Instructional Groupings

Recommendations for Solar System Middle School

Daniel Anderson

ABL Schools Interview

This talk

Focus on

- Evidence
    - and conceptual discussions of methods used to gather evidence
- Substantive findings
- Practical takeaways
- Other potential areas for future exploration


Quick
background

Instructional groups

- SimGroup: Similar performance
- DivGroup: Diverse performance
- Random: Students groups randomly created
- Manual: Teachers create groupings

Quizes

- 10 quizzes administered, each one week apart
- Instructional grouping changed for each quiz
- Scored on a percent correct basis

RQs

> What groups of students, based on common attributes, are consistently underperforming on their quizzes?

> Are there any types of groups, e.g., manual groups vs calculated (SimGroup, DivGroup, Random), that you would recommend teachers use or avoid using?

Primary takeaway

Amongst instructional groups, scores in the **Manual** were clearly the lowest

Race differences were modest, and generally washed out when accounting for instructional grouping

Gender differences were also modest, and generally washed out when accounting for instructional group

- Some evidence **Male** students scored reliably below **Female** students, on average

DivGroup was clearly the highest performing instructional group, and should be preferred over all others

Evidence

Effect sizes

Difference in means, divided by pooled standard deviation

Benefits

- More comparabale across measures
- More comparable across studies
- Can help interpretation by comparing to historical trends

Effect sizes

Instructional Groups

Focal Group	Reference Group	d
DivGroup	Manual	1.45
DivGroup	Random	1.11
DivGroup	SimGroup	1.15
Manual	Random	−0.36
Manual	SimGroup	−0.42
Random	SimGroup	0.00

Effect sizes

Race

Focal Group	Reference Group	d
Atlantean	Liliputian	0.06
Atlantean	Martian	0.21
Atlantean	Venutian	−0.13
Liliputian	Martian	0.15
Liliputian	Venutian	−0.19
Martian	Venutian	−0.34

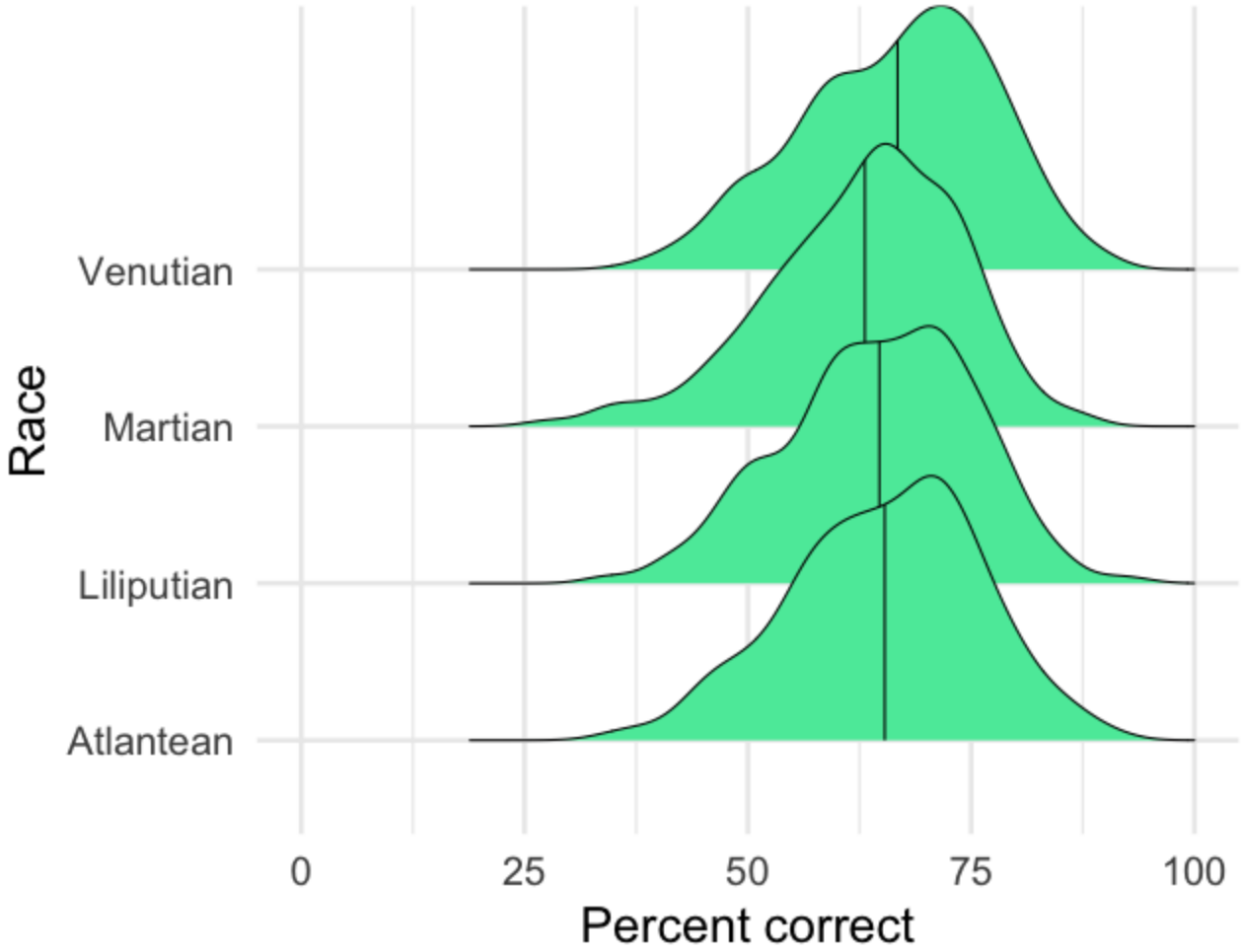
Quick sidebar: "Gap" language

Mean differences

- In education, we often talk about "achievement gaps", implying mean differences
- There is some evidence that this puts greater onus on students, rather than systems
- These **average** differences may also lead teachers to have lower expectations for specific groups of students

But **mean differences** imply little for an **individual** student.

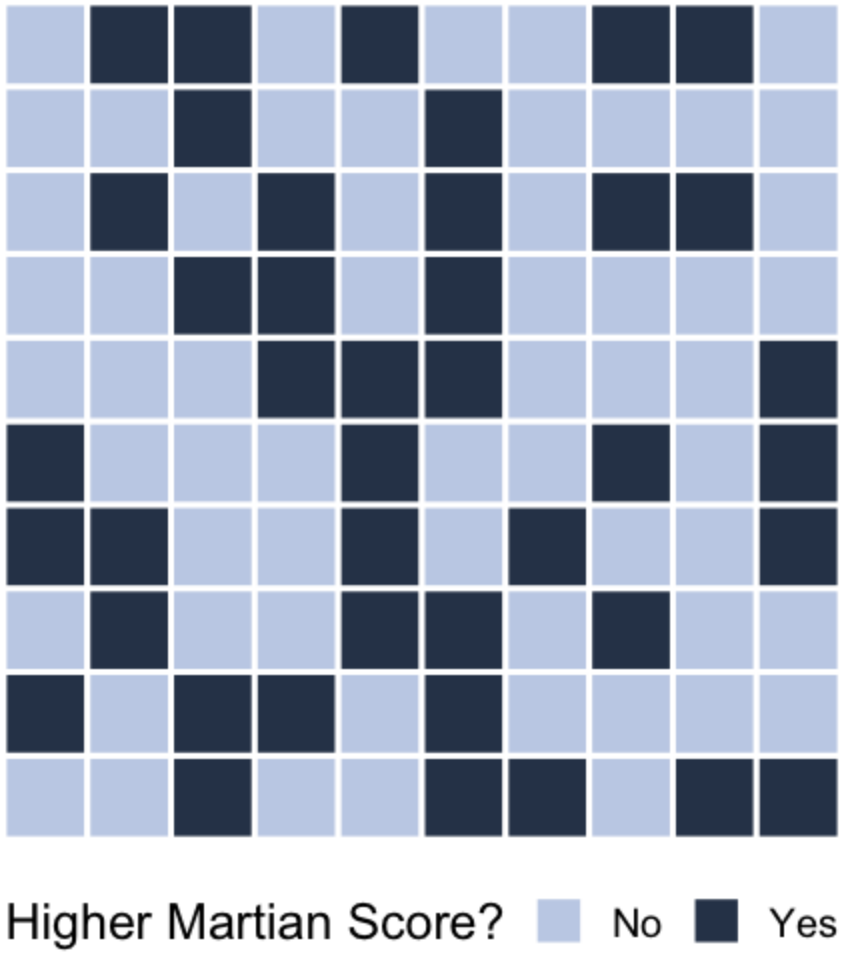
Variability in distributions



An alternative metric

Focal Group	Reference Group	auc
Atlantean	Martian	0.55
Atlantean	Venutian	0.46
Atlantean	Liliputian	0.52
Liliputian	Martian	0.54
Liliputian	Venutian	0.44
Martian	Venutian	0.41

Visually



Gender

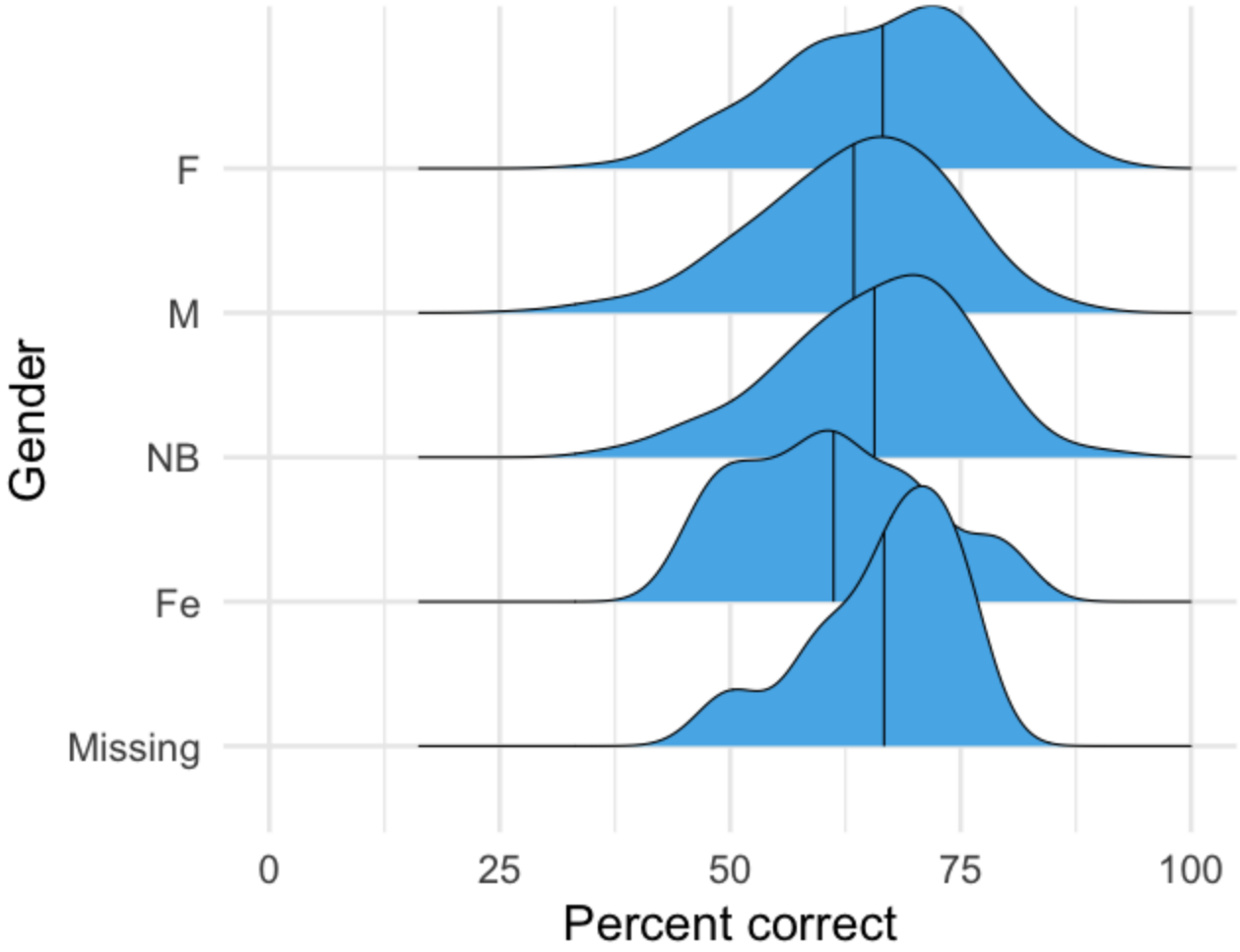
First, let's look at counts

Gender	n
F	98
Fe	1
M	104
NB	36
NA	1

Effect sizes

Focal Group	Reference Group	d
F	M	0.29
F	NB	0.08
M	NB	−0.21

Densities

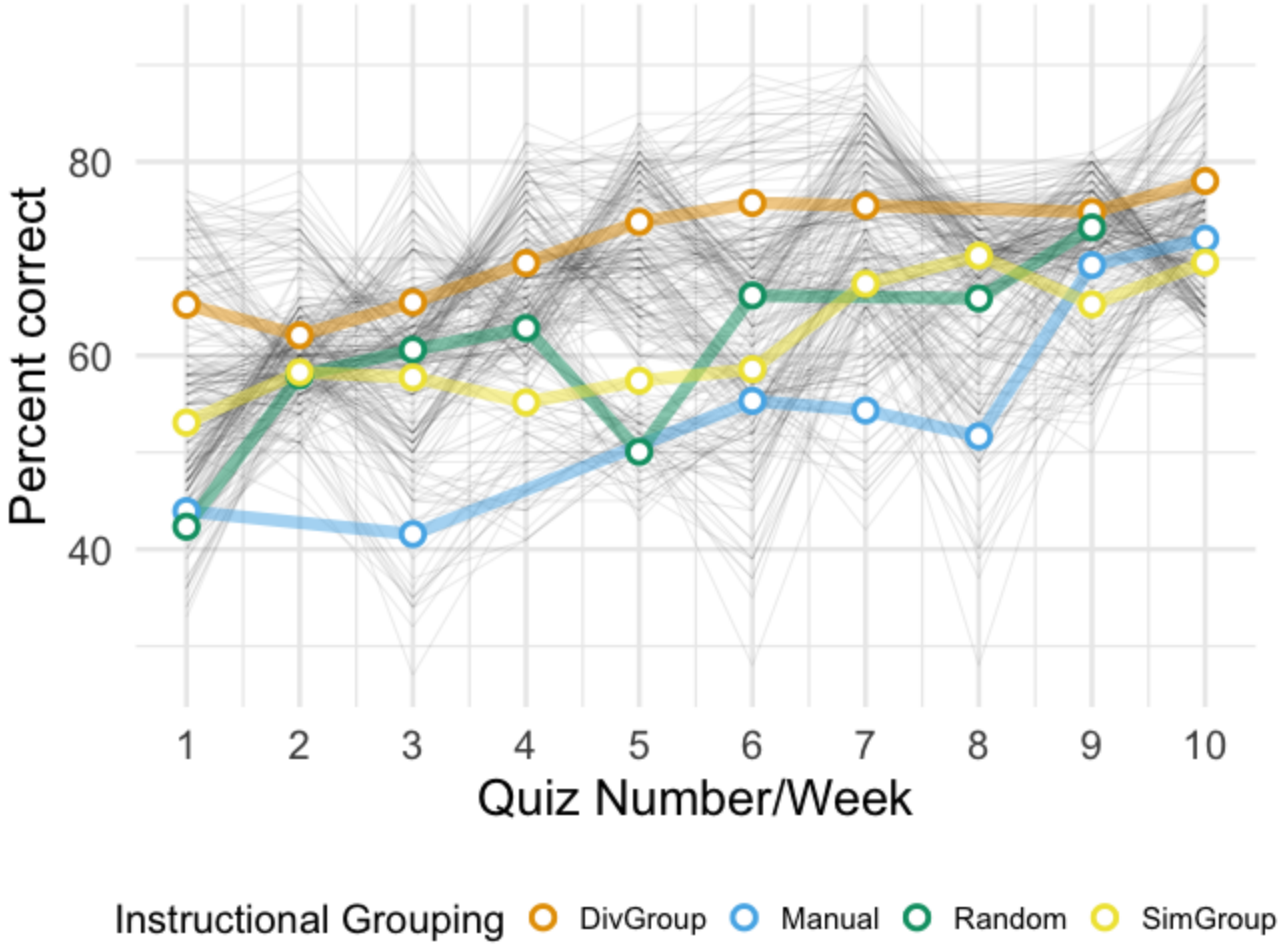


Random selection

Focal Group	Reference Group	auc
F	M	0.58
F	NB	0.53
M	NB	0.44

More complex modeling

Individual change



Process

- First model students' growth only
- Then add in instructional group
- Compare fit of models
- Add *race*, compare, add *gender* compare

If the fit of the model improves, so does our model accuracy

Growth vs Instructional groups

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
m0	4	16646.19	16669.23	−8319.097	16638.19			
m1	7	13416.33	13456.63	−6701.164	13402.33	3235.866	3	0

Lower values indicate better fitting model

Takeaway

Including instructional group *clearly* helps us explain why students get the scores they do

Pairwise comparisons

Contrast	Estimate	Lower Bound	Upper Bound
DivGroup – Manual	17.01	16.36	17.65
DivGroup – Random	12.85	12.20	13.50
DivGroup – SimGroup	12.87	12.38	13.37
Manual – Random	−4.16	−5.00	−3.31
Manual – SimGroup	−4.13	−4.82	−3.45
Random – SimGroup	0.02	−0.67	0.72

Add in Race

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
m1	7	13416.33	13456.63	−6701.16	13402.33			
m2	10	13415.68	13473.26	−6697.84	13395.68	6.64	3	0.08

Takeaway

We are not able to **better** explain why students scored the way they did by including *race*, after already accounting for their instructional group

Pairwise race comparisons

Contrast	Estimate	Lower Bound	Upper Bound
Atlantean – Liliputian	0.04	−3.42	3.49
Atlantean – Martian	1.90	−0.87	4.67
Atlantean – Venutian	−1.14	−3.96	1.68
Liliputian – Martian	1.86	−1.84	5.56
Liliputian – Venutian	−1.18	−4.91	2.56
Martian – Venutian	−3.04	−6.15	0.07

Drop race, add gender

Comparison is to instructional group only (while still accounting for individual student growth)

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
m1	7	13416.33	13456.63	−6701.16	13402.33			
m3	11	13413.03	13476.37	−6695.51	13391.03	11.3	4	0.02

Takeaway

Gender also does not help our explanation beyond what we already know from instructional group

Pariwise comparisons

Some evidence of a reliable male/female difference

Contrast	Estimate	Lower Bound	Upper Bound
Missing – Fe	7.60	–17.29	32.49
Missing – NB	3.82	–14.00	21.65
Missing – M	5.84	–11.83	23.51
Missing – F	2.94	–14.74	20.61
Fe – NB	–3.77	–21.63	14.08
Fe – M	–1.76	–19.46	15.94
Fe – F	–4.66	–22.36	13.04
NB – M	2.01	–1.39	5.41
NB – F	–0.88	–4.31	2.54
M – F	–2.90	–5.37	–0.42

Final takeaways

> What groups of students, based on common attributes, are consistently underperforming on their quizzes?

- Manual instructional grouping
- Little evidence that race is an important predictor
- Some evidence that Male students score lower than Female, on average

Final takeaways

> Are there any types of groups, e.g., manual groups vs calculated (SimGroup, DivGroup, Random), that you would recommend teachers use or avoid using?

- **DivGroup** was consistently the highest performing

Other possible areas to explore

- Could consider collapsing gender in some way
- Could explore interactions
    - It's possible, for example, that we could find greater gender or race differences *within* an instructional grouping
- Probably *should* do a few more robustness checks before making implementation decisions

Thank you!

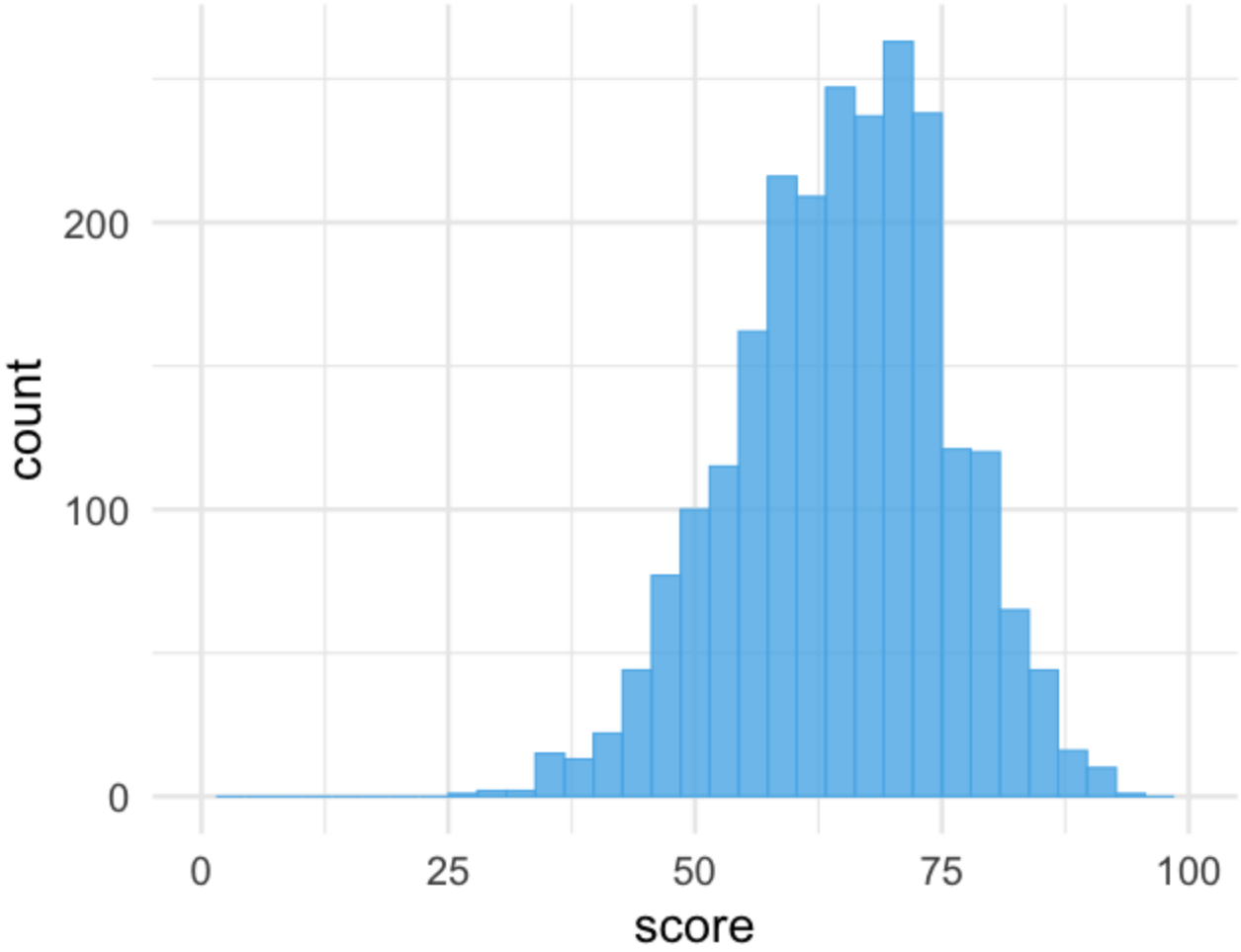
Questions?

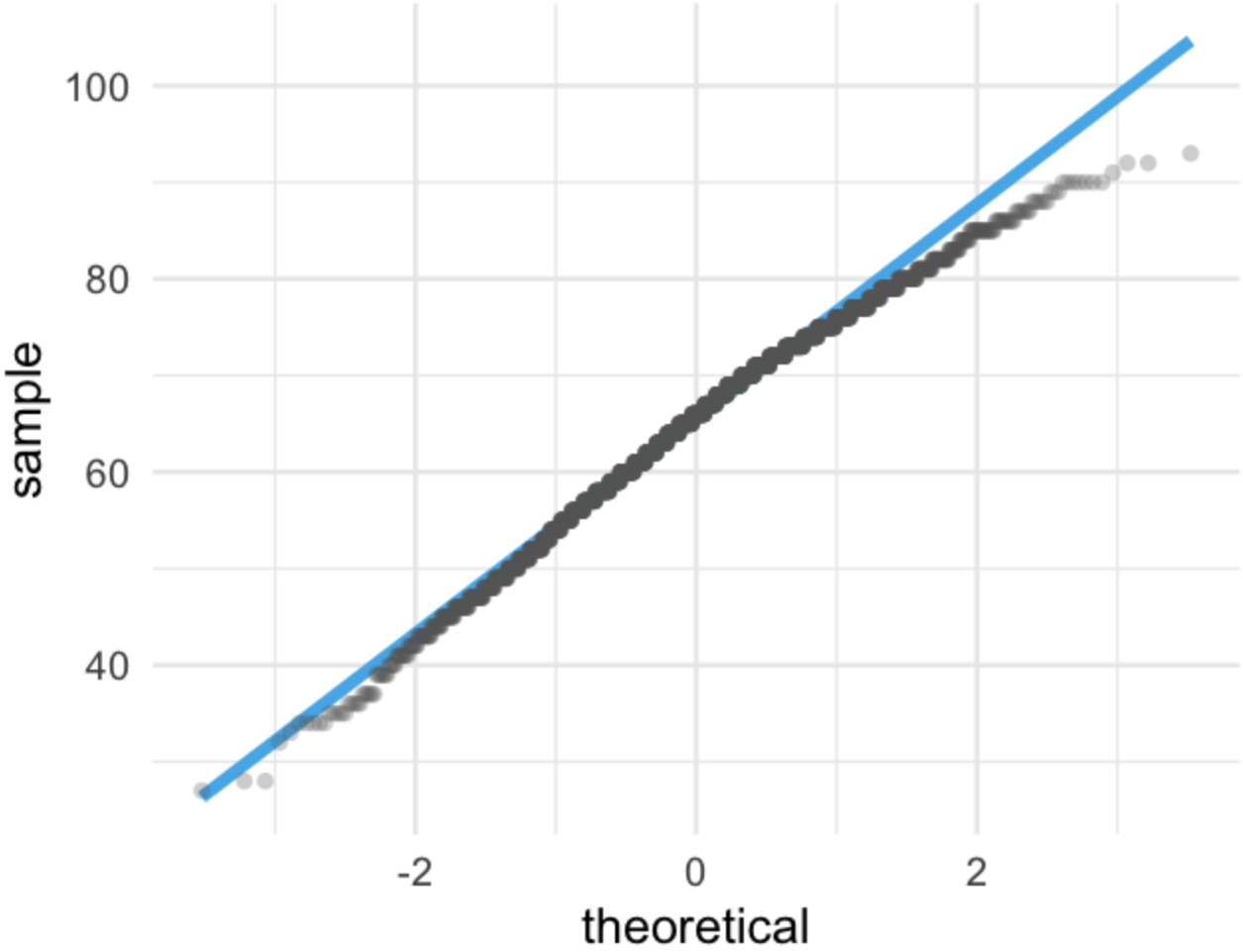
Bonus slides!

A technical note on percents

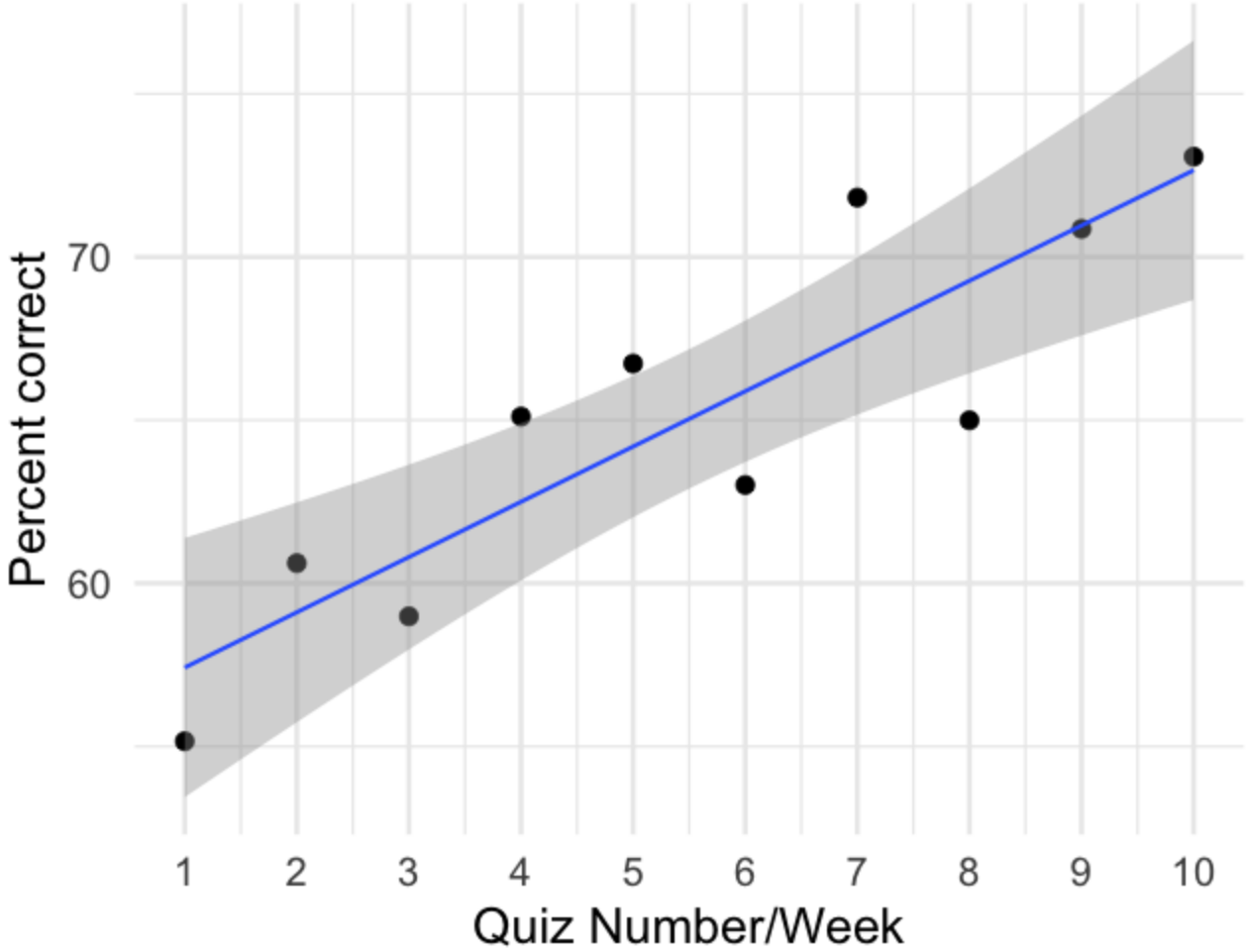
- Regularly modeled as a standard continuous outcome
- This is sometimes problematic
    - Bounded response
- In this case, we can get away with it without much problem

Evidence for this claim




100
80
60
40
sample
-2
0
2
theoretical

Linearity


Percent correct
70
60
1
2
3
4
5
6
7
8
9
10
Quiz Number/Week