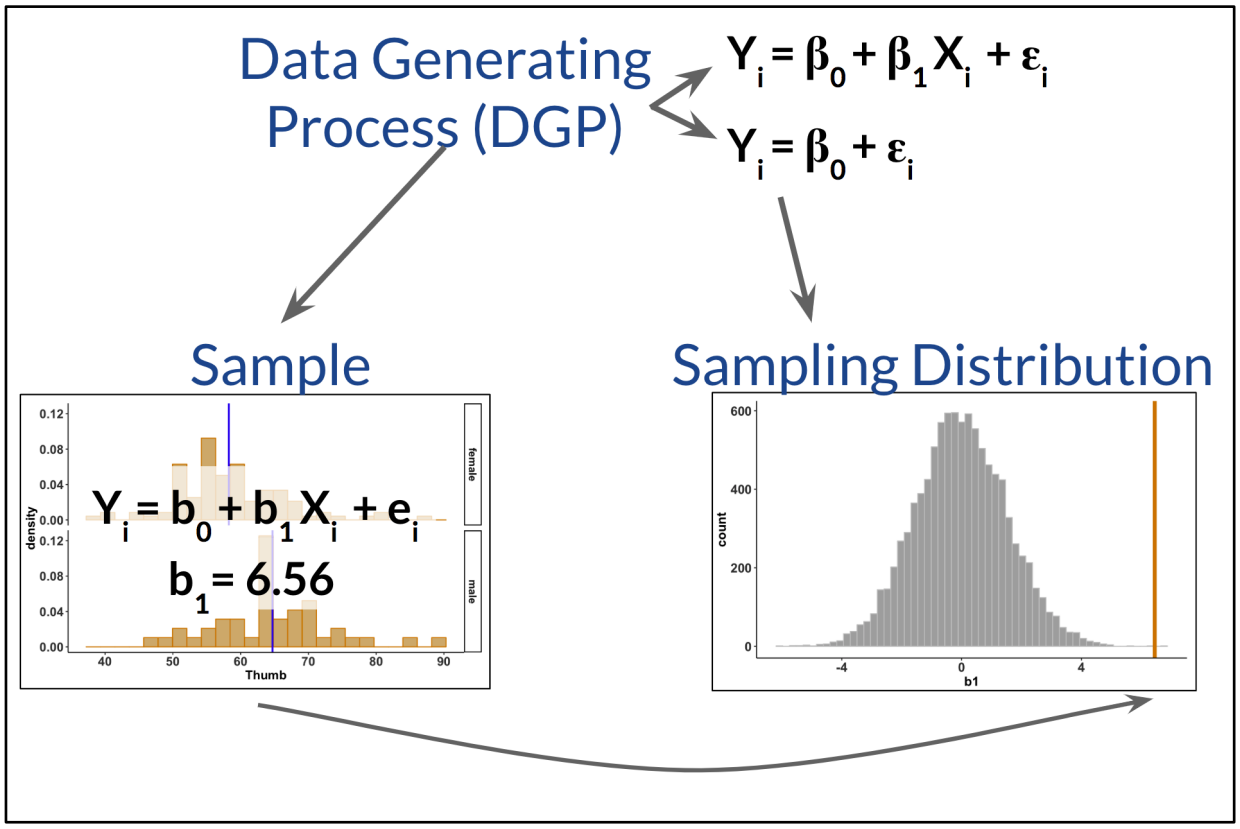

Figure 3. Evaluating models. When "evaluating models," students go beyond the best fitting parameter estimates in their models. They consider how those parameter estimates could potentially vary depending on different DGPs. Students connect the shuffling (randomization) and resampling (bootstrapping) they have done with building up sampling distributions of particular parameter estimates such as the b_1 values (pictured here), or PRE or F.

In this section of the course, students learn to explore the variability among samples. They use techniques such as simulation, bootstrapping, and randomization to generate many random samples from a hypothesized DGP, and use these samples to create sampling distributions of parameter estimates. Figure 3 shows a sampling distribution of 10,000 mean differences generated by randomization (a combination of the do() and shuffle() functions, Pruim, Kaplan, & Horton, 2017). Students can see that the empty model of the DGP does not always produce a sample with a mean difference of 0. The standard deviation of a sampling distribution (or Standard Error) allows us to reason about our parameter estimates using logic like this: If the DGP has the mean, variance, and shape we assumed in our simulation, then how likely is it that we would get a random sample with an estimate as extreme (or more extreme) as one observed in our data? They can also concretely tally up the number of mean differences in their sampling distribution that are more extreme than the difference observed in their sample.

We start by thinking about the distribution of a statistic (such as $b_0$or b_1) and note that any parameter we estimate based on a sample can be thought of as coming from a sampling distribution. We could construct a sampling distribution of PRE or of the F ratio based on the empty model (null hypothesis) and then use that sampling distribution to calculate the probability of getting our data (i.e., $p < .05$) *if the empty model is true*. (Even if the empty model were true, simulations reveal that there would still be variation in Fs such that some samples would have quite high F ratios just by chance.) We can examine the F ratio calculated from our data against simulated, bootstrapped, randomized, and mathematically modeled sampling distributions of F. By using PRE and F to compare a complex model with the empty model, we are preparing students to compare two complex models that differ from each other when neither is the empty model.

Evidence of Understanding: A Pilot Study

We are only at the beginning of what we see as a long-term project. However, given that our goal is for our students to develop a deep understanding of the domain, how would we measure such an outcome? In our practicing connections approach, we are investing a lot of time in developing interconnections among the bits of knowledge that make up the traditional statistics curriculum. We do this because coherent and interconnected knowledge structures should be more flexible and transferable. Do we see any evidence of this?

We conducted a pilot study to examine transfer. There are many ways to assess transfer; indeed, each time we present students with a new data set we are asking them to extend what they have learned in one context to apply in another. Our focus in this study is on preparation for future learning (Bransford & Schwartz, 1999). If students come to understand statistics in a deeper way, and are able to connect what they are learning to core concepts of the domain, then they should be better positioned to learn new statistical techniques. In fact, this is one of the goals we have for our course—that students will be better prepared for more advanced courses.

To assess students' preparation for future learning, we added assessment questions on the final exam that asked students to transfer what they had learned about ANOVA and simple regression to an example of multiple regression with two explanatory variables, one quantitative and one categorical.

We added the multiple regression transfer questions to the final exams of two large introductory statistics classes taught in the psychology department at UCLA, one in the fall of 2018, the other in the winter of 2019. The 10-week classes, which met for four hours each week, were taught by two different instructors with different instructional styles and classroom pedagogies, but both using our new online textbook. There were 265 students that completed the Fall course, and 209 that completed the Winter course.

UCLA is a fairly diverse, though highly selective university. Most students in the course were psychology majors. Seventy-three percent of students identified as female, 26% identified as male, and 1% identified as non-binary or did not answer. When asked to choose which races/ethnicities they identify as, 3% chose *African-American*, 28% chose *White*, 39% chose *Asian*, 20% chose *Latinx*, and 10% chose *Other* or did not respond (the sum is coincidentally 100%, but students were allowed to choose multiple options). Most students were in their second (53%) or third (39%) year of college (first year: 5%; fourth or greater year: 3%). Student ages were not collected.

Students were told that they would get extra credit for completing this section of the exam, but that if they needed the time to complete the main part of the final exam, they would be given the extra credit anyway. We explained to students that we were interested to see if they could figure out what would happen if a second predictor were added to a model. Of the 474 students who completed the course, all but 45 completed the transfer questions (38 in the Fall, 7 in the Winter).

Materials and Procedure

For both classes, students took an in-class comprehensive final exam on an electronic device (e.g., a phone, tablet, or laptop). The two exams covered the same content, were based on the same dataset (which students had not previously encountered), and included very similar questions. The transfer questions were identical across the two exams.

Both the exam and the transfer questions used the candy_rankings dataset, available through the fivethirtyeight R package (Hickey, 2017; Kim, Ismay, & Chunn, 2018). The dataset was based on a survey in which people were shown different pairs of common Halloween candies and asked to choose the candy they liked best (the "winner"). Each of the 85 rows in the dataset was a type of candy (e.g., KitKat). Variables included **winpercent** (the percentage of matchups in which the target candy was selected), **sugarpercent** (the sugar content of the candy), **chocolate** (whether or not the candy included chocolate), and others.

In the preamble to the transfer questions, students were shown the (now familiar) output of the GLM analyses for two separate models of winpercent: the chocolate model (i.e., $winpercent_i = b_0 + b_1 chocolate_i$) and the sugar model (i.e., $winpercent_i = b_0 + b_1 sugarpercent_i$). The outputs, shown in Figure 4, included both the parameter estimates and ANOVA tables. The ANOVA tables were created in the style of Judd, McClelland, and Ryan (2017) using the supernova function from the supernova R package (Blake, Chrabaszcz, Son, & Stigler, 2019). The tables include a column for proportional reduction in error (*PRE)* instead of the more traditional η^2 or R^2.

```
Call:
lm(formula = winpercent ~ chocolate, data = candy_rankings)

Coefficients:
  (Intercept)  chocolateTRUE
         42.1           18.8
```

Chocolate model

```
Analysis of Variance Table (Type I SS)
Model: winpercent ~ chocolate

                              SS df       MS      F    PRE     p
----- ----------------- --------- -- -------- ------ ------ -----
Model (error reduced) |  7368.542  1 7368.542 56.532 0.4052 .0000
Error (from model)    | 10818.492 83  130.343
----- ----------------- --------- -- -------- ------ ------ -----
Total (empty model)   | 18187.034 84  216.512
```

```
Call:
lm(formula = winpercent ~ sugarpercent, data = candy_rankings)

Coefficients:
 (Intercept)  sugarpercent
      44.609         0.119
```

Sugar model

```
Analysis of Variance Table (Type I SS)
Model: winpercent ~ sugarpercent

                              SS df      MS     F    PRE     p
----- ----------------- --------- -- ------- ----- ------ -----
Model (error reduced) |   955.002  1 955.002 4.600 0.0525 .0349
Error (from model)    | 17232.033 83 207.615
----- ----------------- --------- -- ------- ----- ------ -----
Total (empty model)   | 18187.034 84 216.512
```

Figure 4. Model fits and ANOVA tables for the chocolate model (top) and the sugar model (bottom) of winpercent.

After being presented with the outputs shown in Figure 4, students were asked a series of questions. We present these questions below, along with a summary of students' responses.

Questions and Results

Question 1—Inventing GLM notation. Question 1 asked students to consider the possibility of creating a single model that included both chocolate and sugarpercent as predictors, and try to "represent the two--predictor model in GLM notation" using regular

letters (e.g., b1 for b_1, or Xi for X_i). They typed their responses into a standard HTML input box.

Fifty-three percent of the students gave completely correct responses. However, even students who were not completely correct nevertheless included many correct features (see Table 1). In Table 1 we show the proportion of students who included a single outcome variable, a single intercept, two unique predictor coefficients, two unique predictor variables, and a single error term. Most of these features were included by more than 90% of students.

Table 1

The Proportion of Students Including Each Feature in their GLM Notation Response

Feature	**Example**	**Propor-tion**
Completely correct	yi = b0 + b1*X1i + b2* X2i + ei yi = b0 + b1X1i + b2X2i + ei yi = b0 + b1(X1i) + b2(X2i) + ei win_i = b0 + b1*chocolate_i + b2* sugar_i + ei	**0.536**
A single outcome variable	yi = b0 + b1*X1i + b2* X2i + ei *Not: b0 + b1*X1i + b2* X2i + ei*	0.993
A single intercept	yi = b0 + b1*Xi + b2*Xi + ei	0.995
No intercept	yi = b1*X1i + b2*X2i + ei	0.005
Two intercepts	yi = b0 + b1*X1i + b2 + b3* X2i + ei	0.072
Two unique predictor coefficients	yi = b1*Xi + b2*Xi *Not: yi = b1*Xi + b1*Xi*	0.812

Two unique predictor variables	b1*X1i + b1*X2i	0.599
Only used one predictor	yi = b0 + b1*Xi + b2 + e	0.316
Used the same predictor twice	yi = b1*Xi + b2*Xi	0.246
Used more than two predictors	yi = b0 + b1*X1i + b2*X2i + b3* X3i + ei	0.014
A single error term	yi = b0 + ei *Not: yi = b0*	0.949

Note: Correct features are shaded grey in the feature column.

Questions 2-4—Reasoning about PRE, SST, and SSE in the new model. Students were next asked three multiple-choice questions in which they made predictions about PRE, SST, and SSE for the new two-variable model. Questions and a summary of responses are shown in Table 2.

Table 2

Questions, Response Options, and Proportion of Students Choosing Each Option for Questions 2, 3, and 4

Question	**Option 1**	**Option 2**	**Option 3**	**Option 4**
What do you expect the PRE to be for the new model?	Higher than either model	Between the models	Lower than either model	Can't tell
	0.822	0.089	0.039	0.050
What do you expect the SS Error to be for	Higher than either model	The same as the chocolate model	The same as the sugar model	Lower than either model

the new model?	0.121	0.018	0.016	0.84
What do you expect the SS Total to be for the new model?	Higher than either model	The same as both models	Lower than either model	Can't tell
	0.153	0.770	0.059	0.018

Note: Options are presented left-to-right in this table in the same order that they were shown to students. Correct answers and the proportion of students who answered correctly are shaded grey.

Students performed very well on these questions, with the vast majority indicating, correctly, that *PRE* for the combined model should be greater than the single-predictor models, the *SSE* should be lower, and the *SST* should not change. See Table 2 for more details. A chi-square test of goodness-of-fit was performed for each question to determine whether the patterns of responding were different from chance (0.25). Preference for the options was not equally distributed for PRE [$\chi 2$ (3, N = 456) = 842, p < .001], SSE [$\chi 2$ (3, N = 458) = 694, p < .001], nor SST [$\chi 2$ (3, N = 458) = 909, p < .001].

From these data, it is clear that a majority of students have at least a broad understanding of how adding predictors to a model should affect key measures of error. This is a notable achievement given that models with more than one explanatory variable were not covered at all in the course.

Question 5—Why is the SSE for the new model smaller? After answering the previous questions, students moved on to a new page on which they were informed that, in fact, the SSE is lower for the two-predictor model than for either of the single-predictor models. (They were prevented from navigating back to modify their answer to the prior question.) Given this new information, students were asked in an open-response question, why is the SSE for the two-predictor model smaller than the SSEs of the other two models?

Although the answers varied in their focus and in their completeness, the majority of students expressed the main idea we were going for, relating SST, SSM, and SSE to the core concept of DATA = MODEL + ERROR. Seventy percent of students were able to explain that SSE is lower because the complex model explains more error. This was also the most common idea included in the responses. Here are two examples:

> Because with two explanatory variables that both account for some of the variation from the empty model the SS model would be larger than both the sugar percent model and the chocolate model. With a larger SS model and the same SS total the SS error is going to be smaller.

> The total SS is equal to the Model and Error SS. By adding more parameters, we can increase the amount of variation explained, meaning that we can increase the SS Model. Since SS Total will stay the same, this then means that SS Error must decrease.

Most incorrect answers were incomplete because they failed to reference the zero-sum relationship between MODEL and ERROR. For example, a student wrote, "Because the two-predictor model explains more error, there is less leftover error after using the two predictor model." Note that although this response is largely correct, it is incomplete because it does not explicitly note that this is true because the SST is the same for both models. A smaller proportion of incorrect answers included misconceptions (e.g., "combining both… makes the sample larger…"). For example, one student wrote, "It should have a lower SS error as you are combining two explanatory variables together, the amount of error explained by the SS model would add up together and increase, thus reducing SS error." Again, there are some correct ideas in this response but it implies (incorrectly) that the SS model for the two-predictor model is the sum of the SS models for each single predictor model.

Some students gave more sophisticated answers. For example, one student wrote that SSE would be smaller because "it is measuring the distance from the data to the complex model, which should get smaller as more complexity is added."

Question 6—Making predictions with the new model. To test their understanding of model predictions, we gave students the parameter estimates and corresponding ANOVA table for the new two-predictor model (see Figure 5) and asked them, on a multiple-choice question, to predict winpercent for a candy that contains chocolate.

```
Call:
lm(formula = winpercent ~ chocolate + sugarpercent, data = candy_rankings)

Coefficients:
  (Intercept)  chocolateTRUE   sugarpercent
      38.2621        18.2733         0.0857
```

```
Analysis of Variance Table (Type I SS)
Model: winpercent ~ chocolate + sugarpercent

                                    SS df       MS      F    PRE     p
------------ ----------------- --------- -- -------- ------ ------ -----
       Model (error reduced) |  7856.127  2 3928.064 31.178 0.4320 .0000
   chocolate                 |  7368.542  1 7368.542 58.487 0.4163 .0000
sugarpercent                 |   487.585  1  487.585  3.870 0.0451 .0525
       Error (from model)    | 10330.907 82  125.987
------------ ----------------- --------- -- -------- ------ ------ -----
       Total (empty model)   | 18187.034 84  216.512
```

Figure 5. The code, model estimates, and ANOVA table for the linear model that uses chocolate and sugarpercent to predict winpercent.

This was a difficult transfer question for a few reasons. First, although the model included sugar percentage, the question does not explicitly mention sugar. Second, in the ANOVA table, sugar percentage is not a significant predictor (evidenced by the *p* value), which may lead students to discount it. Third, the correct answer was an expression ($38.26 + 18.27 + 0.0857 * sugarpercent$) rather than a single value. Because the model includes sugarpercent and students were not given a specific value for sugar percentage, the correct answer needed to include sugar percentage as a variable. Students had little experience with predictions left as algebraic expressions in the course and were never shown the correct GLM equation for the best fitting two-predictor model. Students were primarily working from their own intuition about what that would be.

Answer options, and the proportion of students selecting each, are shown in Table 3. Despite the intentionally difficult features of this question, the correct answer was the most popular answer. Almost equally popular was Option 1, which completely omitted sugarpercent.

Because we asked this as a multiple-choice question, the thinking of the students who chose Option 1 is unclear. For example, some students may have omitted sugarpercent because it was a non-significant predictor, whereas others may not have included sugarpercent because it was not explicitly mentioned in the question. The popularity of Option 1 suggests that students did not realize that this option (38.26 + 18.27) predicts winpercent when sugarpercent is 0 (38.26 + 18.27 + 0). It would be interesting to probe further about when they would use the two-predictor model versus the chocolate model alone (from Figure 4, 42.1 + 18.8) to make predictions.

Table 3

Question Prompt, Response Options, and Proportion of Students Choosing Each Option for the Model Prediction Question

Question	**Option 1**	**Option 2**	**Option 3**	**Option 4**
What would you predict the win percent to be for a candy that has chocolate?	38.26 + 18.27	38.26 + 18.27 + 0.0857	38.26 + 18.27 + 0.0857*sugarpercent	Can't tell
	0.442	.068	0.478	0.011

Note: Options are presented left-to-right in this table in the same order that they were shown to all students. Option 3 (shaded in grey) was the correct answer.

Question 7—Speculating on why order of entry (chocolate versus sugarpercent) matters. In Question 7, students were told that we re-ran the two-predictor model, but this time reversed the order of the two predictors, putting sugarpercent in first followed by chocolate. They were asked to compare the two outputs (Figure 5 and Figure 6) and to explain why the values in the ANOVA tables might be different.

```
Call:
lm(formula = winpercent ~ sugarpercent + chocolate, data = candy_rankings)

Coefficients:
  (Intercept)   sugarpercent  chocolateTRUE
      38.2621         0.0857        18.2733
```

```
Analysis of Variance Table (Type I SS)
Model: winpercent ~ sugarpercent + chocolate

                                    SS df       MS      F    PRE     p
----------- ----------------- --------- -- -------- ------ ------ -----
       Model (error reduced) |  7856.127  2 3928.064 31.178 0.4320 .0000
sugarpercent                 |   955.002  1  955.002  7.580 0.0846 .0073
   chocolate                 |  6901.126  1 6901.126 54.777 0.4005 .0000
       Error (from model)    | 10330.907 82  125.987
----------- ----------------- --------- -- -------- ------ ------ -----
       Total (empty model)   | 18187.034 84  216.512
```

Figure 6. The code, model estimates, and ANOVA table for the two-predictor model of winpercent in which the order of entry of the two predictors is reversed (with sugarpercent now entered before chocolate**)**. Note that the table shows Type I sums of squares.

This is a challenging question because students would have to invent the concept of Type I SS having only learned about SS in situations with a single explanatory variable. Type I SS are calculated by adding terms into a model sequentially. The first variable will "explain" all the variance that is attributable to it alone *plus* the variance that could also be explained by the second variable (assuming there is some shared variance between the two variables). The second variable will only explain the variance that is attributable to it alone, after subtracting out the variance explained by the first variable.

In this case, the chocolate and sugar percentage variables are correlated (r = ~0.1) so when chocolate is included in the model first, the Type I SS for chocolate includes the variation explained by chocolate uniquely as well as the variation that could also be explained by sugarpercent. Thus, when chocolate is included first, the SS for chocolate is greater than when it is included second.

Some of the students' explanations were quite impressive. For example, one student wrote:

> Since chocolate and sugarpercent are related (amount of chocolate predicts some amount of sugarpercent, and vice versa), the order in which each model is run matters. When running the chocolate model second, there is less variation left because of the sugarpercent model, so the chocolate model explains less variation in this case. Running the chocolate model first would give it more variation to explain away.

Another wrote:

> The variable that goes first explains more variation than if it were to go second. The logic behind this is that the second variable attempts to explain the leftover variation unexplained by the first variable. In other terms, both variables overlap in their process of explaining variation and whichever variable goes first will account for explaining the variation contained in the overlap.

Examination of students' responses revealed four different ideas that we judged to be correct. These ideas (along with the proportion of students who mentioned each) were:

- The first variable gets more of the variation (.308)
- The second variable explains the "left over" variation (.265)
- The variation that is available to be explained is changing during the process (.187)
- The variables overlap in what they can explain (.097).

Importantly, 40.5% of students mentioned at least one of these correct ideas—quite impressive given that multiple regression, not to mention the order of entry of variables into a regression analysis, was never mentioned in the course.

Nevertheless, this question was challenging, with 53.6% giving only incorrect answers (e.g., "the difference is due to sampling variation") or answers deemed irrelevant to the question being asked (e.g., "one variable is better at explaining"), and 5.9% either not attempting to answer or giving answers indicating that they did not understand what the question was asking (e.g., "because they are different variables").

Discussion

In this paper, we have set forth the rationale behind our *practicing connections* approach to teaching introductory statistics, and we have given the reader a guided tour through our interactive online textbook, *Introduction to Statistics: A Modeling Approach*. Our goal throughout is to help students build a coherent representation of the domain of statistics by continually making connections between core concepts, representations, and situations in the world.

The concepts and representations we chose to emphasize—the concept of statistical model, notation of the General Linear Model, and the computational representation of statistical

concepts using R—connect the content of our course with current developments in statistical reasoning. (We are not the first to attempt this approach; see Kaplan, 2017). Although we cover topics such as null hypothesis testing, we try to put all such topics into a model comparison context.

Our pilot data provided encouraging examples of how our sample of mostly non-STEM students managed to apply simple ideas of modeling introduced in the class to the task of reasoning about more complex models that they have not seen before. Instructors teaching with our book in community colleges and regional state universities report that at least some students who had not thought of themselves as being able to learn programming or as being a "math person" engaged successfully with core concepts of modeling.

Although more rigorous data collection is necessary, our current speculation is that the novelty of what students are asked to learn may support the success of a greater variety of students. Whether students have strong mathematical preparation or not, virtually none of our students have ever learned R. Even if some students have taken AP Statistics in high school, almost none of our students have been exposed to the unifying concept of statistical model, or to computational techniques such as simulation, randomization, and bootstrapping. These factors may put students with diverse backgrounds and levels of preparation on more of an even footing. One future direction we are excited to undertake is to more thoroughly assess our students' abilities to differentiate and selectively apply some of the more sophisticated techniques and strategies, such as simulation, bootstrapping, and randomization, after completing just the introductory statistics course.

Of course, we are only at the beginning of our project, and many challenges remain. Although the textbook is freely available, there is great variety in how individual instructors in different institutions implement the course. We do not yet know the best ways to implement the course materials for a large lecture course or a more intimate course; for a general audience or a major specific audience (e.g., statistics taught in economics); for students at a community college or a highly competitive institution; for a course taught with face-to-face meetings or fully online (synchronously or asynchronously). For this reason, we invite interested readers to join our networked improvement community working to improve the online textbook and its implementation (see Stigler et al., in press). Readers can preview the current version of the book at CourseKata.org. Instructors can sign up to teach using the online book (available at no charge), sharing important data back with the core research and development team. Researchers can join to conduct research to increase our understanding of the teaching and learning of statistics while at the same time improving the quality of the book.

Although we find our transfer results to be encouraging, the reader may wonder why we haven't compared our students to a control group receiving more traditional instruction. In fact, we want to do that. Given how different our course is from more traditional courses, however, we have struggled to find the best ways to measure transfer that would be valid for both conditions. For example, we want our students to cast increasingly complex situations into a model comparison framework, even though it would be meaningless to test the traditional student on concepts related to model comparison. Still, we want to try to develop valid transfer measures in future studies that can be used across different courses. We also want to examine how transfer relates to other important measures of learning such as attitudes

and motivation (e.g., Schau et al., 1995) and well-studied measures of basic statistics knowledge (delMas, Garfield, Ooms, & Chance, 2007; Whitaker, Foti, & Jacobbe, 2015).

At the same time, we want to stress the importance of improvement research as an important approach to education research and development (c.f., Bryk, Gomez, Grunow, & LeMahieu, 2015.). Too often, innovative programs are compared to traditional ones prematurely, before the potential of the innovation can be fully tested. We believe that it is important to understand variation within our course first before trying to look for average differences between our course and more traditional ones.

At this stage in our project, our goal is not to determine whether or not our approach is effective, but instead what it would take to make the approach effective for a diversity of students. This question, which we view as an important stage in the development of education innovations, is not one we can answer by ourselves. We will need the collaboration of instructors, their students, talented curriculum designers and developers, and other researchers to realize the potential of our approach.

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature, 567*, 305-307. doi: 10.1038/d41586-019-00857-9

Bassok, M., Wu, L. L., & Olseth, K. L. (1995). Judging a book by its cover: Interpretive effects of content on problem-solving transfer. Memory and Cognition, 23, 354–367.

Blake, A. B., Chrabaszcz, J. R., Son, J. Y., & Stigler, J. W. (2019). Supernova. Retrieved from https://github.com/UCLATALL/supernova

Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24*(1), 61-100.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better.* Cambridge, MA: Harvard Education Press.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121-152.

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum?. *Technology innovations in statistics education*, *1*(1).

Cohen, J. (1994). The earth is round (p<. 05). In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 69-82). Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7-29.

delMas, R., Garfield, J., Ooms, A., & Chance, B., (2007). Assessing Students' Conceptual Understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28-58.

Fries, L., Son, J. Y., Givvin, K. B., & Stigler, J. W. (2020). The practicing connections hypothesis: A framework to guide instructional design in complex domains. *Educational Psychology Review. https://doi.org/10.1007/s10648-020-09561-x.*

Geller, E. H., Son, J. Y., & Stigler, J. W. (2017). Conceptual explanations and understanding fraction comparisons. *Learning and Instruction*, *52*, 122-129.

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198-218.

Givvin, K. B., Stigler, J. W., & Thompson, B. J. (2011). What community college developmental mathematics students understand about mathematics, Part II: The interviews. *The MathAMATYC Educator*, *2*(3), 4-18.

Hesterberg, T. (2015). What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum, *The American Statistician, 69*(4), 371-386. DOI: 10.1080/00031305.2015.1089789

Hickey, W. (2017, October 27). The Ultimate Halloween Candy Power Ranking. Retrieved July 12, 2019, from https://fivethirtyeight.com/features/the-ultimate-halloween-candy-power-ranking/

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, ANOVA, and beyond* (3rd ed.). New York: Routledge.

Kaplan, D. (2017). *Statistical modeling: A fresh approach (2nd Edition)*. Project MOSAIC books.

Kaplan, D., & Pruim, R. (2019). ggformula: Formula Interface to the Grammar of Graphics. R package version 0.9.2. https://CRAN.R-project.org/package=ggformula

Kim, A. Y., Ismay, C., & Chunn, J. (2018). The fivethirtyeight R Package: 'Tame Data' Principles for Introductory Statistics and Data Science Courses. *Technology Innovations in Statistics Education, 11*.

Lock, R. (2017). Lock5Data: Datasets for "Statistics: UnLocking the Power of Data. R package version 2.8. https://CRAN.R-project.org/package=Lock5Data

McGowen, M. A., & Tall, D. O. (2010). Metaphor or Met-Before? The effects of previous experience on practice and theory of learning mathematics. *The Journal of Mathematical Behavior*, *29*(3), 169-179.

Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General*, *131*(1), 5.

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures.* National Academies Press.

Nolan, D., & Temple Lang, D. (2010). Computing in the statistics curricula. The American Statistician, 64(2), 97-107.

Pruim, R., Kaplan, D.T., & Horton, N.J. (2017). The mosaic package: Helping students to think with data using R *The R Journal, 9,* 77-102.

Sagi, E., Gentner, D., & Lovett, A. (2012). What difference reveals about similarity. *Cognitive Science*, *36*(6), 1019-1050.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: a quiet methodological revolution. American Psychologist, 65(1), 1.

Son, J. Y., Ramos, P., DeWolf, M., Loftus, W., & Stigler, J. W. (2018). Exploring the practicing-connections hypothesis: Using gesture to support coordination of ideas in understanding a complex statistical concept. *Cognitive Research: Principles and Implications*, *3*(1), 1.

Son, J. Y., Smith, L. B., & Goldstone, R. L. (2011). Connecting instances to promote children's relational reasoning. *Journal of Experimental Child Psychology*, *108*(2), 260-277.

Son, J. Y., & Stigler, J. W. (2017-19). *Introductory statistics: A modeling approach.* Retrieved from https://coursekata.org/preview/default/program.

Stigler, J. W., Son, J. Y., Givvin, K. B., Blake, A., Fries, L. C., Shaw, S. T., & Tucker, M. C. (in press). The *better book* approach for education research and development. *Teachers College Record, 123*(2).

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–295.

Tintle, N. L., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2014b). "Quantitative Evidence for the Use Simulation and Randomization in the Introductory Statistics Course," in *Proceedings of the Ninth International Conference on Teaching Statistics*, volume ICOTS-9. Available at http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8A3_TINTLE.pdf

Tucker, M., Shaw, S. T., Son, J. Y., & Stigler, J. W. (Under review). Integrating R in a college statistics course improves student attitudes toward programming. Submitted to the Annual Meeting of the American Educational Research Association (Orlando, Florida, April 9-12, 2021).

Tversky, A., & Kahneman, D. (1977). Causal thinking in judgment under uncertainty. In R. E. Butts & J. Hintikka (Eds.), *Basic Problems in Methodology and Linguistics* (167-190). Springer, Dordrecht.

Uttal, D. H., Scudder, K. V., & DeLoache, J. S. (1997). Manipulatives as symbols: A new perspective on the use of concrete objects to teach mathematics. *Journal of Applied Developmental Psychology*, *18*(1), 37-54.

Wild, C. (2006). The concept of distribution. *Statistics Education Research Journal*, *5*(2), 10-26.

Whitaker, D., Foti, S., & Jacobbe, T. (2015). The Levels of Conceptual Understanding in Statistics (LOCUS) Project: Results of the pilot study. *Numeracy: Advancing Education in Quantitative Literacy, 8*(2).

Appendix—Transfer Questions

Below is the **lm()** and **supernova()** output for two models of **winpercent**: one using **chocolate** and the other using **sugarpercent** as the explanatory variables.

chocolate model

```
lm(formula = winpercent ~ chocolate, data = candy_rankings)

Coefficients:
  (Intercept)  chocolateTRUE
        42.14          18.78

Analysis of Variance Table
Outcome variable: winpercent
Model: lm(formula = winpercent ~ chocolate, data = candy_rankings)

                            SS df      MS      F    PRE     p
 ----- ---------------- ------- -- ------- ------ ------ -----
 Model (error reduced) |  7368.5  1 7368.54 56.532 0.4052 .0000
 Error (from model)    | 10818.5 83  130.34
 ----- ---------------- ------- -- ------- ------ ------ -----
 Total (empty model)   | 18187.0 84  216.51
```

sugarpercent model

```
lm(formula = winpercent ~ sugarpercent, data = candy_rankings)

Coefficients:
 (Intercept)  sugarpercent
     44.6094        0.1192

Analysis of Variance Table
Outcome variable: winpercent
Model: lm(formula = winpercent ~ sugarpercent, data = candy_rankings)

                           SS df     MS      F    PRE     p
 ----- ---------------- ----- -- ------ ------ ------ -----
 Model (error reduced) |   955  1 955.00 4.5999 0.0525 .0349
 Error (from model)    | 17232 83 207.61
 ----- ---------------- ----- -- ------ ------ ------ -----
 Total (empty model)   | 18187 84 216.51
```

As you can see, the chocolate model explains more of the variation in winpercent than does the sugarpercent model.

Question 1

Let's say we want to create a more complex model (we will call it the **two--predictor** model) that includes two explanatory variables in a single model: **chocolate** and **sugarpercent**. We could represent this **two--predictor** model in a word equation like this:

winpercent = chocolate + sugarpercent + other stuff

How do you think you would represent the **two--predictor** model in GLM notation? (Enter the notation using regular letters, for example, b1 for b_1; or Xi for X_i.)
<TEXT ENTRY BOX HERE>
Question 2

What do you expect the PRE to be for the new model?

- Between the PRE for the **chocolate** model and that for the **sugarpercent** model
- Higher than either the **chocolate** model or the **sugarpercent** model
- Lower than the PREs for either the **chocolate** model or the **sugarpercent** model
- Can't tell from the information given

Question 3

What do you expect the SS Total to be for the new model?

- Lower than that of the **chocolate** model and the **sugarpercent** model
- Higher than that of the **chocolate** model and the **sugarpercent** model
- The same as both the **chocolate** model and the **sugarpercent** model
- Can't tell from the information given

Question 4

What do you expect the SS Error to be for the new model?

- The same as that of the **chocolate** model
- Higher than that of either the **chocolate** model or the **sugarpercent** model
- The same as that of the **sugarpercent** model
- Lower than that of either the **chocolate** model or the **sugarpercent** model

<PAGE BREAK>
Question 5

Here is the answer to the previous question: The **two--predictor** model should have a lower SS error than either the **chocolate** model or the **sugarpercent** model. Why do you think this is true?

<TEXT ENTRY BOX HERE>

Question 6

We fit the **two--predictor** model using **lm(winpercent ~ chocolate + sugarpercent, data=candy_rankings)**. The resulting parameter estimates and anova table are presented below.

winpercent ~ chocolate + sugarpercent

```
lm(formula = winpercent ~ chocolate + sugarpercent, data =
candy_rankings)

Coefficients:
  (Intercept)  chocolateTRUE   sugarpercent
     38.26211       18.27331        0.08567

Analysis of Variance Table
Outcome variable: winpercent
Model: lm(formula = winpercent ~ chocolate + sugarpercent, data =
candy_rankings)

                             SS df      MS       F    PRE     p
 ----- ----------------- -------- -- ------- ------- ------ -----
 Model (error reduced) |  7856.13  2 3928.06 31.1784 0.4320 .0000
       chocolate       |  7368.54  1 7368.54 58.4867 0.4052 .0000
       sugarpercent    |   487.59  1  487.59  3.8701 0.0268 .0525
 Error (from model)    | 10330.91 82  125.99
 ----- ----------------- -------- -- ------- ------- ------ -----
 Total (empty model)   | 18187.03 84  216.51
```

Based on this output, what would you predict the winpercent to be for a candy that contains chocolate?

- 38.26 + 18.27
- 38.26 + 18.27 + 8.57
- 38.26 + 18.27 + 8.57**sugarpercent*
- Can't tell from the output above

Question 7

Finally, we ran the **two--factor** model in two ways (see output below). On the top is the result of running the model with **chocolate** first, followed by **sugarpercent** (just like we did before). On the bottom, we reversed the order, putting **sugarpercent** first, followed by **chocolate**.

winpercent ~ chocolate + sugarpercent

```
lm(formula = winpercent ~ chocolate + sugarpercent, data =
candy_rankings)

Coefficients:
  (Intercept)  chocolateTRUE   sugarpercent
     38.26211       18.27331        0.08567

Analysis of Variance Table
Outcome variable: winpercent
Model: lm(formula = winpercent ~ chocolate + sugarpercent, data =
candy_rankings)

                             SS df      MS       F    PRE     p
----- ----------------- -------- -- ------- ------- ------ -----
Model (error reduced) |  7856.13  2 3928.06 31.1784 0.4320 .0000
      chocolate       |  7368.54  1 7368.54 58.4867 0.4052 .0000
      sugarpercent    |   487.59  1  487.59  3.8701 0.0268 .0525
Error (from model)    | 10330.91 82  125.99
----- ----------------- -------- -- ------- ------- ------ -----
Total (empty model)   | 18187.03 84  216.51
```

winpercent ~ sugarpercent + chocolate

```
lm(formula = winpercent ~ sugarpercent + chocolate, data =
candy_rankings)

Coefficients:
  (Intercept)   sugarpercent  chocolateTRUE
     38.26211        0.08567       18.27331

Analysis of Variance Table
Outcome variable: winpercent
Model: lm(formula = winpercent ~ sugarpercent + chocolate, data =
candy_rankings)

                            SS df      MS       F    PRE     p
----- ----------------- ------- -- ------- ------- ------ -----
Model (error reduced) |  7856.1  2 3928.06 31.1784 0.4320 .0000
      sugarpercent    |   955.0  1  955.00  7.5802 0.0525 .0073
      chocolate       |  6901.1  1 6901.13 54.7766 0.3795 .0000
Error (from model)    | 10330.9 82  125.99
----- ----------------- ------- -- ------- ------- ------ -----
Total (empty model)   | 18187.0 84  216.51
```

As you can see, a lot of things are similar across these two models: SS model, SS error, SS total, and all three of the parameter estimates. On the other hand, SS for **chocolate** and for

sugarpercent are quite different across these models. Why do you think this might be? What's happening here?
<TEXT ENTRY BOX HERE>