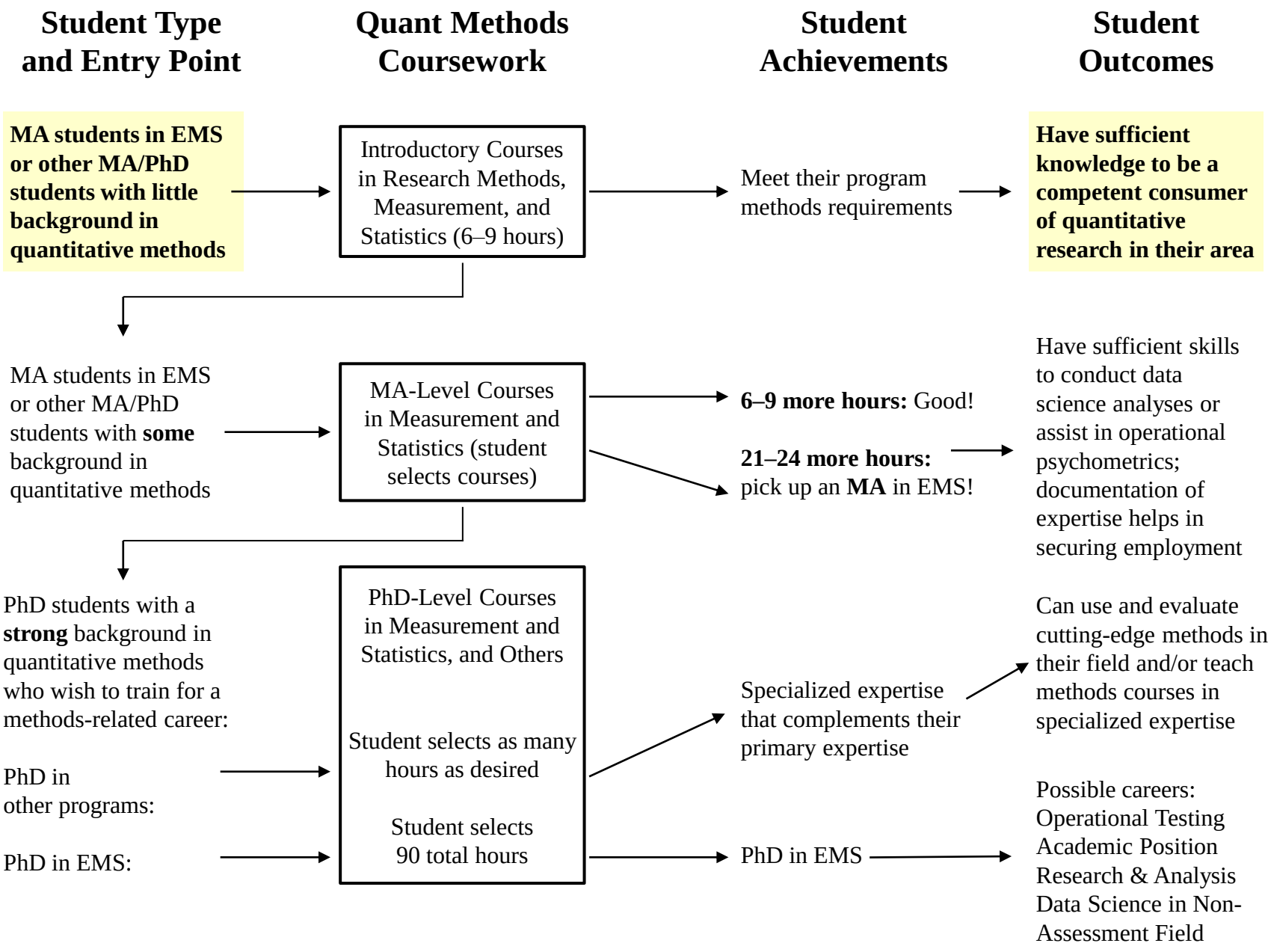
Introduction to PSQF 6243 (and to Quantitative Methods)

- Topics:
  - Why might you be here?
  - The truth about "statistics"
  - Course requirements, responsibilities, and your experience
  - About the statistical software used in this course
  - What you are supposed to know already (or should review)
  - What we will cover this semester (and what could be next)

Two Reasons Why You Are Here

1. "This class fulfills a requirement" (and I just need to pass it).
   - I get it—it's ok if this is the *only* reason you are here, but I hope to convince you otherwise!

2. "I want to learn more about data analysis using **quantitative methods**" (yes, me too)!
   - One method by which to answer questions—in real life or in research settings—is by collecting quantitative data
   - The process of summarizing that data—by finding patterns in order to answer questions—requires statistical models
   - **Quantitative methods = Quantitative data + application of statistical models to answer questions**
   - Let's examine the levels of expertise you can acquire…


Student Type and Entry Point
Quant Methods Coursework
Student Achievements
Student Outcomes
MA students in EMS or other MA/PhD students with little background in quantitative methods
Introductory Courses in Research Methods, Measurement, and Statistics (6–9 hours)
Meet their program methods requirements
Have sufficient knowledge to be a competent consumer of quantitative research in their area
MA students in EMS or other MA/PhD students with some background in quantitative methods
MA-Level Courses in Measurement and Statistics (student selects courses)
6–9 more hours: Good!
21–24 more hours: pick up an MA in EMS!
Have sufficient skills to conduct data science analyses or assist in operational psychometrics; documentation of expertise helps in securing employment
PhD students with a strong background in quantitative methods who wish to train for a methods-related career:
PhD in other programs:
PhD in EMS:
PhD-Level Courses in Measurement and Statistics, and Others
Student selects as many hours as desired
Student selects 90 total hours
Specialized expertise that complements their primary expertise
Can use and evaluate cutting-edge methods in their field and/or teach methods courses in specialized expertise
PhD in EMS
Possible careers:
Operational Testing
Academic Position
Research & Analysis
Data Science in Non-Assessment Field

Intermediate Statistical Methods: What Will "Statistics" Mean to Us?

- Statistics = **applied math** used for a **relevant** purpose!

- Competent consumers and users of quantitative methods must learn the **logic behind the uses of statistical models**

- **This will NOT require anxiety-provoking behaviors like:**
  - Calculating things by hand—computers are always better, and more advanced statistical models cannot be implemented by hand anyway
  - Deriving formulas or results—it's ok to trust the people who specialize in these areas to have gotten it right and use their work (for now, at least)
  - Memorizing formulas—it's ok to trust the computer programmers who have implemented various estimation techniques (for now, at least)

The Truth about "Statistics"

- The hardest part about learning statistics is **not the math**… it is the **working memory load** of new language + logic!

- **Language:** Ideas will be expressed through words, notation (symbols and equations), and computer code ("syntax")

- **Logic:** Decision guidelines for matching data types and questions to statistical models (and then "estimating" models)

- **Working memory load** is reduced through frequent exposure, mindful repetition, and engagement → automaticity
  - This is one the main reasons we will meet twice a week
  - The material builds cumulatively, so staying checked in will help!

How I Will Help You Acquire the Language and Logic of Statistical Modeling

- I believe that **everyone is capable** and **can significantly benefit**** from learning how to use quantitative methods!
- **Philosophy:** Focus on accessibility + mastery learning
- **Materials:** Unit = (wordy) lecture + example(s); 6–7 planned
  - **Lecture** slides present concepts—the **what** and the **why**
  - **Example** documents: reinforce the concepts and demonstrate the **how using software**—SAS, STATA, or R (stay tuned)
    - Unit 1 will have software demo videos instead of an example document
  - All available at the course website (hosted outside of ICON)

** ***Benefits*** *include but are not limited to: Better research, more authorship opportunities, and actual money*

How I Will Help You Acquire the Language and Logic of Statistical Modeling

- I will NOT:
  - Present statistics as a series of unrelated ideas and formulae
  - Use infrequent high-stakes tests to assess your learning
  - Ask you to conduct extensive calculations by hand (or in excel)

- I WILL:
  - Present statistics by linking data, questions, and models explicitly
  - Use **formative assessments (in ICON)** to help you review concepts (6 planned; 12 points for **completing them at all**)
  - Use **homework (in my custom online system)** to give you hands-on software practice (6 planned; 88 points for **accurately completing** both computation and interpretation questions)

More About the Course Requirements

- **Everything** is take-home, open-note, and untimed
- Late* work will be accepted (with small penalties)
  - **Extensions granted if requested at least 2 weeks in advance*
  - HW due dates **may be pushed later** (to ensure approximately 1 week after covering the material before it's due), but never sooner

- **Formative assessments:** Big-picture questions for a structured review (will go over answers at the next class)

- **Homework assignments:** Practice doing data analysis
  - Based directly on examples given (no googling required)
  - You will each have a unique dataset (made with a common story)
  - **Computation** sections: Instant feedback, infinite attempts
  - **Results** (interpretation) sections: Delayed feedback, single attempt (but repetition of concepts across the semester)

Our Other Responsibilities

- My job (besides providing materials and assignments):
  - **Answer questions** via email, in individual meetings, or in group-based zoom office hours—you can each work on homework during office hours and get (near) immediate assistance (and then keep working)

- Your job (in descending order of timely importance):
  - **Ask questions**—preferably in class, but any time is better than none
  - **Frequently review** the class material, focusing on mastering the vocabulary (words and symbols), logic, and procedural skills
  - Don't wait until the last minute to start homework, and don't be afraid to **ask for help if you get stuck** on one thing for more than 15 minutes
    - Please email me a screenshot of your code + errors so I can respond easily
  - **Read the textbook** for a broader perspective and additional examples (best after lecture; readings are for the whole unit, not just that day)
  - **Practice** using the software to implement the techniques you are learning **on data you care about**—this will help you so much more!

More About Your Experience in this Class

- **Attendance**: Strongly recommended but not required
  - **You choose** (for any reason): In-person or zoom
  - **Masks** are *STRONGLY encouraged* for in-person attendees
  - **Please do not attend in-person if you were exposed to Covid!**
  - You won't miss out: I will post **YouTube-hosted recordings** (audio + screenshare only) for each class at the course website
  - **Ask questions aloud or in the zoom chat window (+DM)** (even if you are attending class in-person)

- **Changes** will be sent via email by 9 am on class days
  - I will change to zoom-only if I am exposed to Covid!
  - I will change to zoom-only for dangerous weather
  - Nothing is more important than our health and safety…

Class-Sponsored Statistical Software

- To help address the needs of different Iowa degree programs, I will show examples using **SAS, STATA, and R** software
  - **SAS** = "Statistical Analysis System"
  - **STATA** (aka, Stata) = "Software for Statistics and Data Science"
  - **R** = free implementation of what was initially the "S" language

- **Why not SPSS**? Because it doesn't have as much room to grow (and thus isn't used in any EMS advanced classes)
  - As in SPSS, drop-down windows can also generate syntax in STATA and in SAS "enterprise" (which I don't use, and you won't need to)
  - Btw, SPSS is used in the textbook, and it can do *most* of our content

- **My story:** After SPSS, I became a heavy-duty **SAS enthusiast** who:
  - Picked up enough STATA initially to teach workshops using it, and I am learning it better now that I teach it in my classes
  - Is (begrudgingly) learning enough (base) R to add it to my classes
  - So if you have **STATA or R tips**, please share them with me!

Which Program: SAS, STATA, or R?

- **Yes, you will need to learn to use at least one of these!**
  - Each is available (with VPN) in the free U Iowa Virtual Desktop
  - More programs = more "technical skills" for your CV; easier collaboration with colleagues (who only know one program)

- **To consider** when choosing which program to focus on:
  - Future use: R can be freely installed on your own machine; SAS has a free web-based SAS OnDemand; STATA install = $$$
  - **STATA** is popular in fields that use **large, weighted survey data** (e.g., sociology, political science, public health, EPLS at Iowa)
  - **R** will be used exclusively in classes by Drs. Aloe, LeBeau, or Templin, and it has become increasingly mainstream, **but**:
    - R packages are only as good as their authors (so little quality control)
    - Syntax and capabilities are idiosyncratic to the packages (grrrrrr)

SAS vs. STATA: My Opinion

Activity	Winner	Commentary
Working with many raw data files or multiple datasets	SAS, hands down	Without their newer "frames", only one dataset can be open at once—problematic for data management
Within-dataset manipulations	Tie, but STATA for some tasks	STATA wins for group-centering, stacking, and unstacking data (i.e., as used for multilevel models)
Data analysis	Tie, but SAS for some tasks	I've had estimation problems in STATA for certain advanced model variants (within multilevel models)
Post-estimation (i.e., predicted outcomes or simple slopes)	STATA, hands down	STATA has simple yet powerful options for doing these tasks in bulk that SAS doesn't have
Automating data tasks (i.e., loops)	Tie	Both programs have ways to do this, but I only know how in SAS…

Overview of Syntax by Program

- SAS, STATA, and R differ greatly in their syntax structure (syntax = codes you type to make the program go)
- Syntax in both **STATA** and **R** is **case-sensitive**! SAS is not.
  - When possible (in SAS), I use UPPER-CASE letters for recognized program commands, and lower-case (or Title Case) for user-specific info to be changed (like names of datasets and variables)

- **SAS syntax** has **two formats** used for almost everything
  - "Data step" (for managing data) and "PROC" (for analysis)
  - Semi-colons are line terminators (how you say the command is done)
- **STATA syntax** is also regularized, but it uses **fewer words**
  - Quicker to type, but also much less transparent
  - Line end is command termination (so must use /// as a line continuer)
- **R syntax** is composed (almost?) entirely of **calls to functions**
  - May look familiar to coders, but foreign to the rest of us
  - R is much, much easier when used within the Rstudio interface

A Colloquial Demo of Program Syntax

- Imagine you were asked how your dinner was… and you'd like to answer "It's fine, not too spicy" in each program
- Text in green are comments (= notes only to yourself)

```
* Answer question about dinner dataset using SAS;
PROC ANSWER DATA=work.dinner;
     MODEL response = fine / SPICY=NO; * Options after /;
RUN; * RUN makes it go and print (like EXECUTE in SPSS);
```

```
// Answer question about dinner (only open) dataset using STATA
answer response fine, nospicy  // Options after comma;
// Result is printed after execution without analog to RUN
```

```
# Answer question about dinner dataset using answeR package in R
myanswer = answeR(data=dinner, formula~response=fine, spicy=FALSE)
summary(myanswer) // Print of saved result requested separately
```

How am I ever going to learn this???

- I will demonstrate how to access the Virtual Desktop and how to use each software program in videos (coming soon)
- Don't worry: I DO NOT need you to memorize syntax, ever!

- Instead, you can do exactly what I (still) do:
  - **Find the example I gave you** of what you need to do
  - Figure out how to **modify it** to work for your homework
  - **Copy** (control+C), **paste** (control+V), and **find and replace** (control+H) are your friends (Mac: swap control for command)

- Colors will help you troubleshoot (e.g., in SAS, red=wrong)
- Don't hesitate to ask for help (i.e., email me a screenshot)
- It will get easier with practice, I promise!!!

What You Are Supposed To Know Already

- Listed pre-requisite: PSQF 4143 or equivalent

- Working pre-requisites are familiarity with:
  - Descriptive statistics (e.g., frequency, mean, variance)
  - Bivariate associations (e.g., Pearson correlation)
  - Statistical concepts (e.g., null hypothesis testing)
  - Use of some (non-excel) software for all of the above

- We will quickly review these concepts in units 1–2
  - For a more thorough treatment, review the first five weeks of materials + videos of PSQF 6242

- Most of this class will focus on the **GLM**… so what's that?

What We Will Cover This Semester

Intro to **General Linear Models** (GLMs) as a one-stop shop *for predicting one conditionally normal outcome per person*

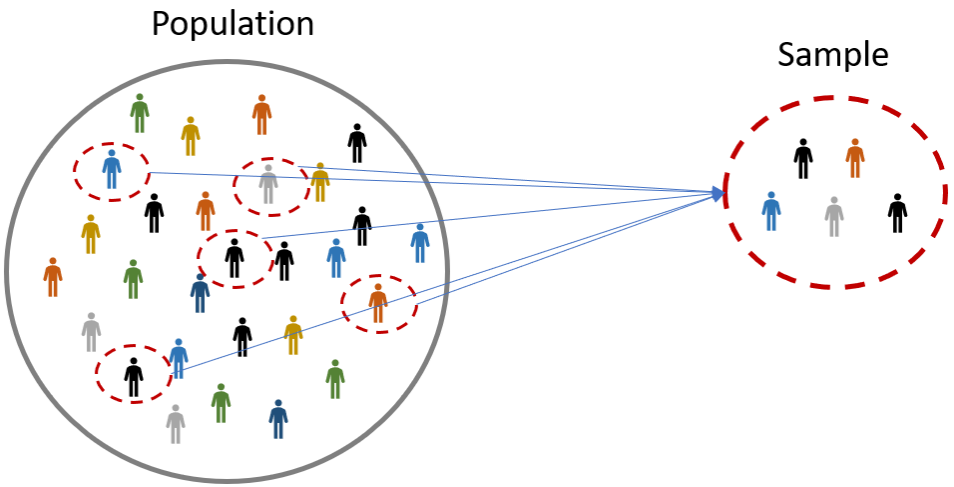
- Quantitative predictors = *"(linear) regression"*
  - 1 numeric predictor variable = *"simple (linear) regression"*
  - 2+ numeric predictor variables = *"multiple (linear) regression"*
  - We will cover both linear and nonlinear prediction
- Categorical predictors = *"analysis of variance (ANOVA)"*
  - 1 two-group predictor variable = *"independent-samples t-test"*
  - 1 three-or-more-group predictor variable = *"one-way ANOVA"*
  - 2+ group predictor variables = *"two-way (or factorial) ANOVA"*
- Both kinds of predictors = *"analysis of covariance (ANCOVA)"*
- We will cover moderation (via interactions) of all kinds, too!

I present this material in a way that builds into future courses…

So What Kind of Data Can Use GLMs?
Let's Review Some Sampling Vocabulary…

- Who are we trying to know about, more generally? → To what **population** do we want to make inferences?

- Accordingly, from whom should we collect data? → And what info should we collect in our selected **sample**?
  - **Variables** are characteristics that **differ across units*** in a sample



* Units = persons, organizations, animals, etc.

Image borrowed from: https://www.sigmamagic.com/blogs/online-sample-size-calculators/

Where to Begin? Sampling Vocabulary

- Example: Let's say a researcher wants to examine graduate student life, so they use a survey to collect self-report info on program membership, stress levels, and well-being.

- So what **type of sample** should we collect? For instance:
  - Collect data for multiple students from the same program only? Then **program would be a *constant*, not a *variable***
  - To examine **differences between programs**, we'd need to sample multiple programs from the same college, at a minimum
  - But would it help our **generalizability** to include multiple colleges from the same university, or even from multiple universities?
  - Should we survey each student once? Or would **several times** be better?
  - Should we also try to collect **corresponding data** from other people who know each student well (e.g., their partners, friends, family)?
- These questions address **independent** versus **dependent** sampling…
  - The latter cases are also known as "**dependent data**"

Independent vs. Dependent Samples

- Example of a (maybe) **independent sample**: One occasion of measurement each from students in the same program
  - If program is a **constant**, not a **variable**, it can't be part of any research questions (but then program differences are controlled)

- Examples of **dependent** (= naturally related) **samples** (in which your analyses must account for common sampling):
  - Sample lots of programs (e.g., >20) from same university
    - e.g., Stress rates of persons from the same program may be more related (dependent) than those of persons from different programs
    - This is known as "**clustered**" or "**nested**" data
  - Sample each person more than once
    - e.g., Stress rates at occasions from the same person may be more related (dependent) than those of occasions from different persons
    - This is known as "**repeated measures**" or "**longitudinal**" data
    - Collect both self-report and another-report ratings → "**dyadic**" data

PSQF Courses that Cover Analysis of Independent and Dependent Samples

- **PSQF 6243** will only be able to cover analysis of quantitative data from **independent** samples via general linear models
  - Using "**univariate**" statistical models (of one observation per variable per person) predicting a (conditionally normal) **numeric variable**

- My next courses are extensions for **dependent** samples:
  - PSQF 6270 Generalized Linear Models: models for predicting **other kinds of variables**, as well as "**multivariate**" statistical models for predicting **multiple outcomes at once** (and testing mediation)
  - PSQF 6271 Longitudinal Multilevel Models: multivariate (mixed-effects) models for **repeated measures** data (of occasions nested in persons)
  - PSQF 6272 Clustered Multilevel Models: multivariate (mixed-effects) models for **clustered/nested data** (of persons nested in many groups)
  - But **GLMs are the key building block** of all of these advanced models!

Wrapping Up

- **End goal of this semester**: Learn how to use general linear models [**GLMs**; with variants known as regression, analysis of (co)variance] to analyze quantitative research data
  - Requires learning **new language** (words, symbols, and syntax) and **logic** by which to link data, questions, and models
  - Begins by reviewing how to summarize variables (lecture 1) so that you can get to the know the software using familiar ideas
  - Continues with GLMs: statistical models for predicting numeric variables from any kind of variable in **independent samples** *(which need extensions to be covered elsewhere for predicting other kinds of variables or for use in dependent samples)*

- We will estimate GLMs using **SAS, STATA, or R** software
  - I will provide examples of what you will need to do to complete the homework assignments (and for your future reference)