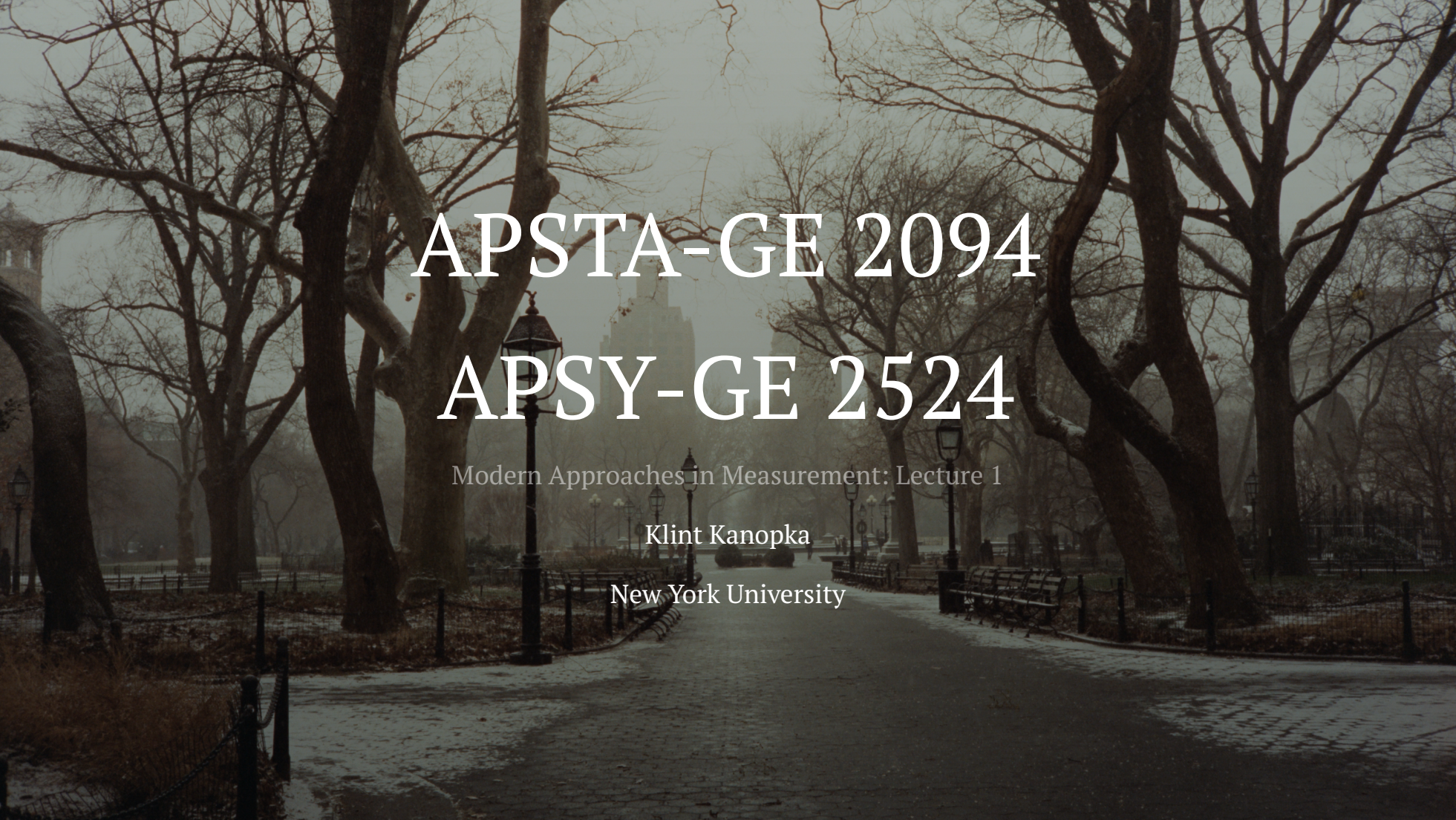


APSTA-GE 2094
APSY-GE 2524

Modern Approaches in Measurement: Lecture 1

Klint Kanopka

New York University

Table of Contents

Modern Approaches in (Psychological) Measurement

Your Instructor

- Klint Kanopka
- Third year assistant professor of applied statistics
- I'm *mostly* a psychometrician, but I have a background in computational physics and machine learning
- Please call me Klint if you feel comfortable doing so
- This is the third time this course has ever existed, so I'm *very* interested in feedback
- Email me: klint.kanopka@nyu.edu
- Office Hours: W 2p-3p @ Kimball, Rm 205W

Fun Facts

- I have a lot of idiosyncratic coding preferences and habits
- I'm a dedicated `ggplot()` user
- I love the `tidyverse` for cleaning data
- I hate the `tidyverse` for absolutely everything else
- I think everything we do should be presented four ways:
  1. Intuitively
  2. Mathematically
  3. Computationally
  4. Applied
- My problem sets have a pretty clear "style"
- I am going to be really picky about some things and then not picky at all about basically everything else (I'll try my best to warn you)
- Questions?

Course Business

Prerequisites

- You should know right now:
  - `R`, the programming language
  - Data Visualization
  - Regression
  - Probability
- One (or both) of these are helpful, but definitely not required:
  - Survey methods
  - Unsupervised machine learning

Course Grades

- Three-credit course
- Eight equally-weighted problem sets (PS0-PS7)
- Participation is basically a subjective assessment of "are you doing class"
- Extra credit points are added directly to the total for problem sets
- I probably set up the Brightspace gradebook wrong
- Category weights:

Category	p
Problem Sets	0.9
Participation	0.1

Grading Scale

	G^-	G	G^+
A	$[.895, .945)$	$[.945, 1]$	
B	$[.795, .825)$	$[.825, .865)$	$[.865, .895)$
C	$[.695, .725)$	$[.725, .765)$	$[.765, .795)$
D	$[.600, .640)$	$[.640, .670)$	$[.670, .695)$
F		$[0, .600)$	

Problem Sets

- Released on (or before) class on Thursdays
- Due on Fridays @ 11.59p
- PS0 is a one week assignment
- PS1-PS6 are two week assignments
- PS7 is longer, due during finals week, and more like a "project"
- Use my `.qmd` template and submit the compiled `.pdf` file on Gradescope
  - You should be prepared to submit the `.qmd` upon request
- Do not call `install.packages()`
- For more specifics, see syllabus

Participation

- If I remember who you are, see you in class, and you're reasonably engaged in the stuff we do, you'll get full credit
- Understand that this is partially subjective
- Attendance will be taken in class using PollEverywhere and you must enable your device's location settings to receive credit for attending

Extra Credit

- One point to the first person to identify a typo or mistake in **any** of the course materials I write
  - Problem sets
  - Book
  - Slides
- If there are code mistakes or more serious issues and you provide fixes, you'll get more than one point, consistent with effort required
- "Valuable suggestions" include proposed examples, datasets, ways to increase clarity, new topics, and applications
  - If I *plan* to adopt a suggestion, you'll get extra credit
  - If I *actually* adopt a suggestion, you'll get acknowledgement in course materials

Collaboration Policy

- You should form study groups and work together on problem sets!
- You must:
  - Report at the top of your assignment who you consulted with
  - Write your code, answers, and solutions independently
  - Understand your work well enough to reconstruct it entirely on your own
- What if you don’t have friends in the class?
  - PRIISM has a great lounge, and it’s a great place to co-work!
  - It’s the area right outside my office, so maybe hang out during office hours?
  - Working with others is mutually beneficial

AI Tool Policy

- See the syllabus for the full policy
- Don’t use generative AI to automatically write your code, analyze your data, or write text for you
- What do you do if you’re stuck?
  - Stack Overflow
  - Stack Exchange
  - Ask your classmates!
  - Come to office hours!

Syllabus Wrap Up

- Hopefully this all makes sense?
- If not ask me questions!
- I am super excited for this class
  - I hope you are too!
  - I will make mistakes
  - Tell me if you need things

Check-In

- PollEv.com/klintkanopka

Measurement

Measurement

- The quantification of attributes of an object or event, which can be used to compare with other objects or events
- The fundamental building block of all quantitative science
- A measurement has four parts:
  - **Level** (or type): ratio, difference, ordinal, nominal[1]
  - **Magnitude**: the number assigned by an *instrument*
  - **Unit**: the scale factor that allows comparison between different measurements
  - **Uncertainty**: a quantification of the error inherent to making the measurement, often a function of the properties of the instrument itself

---

1. Nominal quantities are often *not* considered measurements, but we are going to talk about them in the second half of the course ↩

Units and Scales

- Meters
- Kilograms
- Proof
- Scoville heat units
- Mohs scale
- Schmidt pain index

Psychometrics

- Psychometrics is the study and practice of measuring psychological constructs
- *"All that exists, exists in some amount and can be measured" - E.L. Thorndike, 1918*
- Big problem:
  - Psychological constructs aren't directly observable
  - We call these *latent* constructs
  - How do we measure what we can't observe?
- You put people in carefully crafted situations and see what they do
  - Tests
  - Surveys
  - Behavioral screeners
  - Observation protocols
  - We will blanket refer to these types data as *item response data*

Paradigm Shift

- Often survey research cares about *population* level inferences
- Here we care about and develop tools for inference and comparison at the *individual* level

The "Modern" Part: Unsupervised Machine Learning

- Machine learning is often used in a *supervised* context
  - We have known outcomes
  - We can measure performance on out of sample data
  - We can predict outcomes for new observations
- Measurement contexts aren't like that!
- We can, however, think of them as *unsupervised* machine learning problems
  - There is no known outcome
  - The model is supposed to learn patterns from the data
  - The constraints we put on the model dictate the types of patterns it will be sensitive to
  - Things like person abilities, item difficulties, and weights of individual items are all unknown and need to be estimated simultaneously
  - To do this, we'll use the EM algorithm (which will be a party in a future lecture)

Scoring

Scoring

- Scoring is the process of summarizing individual item responses
- We want to think deeply about how scoring is done, as it is the fundamental process of measurement!
- For each of the assessments on the next slide, think:

1. How could we score each of these?
2. What are potential pros and cons of different scoring practices?

Scoring

1. Depression Screener

1. I feel down, depressed, or hopeless
2. I have trouble getting out of bed in the morning
3. Every single day of my life is worse than the day before it, meaning every day you see me is on the worst day of my life

2. Math Test

1. $4 + 3 = ?$
2. $8 \div 2(2 + 2) = ?$
3. $\int_0^\pi \sin x \, dx = ?$

3. Health Screener

1. Do you have diabetes?

Person and Item Properties

Some Item Response Data

- The dataset `frac20_noq.csv` (downloadable here) contains a selection of responses to math items about fractions. These data are stored in *long form*, with three columns:
  - `id` contains a `numeric` code that uniquely identifies each respondent.
  - `item` contains a `string` that identifies each individual item
  - `resp` contains a `numeric` code for each respondent-item pair, containing `1` if that respondent answered correctly and `0` if they answered incorrectly

Converting Between Long and Wide Form

```
1 library(tidyverse)
2
3 d <- read_csv('frac20_noq.csv')
```

Converting Between Long and Wide Form

```
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider()
```

Converting Between Long and Wide Form

```
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider(data =, id_cols =, names_from =, values_from =)
```

Converting Between Long and Wide Form

```
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider(
  data = ,
  id_cols = ,
  names_from = ,
  values_from =
 )
```

Converting Between Long and Wide Form

```
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider(
  data = d,
  id_cols = ,
  names_from = ,
  values_from =
 )
```

Converting Between Long and Wide Form

```
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider(
  data = d,
  id_cols = id,
  names_from = ,
  values_from =
 )
```

Converting Between Long and Wide Form

```
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider(
  data = d,
  id_cols = id,
  names_from = item,
  values_from =
 )
```

Converting Between Long and Wide Form

```
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider(
  data = d,
  id_cols = id,
  names_from = item,
  values_from = resp
 )
```

Converting Between Long and Wide Form

```r
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider(
  data = d,
  id_cols = id,
  names_from = item,
  values_from = resp
)

d_long <- pivot_longer()
```

Converting Between Long and Wide Form

```r
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider(
  data = d,
  id_cols = id,
  names_from = item,
  values_from = resp
)

d_long <- pivot_longer(data =, cols =, names_to =, values_to =)
```

Converting Between Long and Wide Form

```r
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider(
  data = d,
  id_cols = id,
  names_from = item,
  values_from = resp
)

d_long <- pivot_longer(
  data = ,
  cols = ,
  names_to = ,
  values_to =
)
```

Converting Between Long and Wide Form

```
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider(
  data = d,
  id_cols = id,
  names_from = item,
  values_from = resp
)

d_long <- pivot_longer(
  data = d_wide,
  cols = ,
  names_to = ,
  values_to =
)
```

Converting Between Long and Wide Form

```r
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider(
  data = d,
  id_cols = id,
  names_from = item,
  values_from = resp
)

d_long <- pivot_longer(
  data = d_wide,
  cols = -id,
  names_to = ,
  values_to =
)
```

Converting Between Long and Wide Form

```r
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider(
  data = d,
  id_cols = id,
  names_from = item,
  values_from = resp
)

d_long <- pivot_longer(
  data = d_wide,
  cols = -id,
  names_to = 'item',
  values_to =
)
```

Converting Between Long and Wide Form

```r
library(tidyverse)

d <- read_csv('frac20_noq.csv')

d_wide <- pivot_wider(
  data = d,
  id_cols = id,
  names_from = item,
  values_from = resp
)

d_long <- pivot_longer(
  data = d_wide,
  cols = -id,
  names_to = 'item',
  values_to = 'resp'
)
```

Describing People and Items

- Using the `frac20_noq.csv` dataset, take 15 minutes and answer the following questions in a group of 3 ± 1:
  1. How would you quantify individual math ability? Come up with a procedure and find the top and bottom three respondents according to it.
  2. How would you quantify item difficulty? Come up with a procedure and find the three easiest and three hardest items according to it.
  3. How would you quantify how good an item is at distinguishing between high and low ability respondents? What are the best and worst three items for this?
  4. Some items may exhibit between group biases (for example, by age or gender). How might you detect items with this property?

Classical Test Theory

Classical Test Theory

- Classical Test Theory is a theory of *measurement error*
- In CTT, the object of measurement is the *true score*, T
- We can never observe T, we only ever get the observed score, S
- The problem is that the observed score is the sum of the true score and measurement error:
- $S = T + M$
- Measurement error can be due to a ton of different factors:
  - Bad items
  - Environmental factors
  - Test is too short
  - Cheating

Describing People and Items

- **Ability** - (person level) a numerical description of the latent trait of interest
- **Difficulty** - (item level) a numerical description of the item-side driver of response probability
- **Discrimination** - (item level) a numerical description of how well an item differentiates (*discriminates*) between individuals of high and low ability
- **Differential Item Functioning (DIF)** - (item level) occurs when the probability of response depends on group membership

Ability

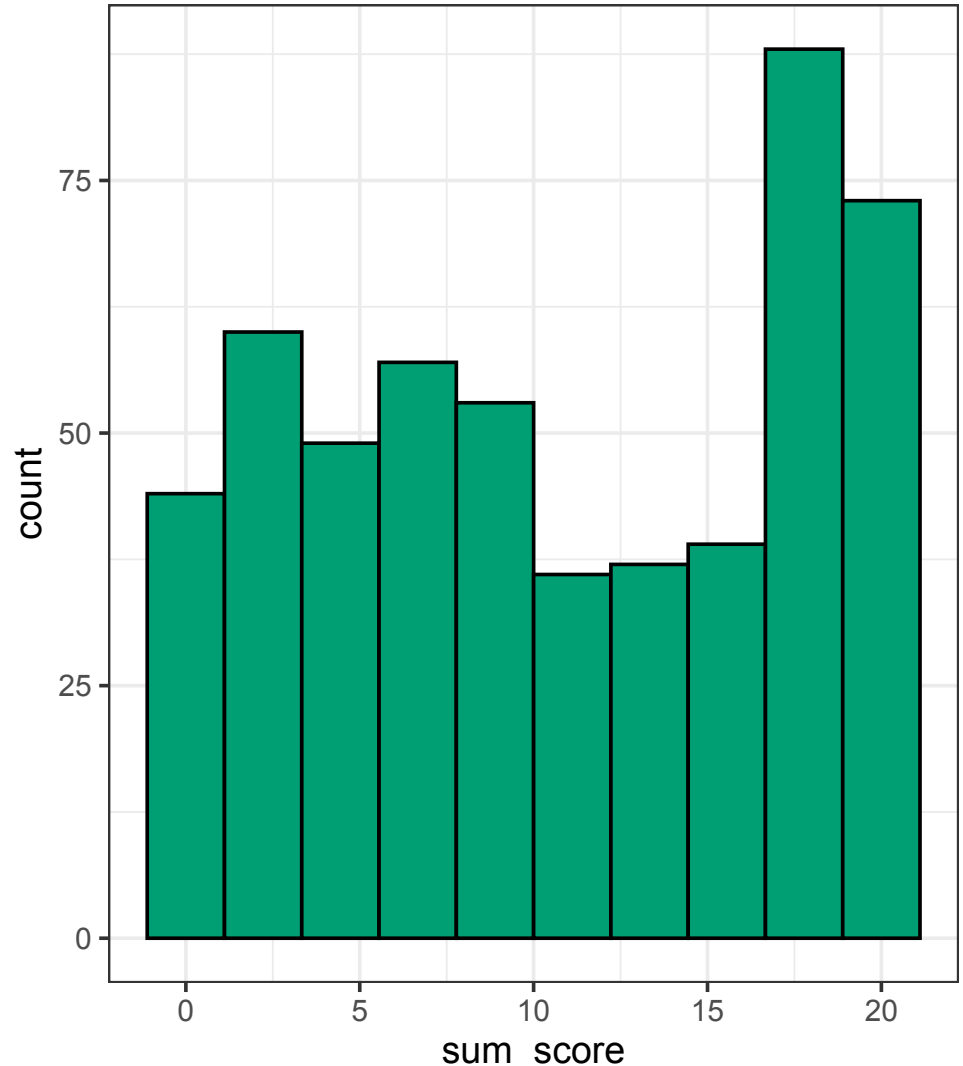
- In CTT, ability is operationalized as the *true score*, T
- Recall the true score is unobservable, as the observed score, S, is confounded with measurement error, M
- Again: $S = T + M$
- Conceptually, this is kind of strange!
  - Observed scores are *samples* from the error-confounded distribution
  - Individual tests are samples from a *universe* of possible tests
  - We will return to this sampling idea with tests!

Ability

Sum Score

```r
sum_scores <- d |>
  group_by(id) |>
  summarize(sum_score = sum(resp))

ggplot(sum_scores, aes(x = sum_score)) +
  geom_histogram(
    bins = 10,
    color='black',
    fill = okabeito_colors(3)
  ) +
  theme_bw()
```



Difficulty

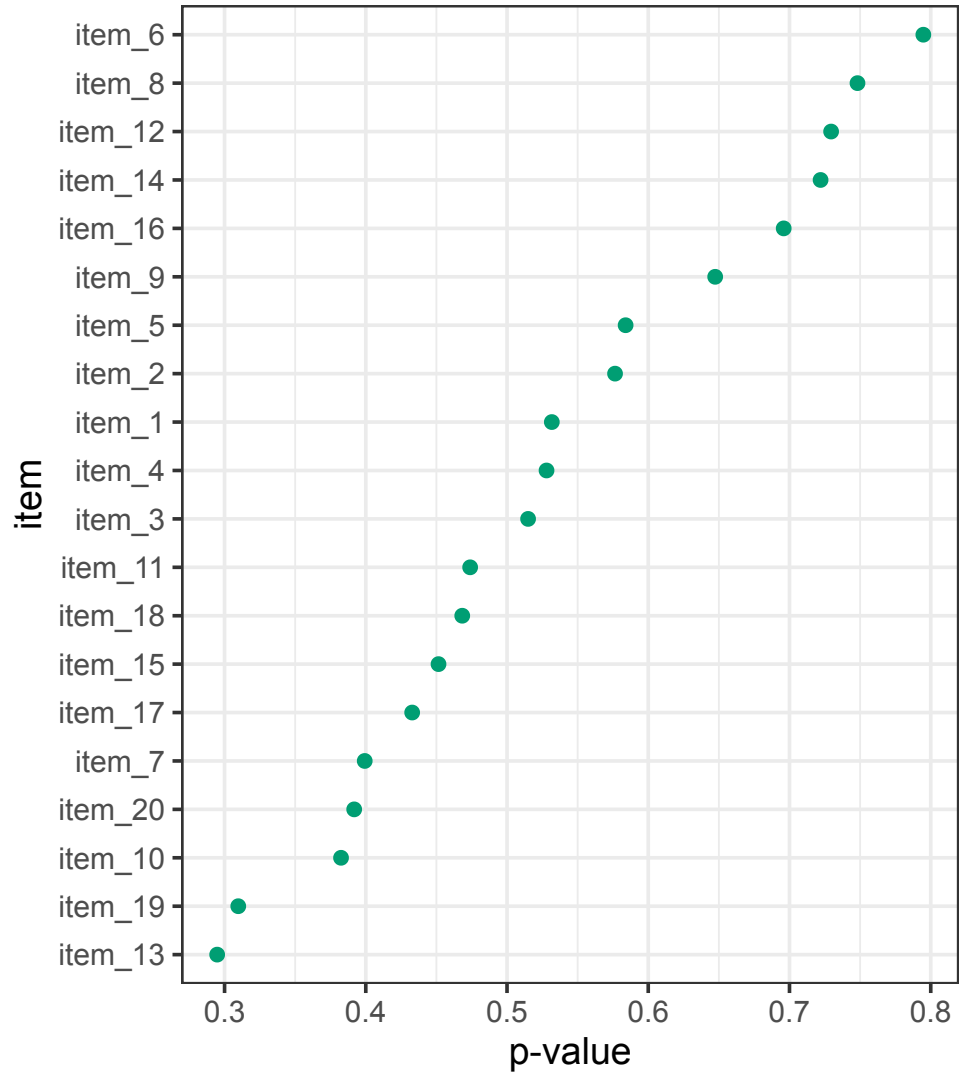
- Difficulty is an observed property of items
- In CTT, we describe this as the p-value
- This is the proportion of respondents giving a particular (usually correct or affirmative) response
- Low p-values mean items are less likely to be correct or affirmed, meaning they are "harder"

Difficulty

p-value

```
diff <- d |>
  group_by(item) |>
  summarize(p = mean(resp))

ggplot(diff, aes(x = p, y=reorder(item,p))) +
  geom_point(color = okabeito_colors(3)) +
  labs(x = 'p-value', y = 'item') +
  theme_bw()
```



Discrimination

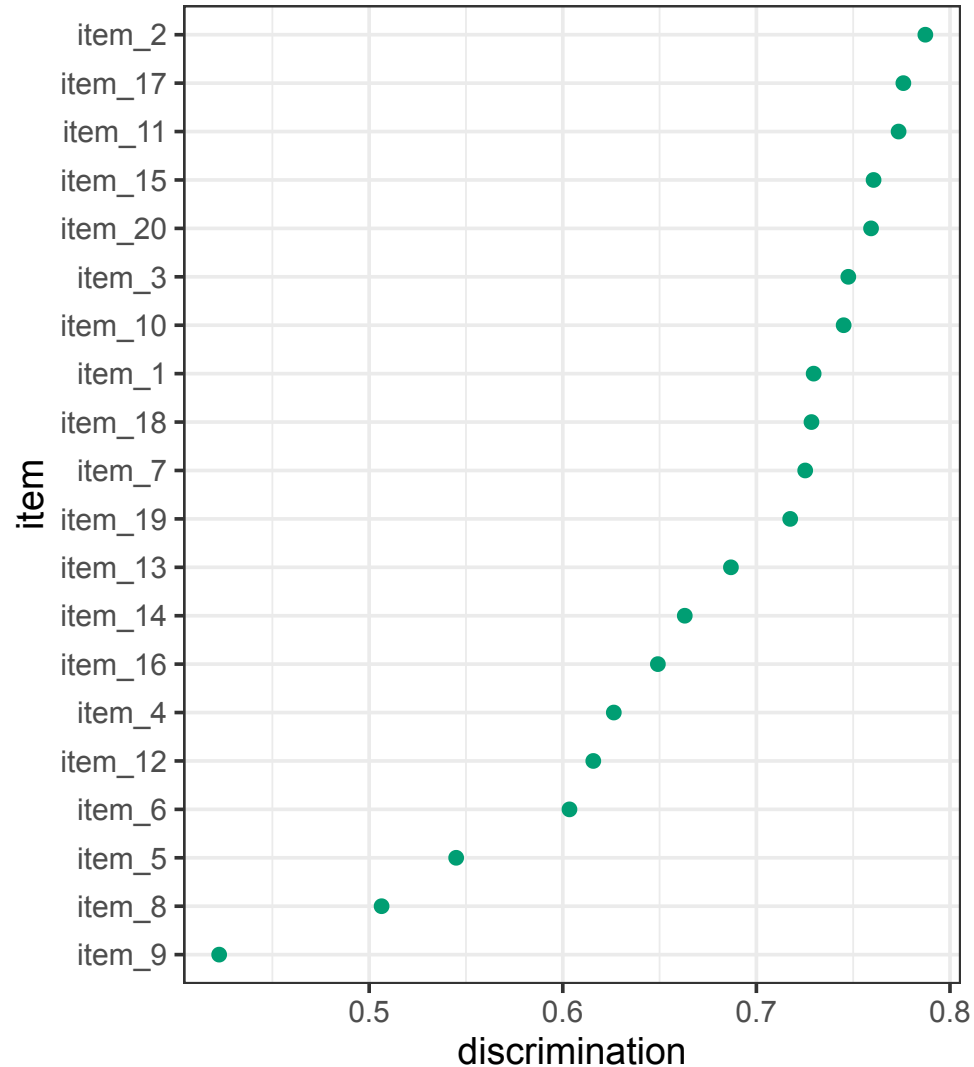
- How well an item differentiates between individuals of high and low ability
- In CTT, this is the item-total correlation
  - Correlation between the response to an item and the sum score
  - For dichotomous items, this is called the *point-biserial* correlation
  - Cheat code: if you mark incorrect/correct responses as 0/1, this is exactly the Pearson correlation
- High discrimination means that you can more clearly divide between high and low ability respondents based upon only the response to that item
  - The item provides higher information about the individual respondent's latent trait
  - Low discrimination items provide less information about the respondent

Discrimination

Item-Total Correlation

```r
disc <- left_join(d, sum_scores, by='id') |>
  group_by(item) |>
  summarize(a = cor(resp, sum_score))

ggplot(disc, aes(x = a, y = reorder(item, a))) +
  geom_point(
    color = okabeito_colors(3)
  ) +
  labs(x = 'discrimination', y = 'item') +
  theme_bw()
```



Differential Item Functioning

- Some items may function differently for respondents in different groups
- Typically we look to see if probability of correct response, conditioned on ability, differs by group membership
  - Often done by regressing the item response on the sum score (minus the item in question) and a group membership variable
  - A significant coefficient on the group membership variable indicates the presence of DIF
- This has two key assumptions:
  1. Some items (called *anchor items*) do not exhibit DIF and allow for comparable ability estimates
  2. Underlying ability distributions are the same for both groups
- DIF can be caused by linguistic bias, cultural bias, or some other item malfunction and is a source of construct-irrelevant variance

Reliability

Reliability

- In groups of 3 ± 1, discuss the following questions:
  - What does it mean for something to be reliable?
  - In what contexts do you think about or hear the term "reliability?"
  - What could it mean for a test to be reliable?
  - What strategies could you use to *quantify* the degree to which a test is reliable?

Reliability

- Reliability is the overall consistency of scores on a test
  - "Scores that are highly reliable are precise, reproducible, and consistent from one testing occasion to another. That is, if the testing process were repeated with a group of test takers, essentially the same results would be obtained."[1]
- Estimated a few different ways:
  - **Test-retest reliability** - If you give the test repeatedly, does it produce the same scores?
  - **Parallel-forms reliability** - If you make two versions of the same test, are the scores across forms correlated?
  - **Split-half reliability** - If you give the test once, are scores on half the items correlated with scores on the other half? (Often we correlate odd and even items)
  - **Internal consistency** - Are responses to each item correlated with responses to each other item? (This is Cronbach's Alpha - it's the mean split-half reliability across all possible splits of a test)

---

1. see the *Standards for Educational and Psychological Testing* ↩

Validity

- On the most basic level, the idea that we have confidence we are measuring the thing we think we are measuring
- Starts with the definition of a *construct*, or the unobservable thing we care about measuring
- Requires some amount of theory and some amount of empirical study to get at
  - The three papers we'll discuss next week are all about this specifically
- Reliability is a necessary *but insufficient* condition for validity!
- We'll discuss this more next week after you read the Cronbach & Meehl, Kane, and Boorsboom, Mellenbergh, & van Heerden papers!

Wrap Up

Recap

- This course is concerned with the measurement of individual-level latent variables
- Scoring is the fundamental process of measurement
- Validity is the idea that we're actually measuring something that makes sense
  - We'll discuss this more next week
- Reliability is a necessary (but insufficient) condition for validity
- Classical Test Theory is mostly concerned with sources of measurement error and provides some intuitive tools for understanding both respondents and items

Problem Set 0

- Currently released, must be submitted on Gradescope
  - Due Friday, January 30th at 11.59p
  - Give yourself some time for the actual submission process if you've never used Gradescope before!
- Pretty typical of what you can expect from my assignments, just shorter
- First question gives you an item response data set, and you will:
  - Recode it, rename variables, and make a table
  - Answer a few questions with some plots
  - Fit a regression and interpret the coefficients
- Second question walks you through a Monte Carlo simulation to explore the "Birthday Problem"
  - Write two functions to carry out the simulation
  - Put them together to do one iteration of the simulation
  - Loop over a bunch of conditions and make a plot

Check-Out

- PollEv.com/klintkanopka