Foundations of Evaluation

Katrina Ballard, Office of Migrant Education, Data and Evaluation Subject Matter Expert

Erin Pollard, Institute of Education Sciences, What Works Clearinghouse and ERIC

The mission of the Office of Migrant Education is to provide excellent leadership, technical assistance, and financial support to improve the educational opportunities and academic success of migratory children, youth, agricultural workers, fishers, and their families.

2024 OFFICE OF MIGRANT EDUCATION ANNUAL DIRECTORS' MEETING

DEPARTMENT OF EDUCATION UNITED STATES OF AMERICA

Introduction

- Evaluation requirement
  - Promising evidence in Selection Criteria for HEP and CAMP since 2016
  - Now collecting (2023-24 optional, required moving forward)
- Performance evaluation vs. evaluation producing promising evidence
- Understanding your application is key.
- Think ahead
- Session recording

Big picture for today

- Document and share your evaluations– we want to learn from each other!
- Follow your evaluation plans
- Going forward, think about how to add rigor

Agenda

- What are program evaluations?
- What do we mean by evidence?
- How to pick a good research design?
- What should we watch out for?
- How do we actually do this?
- How do we share our findings?

A program evaluation will tell you:

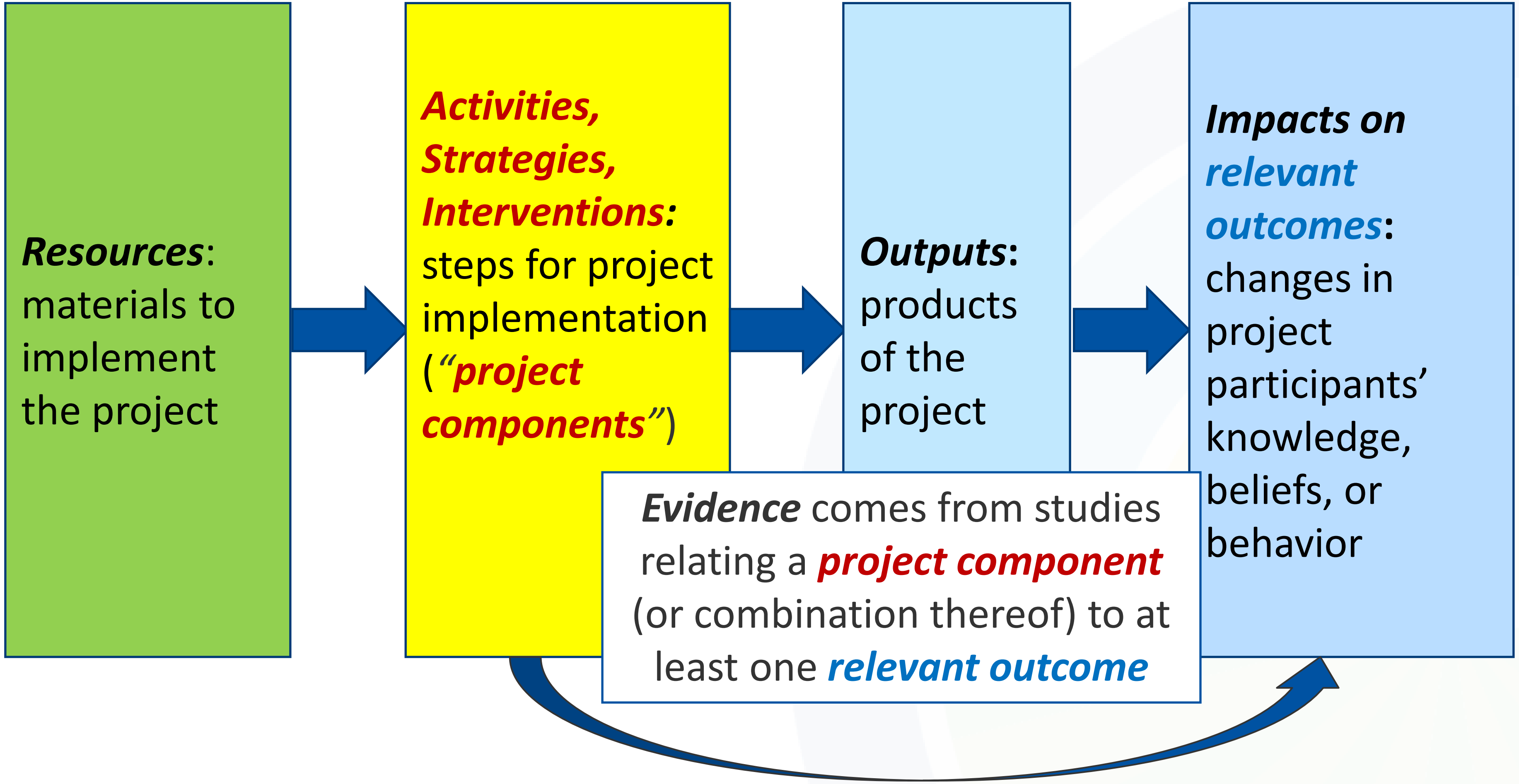
- What is the program?
- How was it implemented?
- Costs and resources necessary to implement the program
- Program participants
- What do the program performance measures tell you?
- **Is the program likely the reason for student successes?**

This contextual data is often the most important part of the evaluation. Documenting the program is essential.

What do you aim to do?

- Think about what your program wants to do:
  - What is your program?
  - Who do you serve?
  - How do you serve them?
- What are you aiming to do?
- How do you know if you are successful?

First think about your logic model– How is your program supposed to work?



Resources: materials to implement the project

Activities, Strategies, Interventions: steps for project implementation (*"**project components**"*)

Outputs: products of the project

Impacts on relevant outcomes: changes in project participants' knowledge, beliefs, or behavior

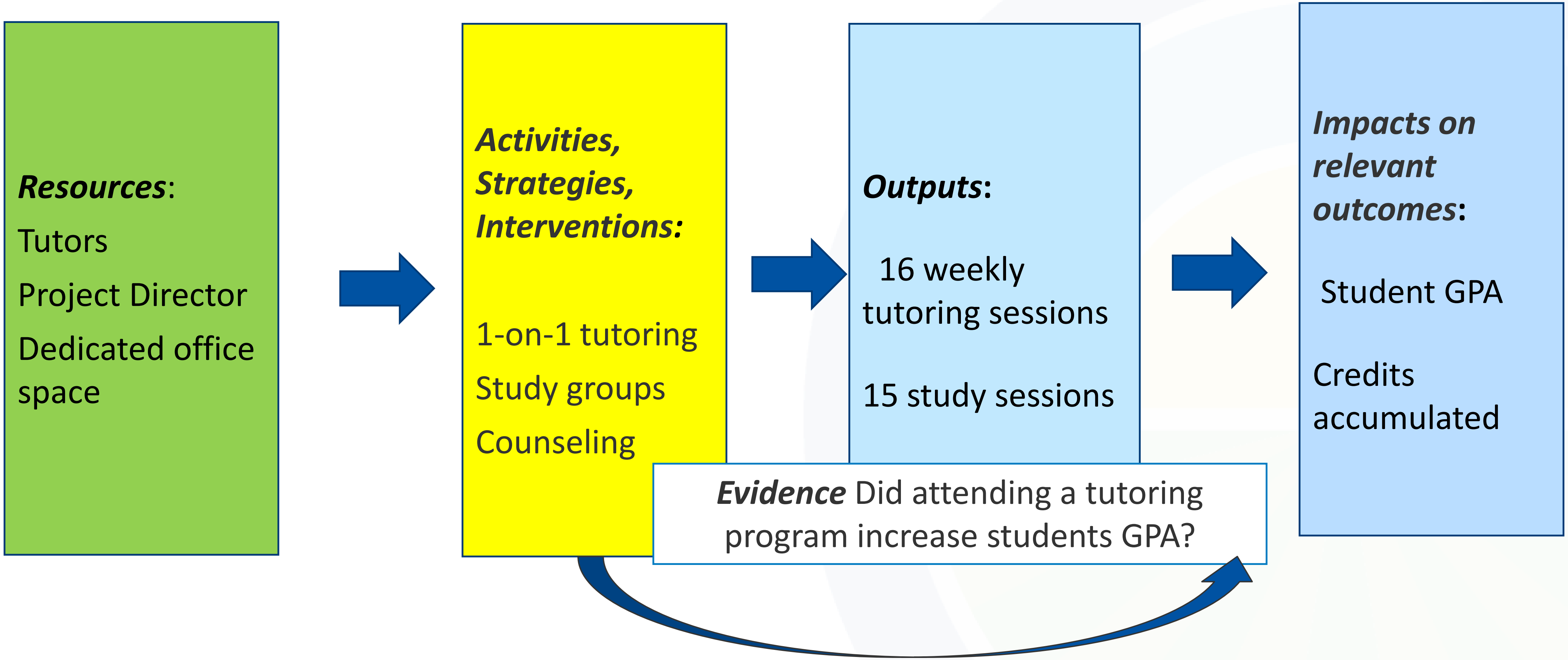
Evidence comes from studies relating a ***project component*** (or combination thereof) to at least one ***relevant outcome***

From: https://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=REL2014025

Example:

- Think about what your program wants to do:
  - What is your program? A set of tutoring activities for students identified at risk for dropping out
  - Who do you serve? Students are at risk of dropping out due to being first generation, low income, and from lower performing high schools
  - How do you serve them? One-on-one tutoring, study sessions, and counseling
- What are you aiming to do? Keep students in school so they graduate
- How do you know if you are successful? Students have the grades and credits to stay in academic good standing. Students graduate.

Example: Tutoring Program



Resources:

Tutors

Project Director

Dedicated office space

Activities, Strategies, Interventions:

1-on-1 tutoring

Study groups

Counseling

Outputs:

16 weekly tutoring sessions

15 study sessions

Impacts on relevant outcomes:

Student GPA

Credits accumulated

Evidence Did attending a tutoring program increase students GPA?

Research Questions

- All good evaluations need a few key questions you aim to answer. Consider these questions:
    - Is the tutoring program effective?
        - What tutoring program?
        - How do you define effectiveness?
    - Do students who participate in Erin’s Tutoring program have higher GPAs than students who do not?

What do we mean by evidence?

Basic Evaluation

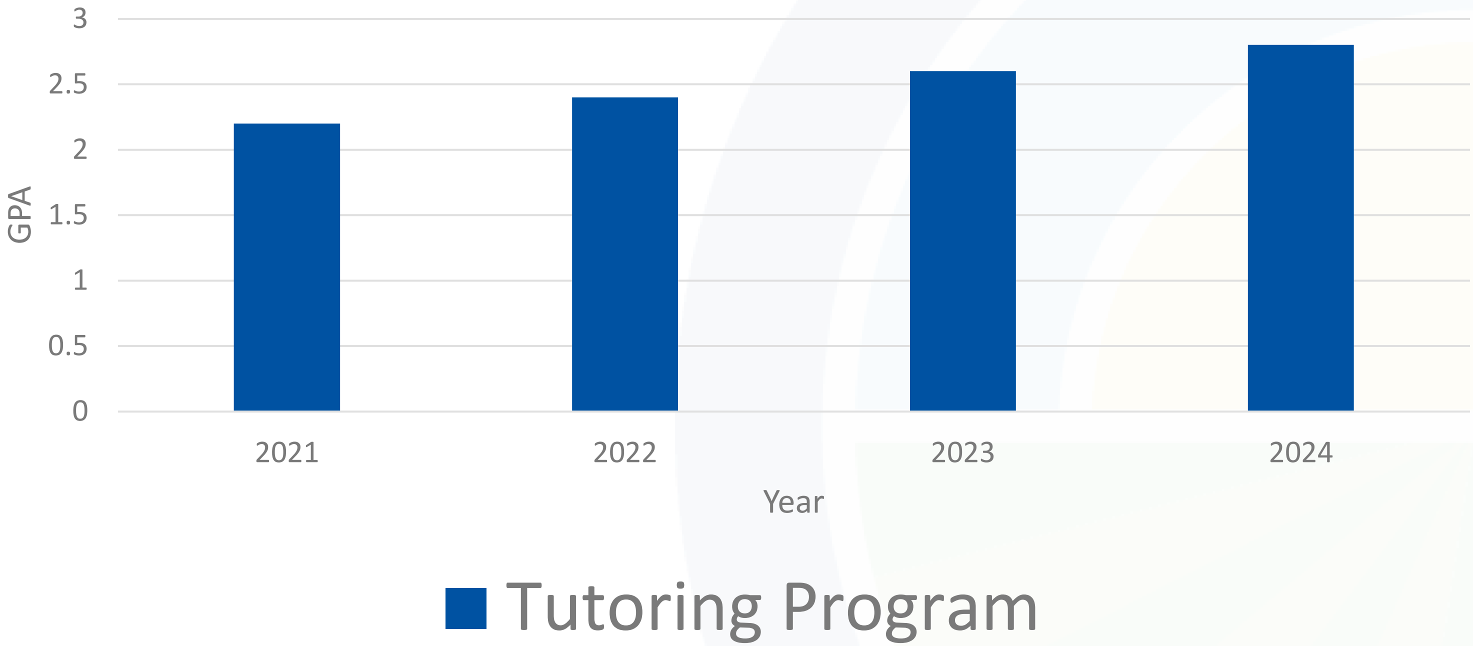
- Did students who received the program do better than those who did not?

What about pre-post studies without a comparison group?

- Programs often do a pre-post study, or looking at changes in student outcomes from the beginning of the intervention to the end.
- This will likely not qualify as promising evidence
- It is better to have an imperfect study with a comparison group than a pre-post design.

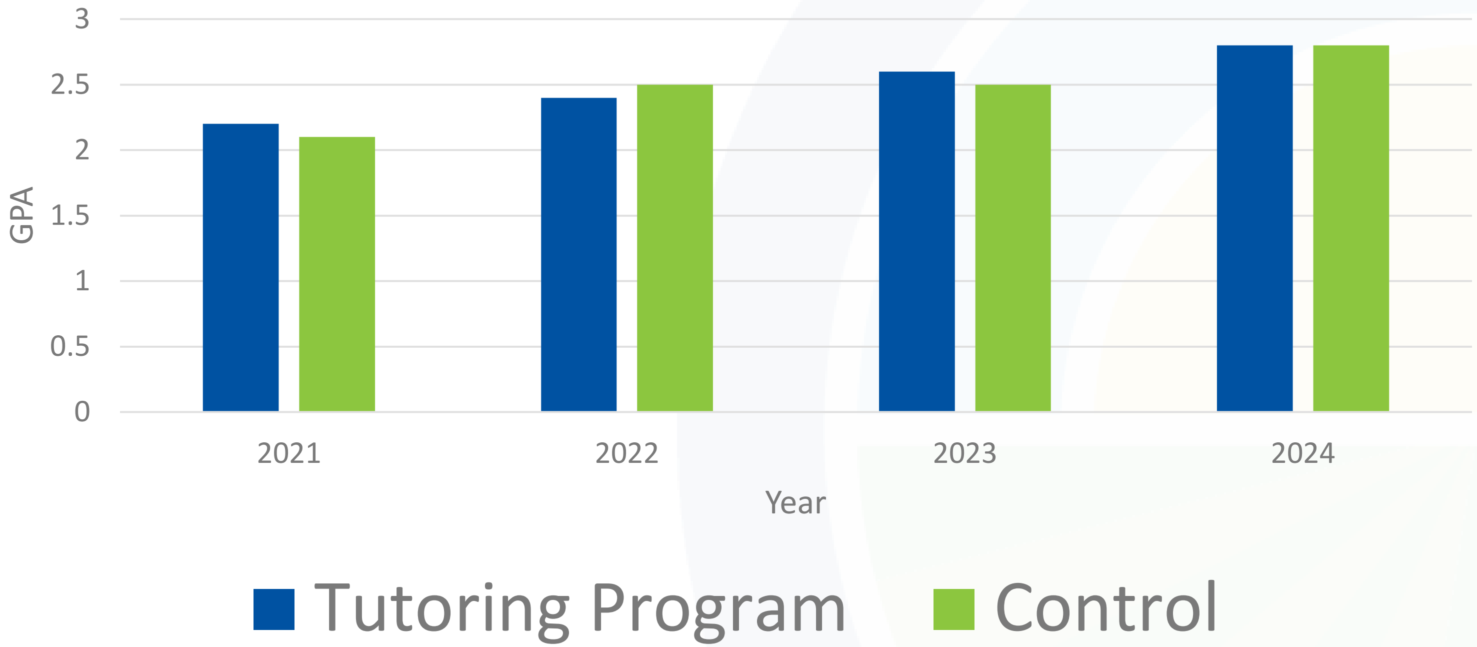
Comparison groups

Allow you to compare two different groups of students, because sometimes differences are not what you think:



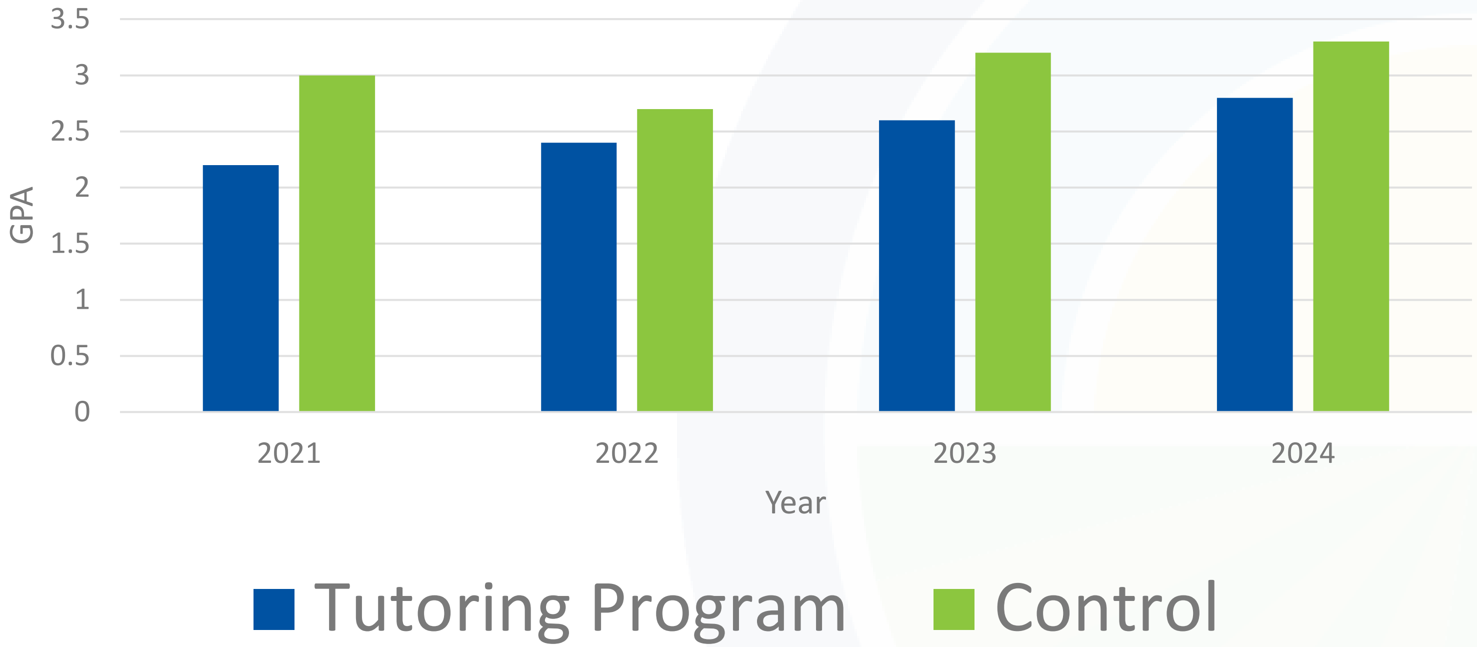
Comparison groups

Allow you to compare two different groups of students, because sometimes differences are not what you think:



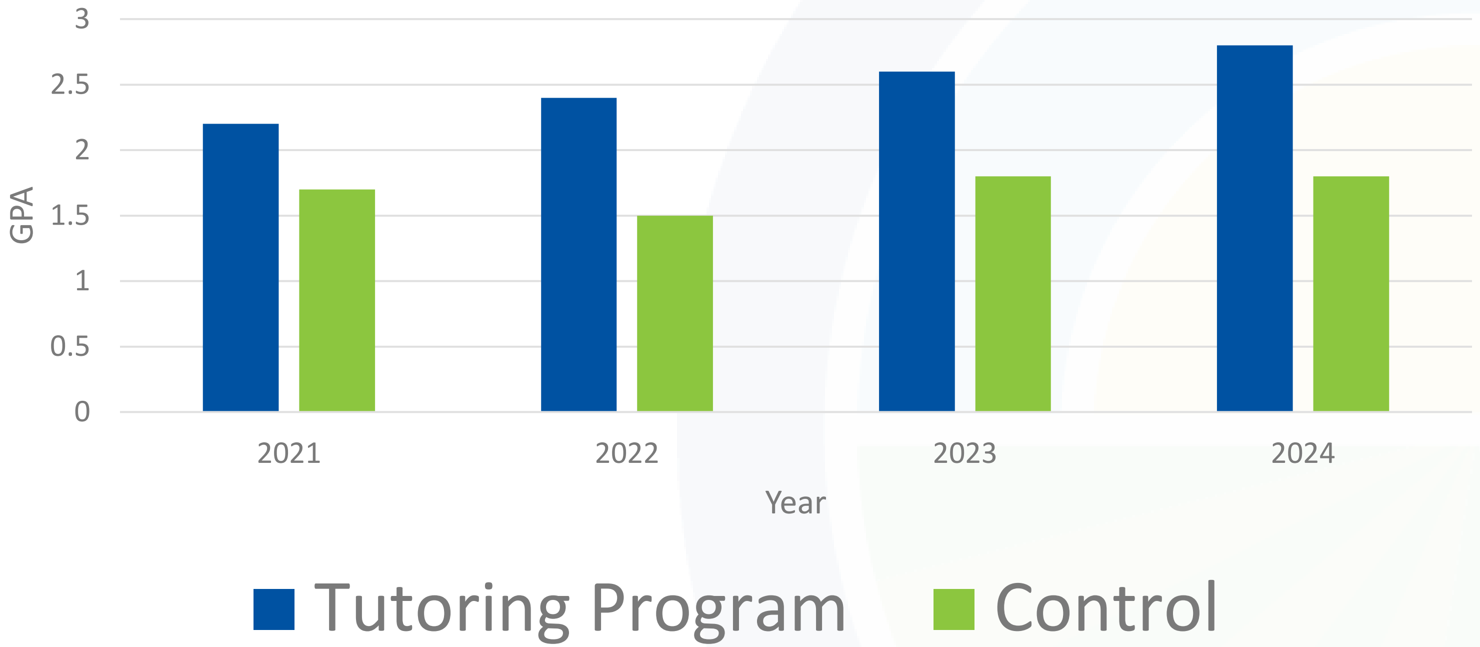
Comparison groups

Allow you to compare two different groups of students, because sometimes differences are not what you think:



Comparison groups

And sometimes the results are exactly what you expect and can help you justify your program works:



Evidence Tier	TIER 1 STRONG	TIER 2 MODERATE	TIER 3 PROMISING	TIER 4 DEMONSTRATES RATIONALE
Study design	Well-designed and well-implemented experimental	Well-designed well-implemen d quasi-experim al	Well-designed and well-implemented correlational design with statistical controls for selection bias	W defined logic model
Positive, statistically significant effect on the outcome	★	★	★	Re d research or luation is anned or nderway
No overriding negative effects	★	★	★	
Large, multisite sample	350+ students across multiple sites	350+ students acros multiple sites		
Relevance to proposed context	Population **and** setting	Population **or** setting		

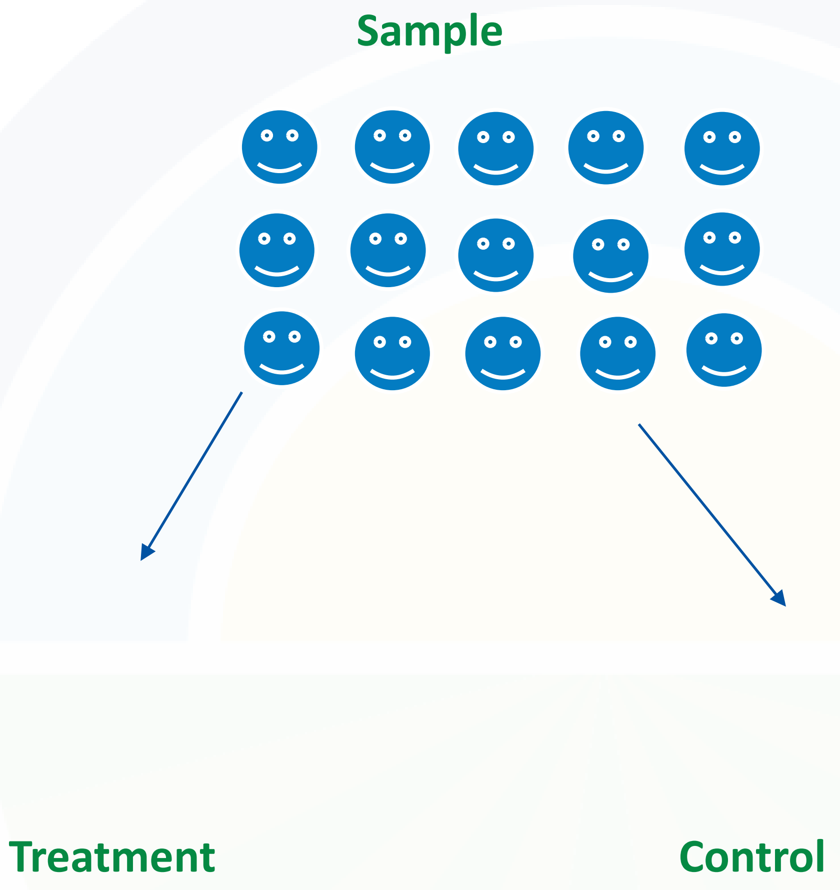
Experimental research designs

- Can tell you “attending our tutoring program causes a 0.5 point increase in GPA”

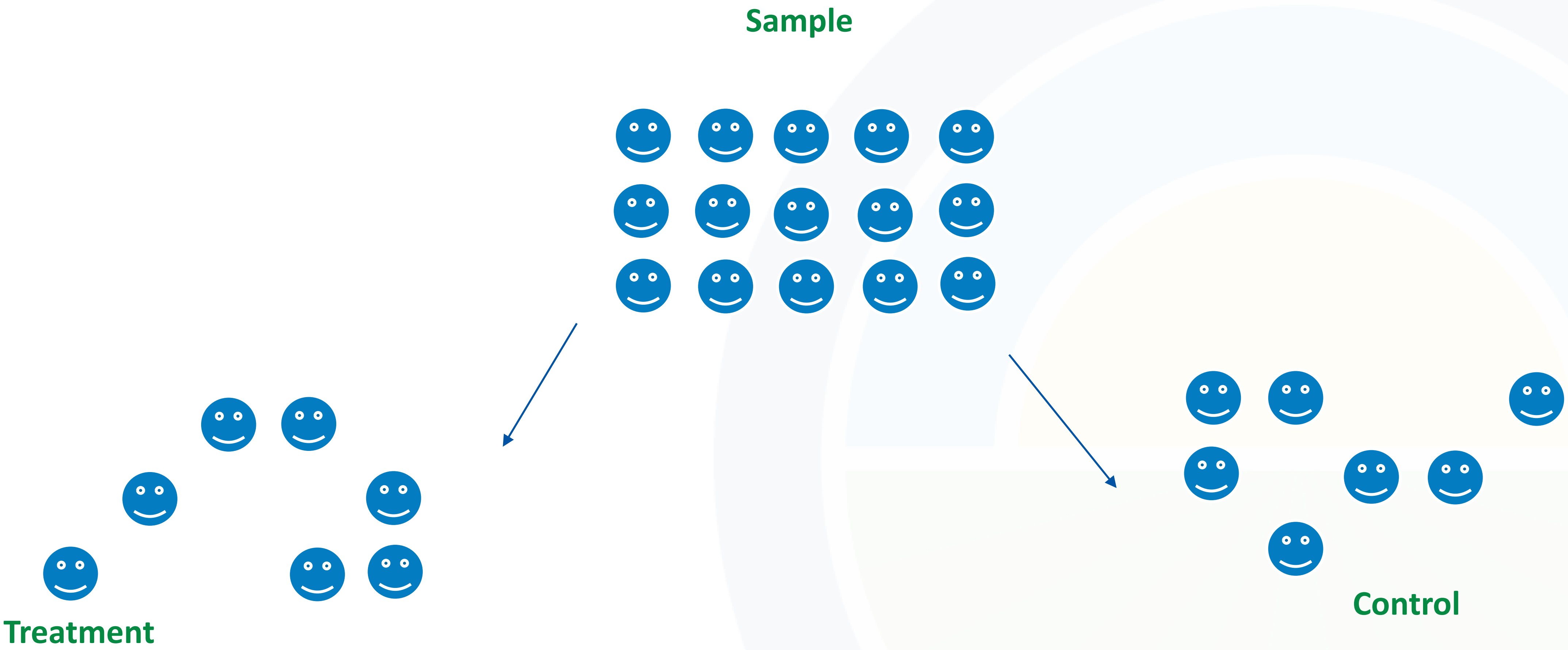
- Types:
  - Randomized control trials
  - Regression discontinuity designs
  - Single case designs

Types of evaluations: Randomized Control Trials

- Two groups of participants assigned at random
- One group gets the intervention, one does not
- Statistically, the is the purest type of assessment
- Practically, it is difficult to do for programs that already exist



How does randomization work?



Little bit of randomization?

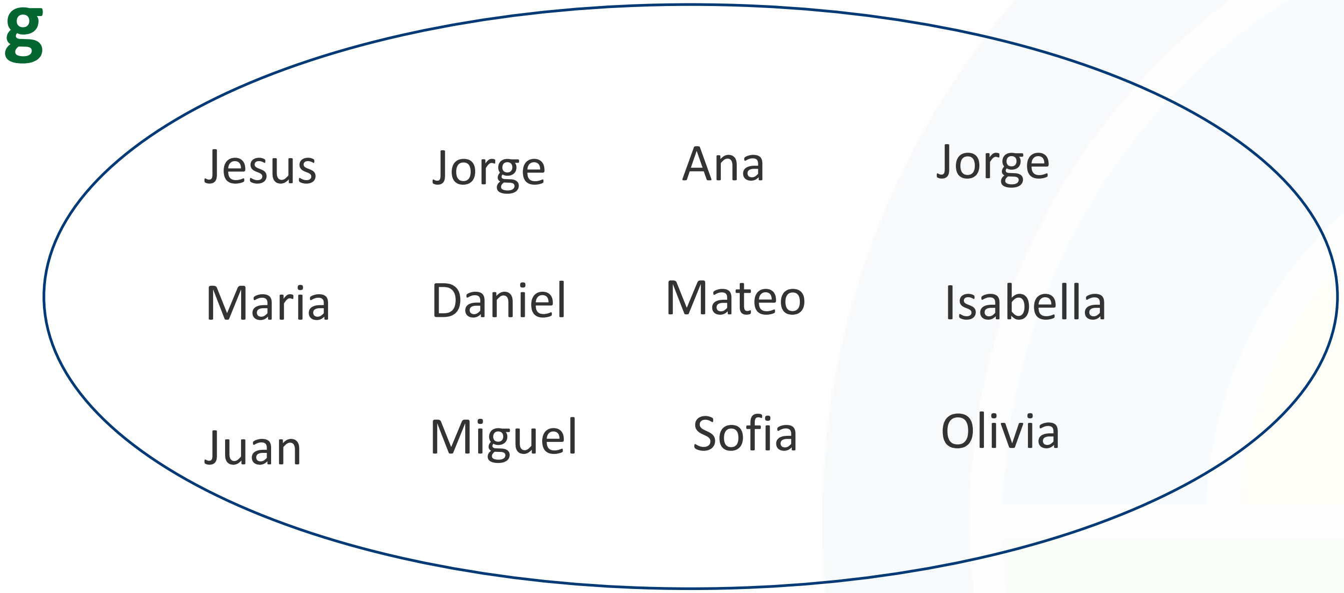
Tutoring

Jorge

Mateo

Jesus

Juan



Jesus Jorge Ana Jorge

Maria Daniel Mateo Isabella

Juan Miguel Sofia Olivia

Study sessions

Ana

Daniel

Miguel

Olivia

Counseling

Miguel Sofia

Maria Isabella

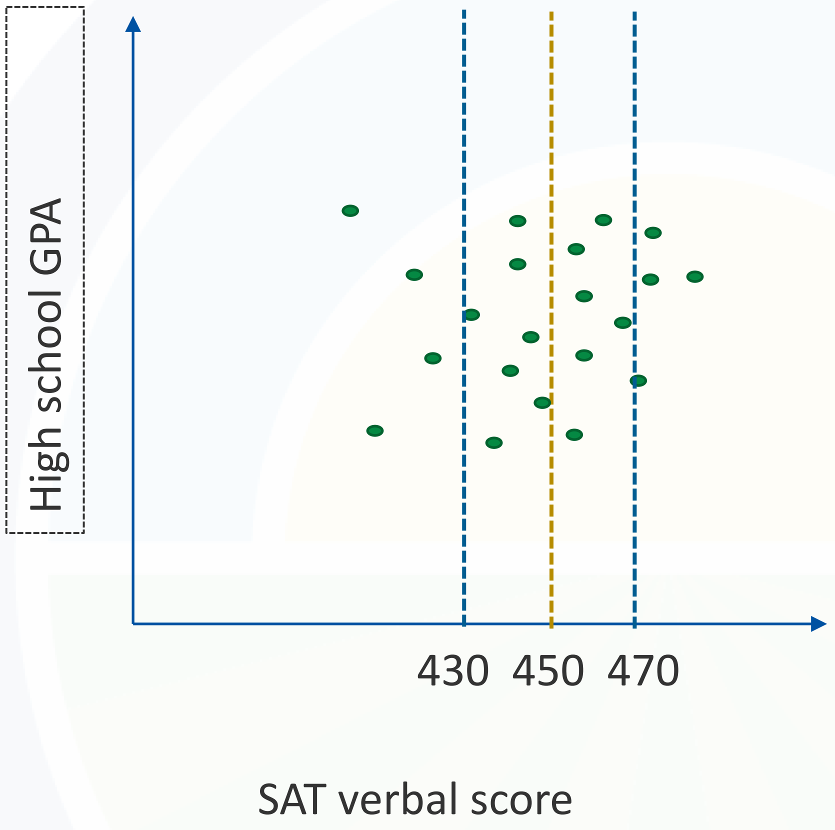
Randomization a different way

Tutoring	Jorge	Mateo	Jesus	Juan
Tutoring + study sessions	Ana	Daniel	Miguel	Olivia
Tutoring + study sessions + counseling	Miguel	Maria	Isabella	Sofia

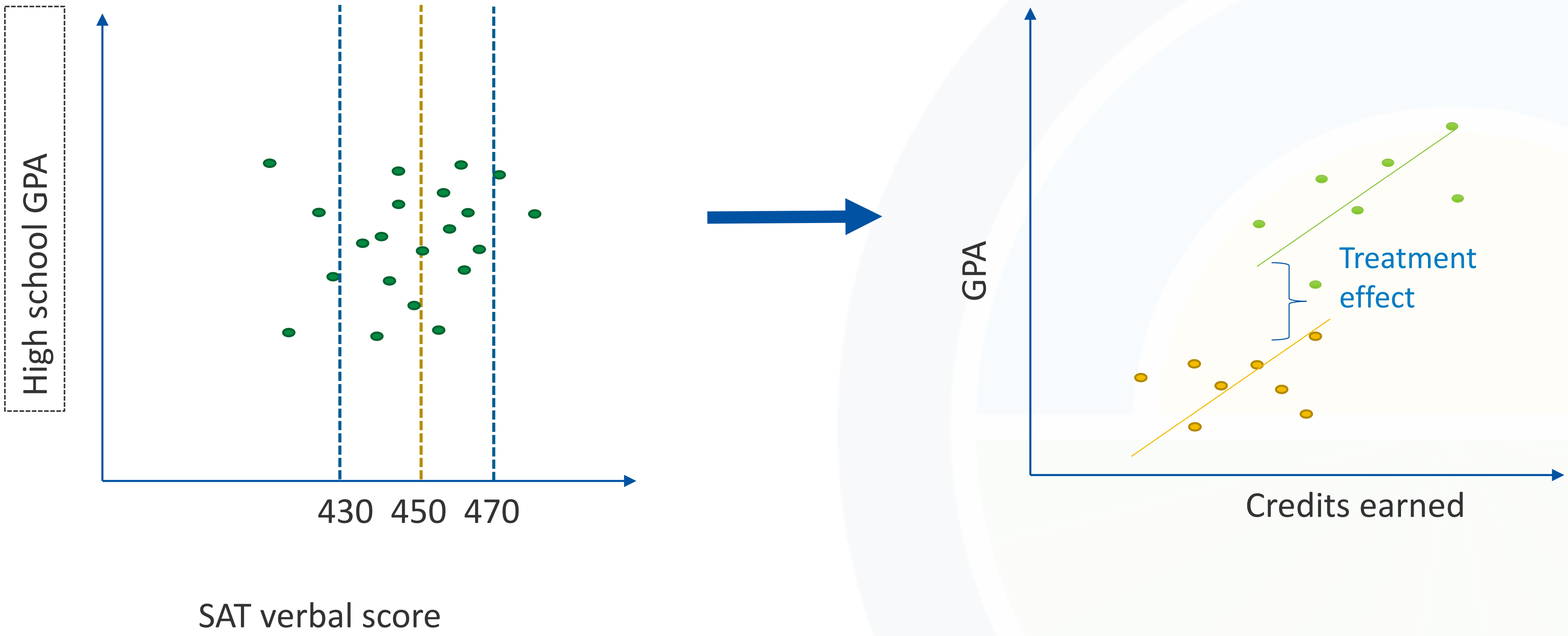
Types of evaluations: Regression Discontinuity Design

- There is some sort of arbitrary cutoff where some students get the program and some do not

- Example: remediation based on SAT verbal score– those below 450 get a tutoring



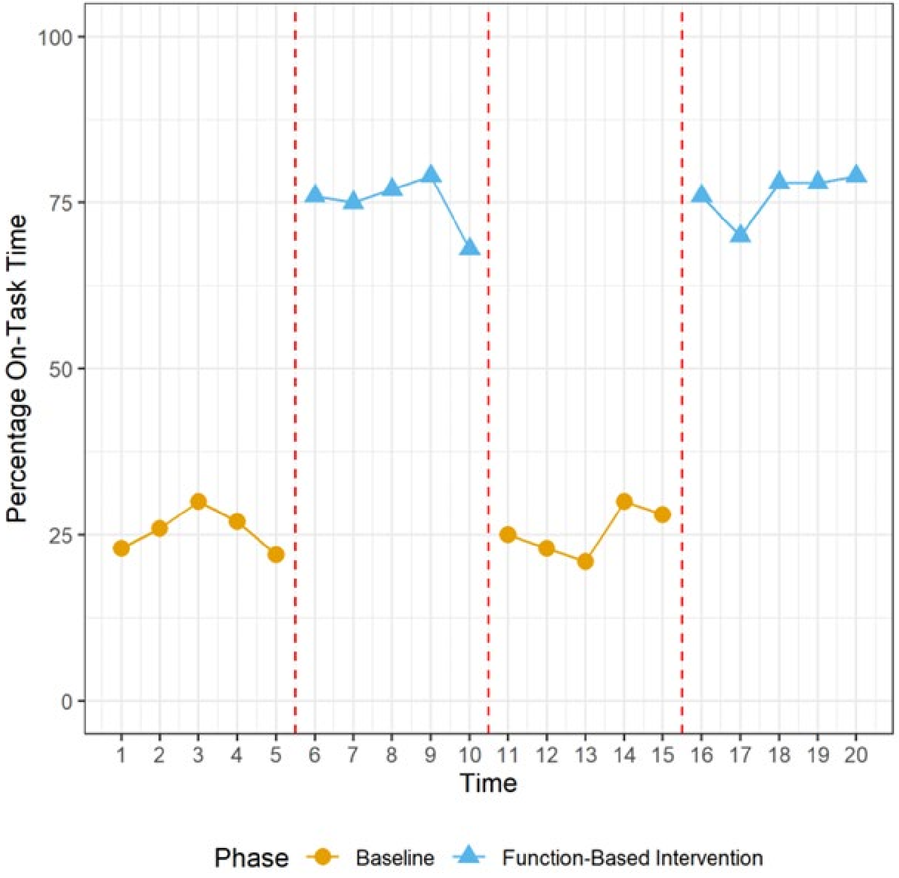
Types of evaluations: Regression Discontinuity Design



Types of evaluations: Single Case Design

- Each unit is repeatedly observed on one or more outcomes multiple times in a series, with observations broken into phases.

- This is normally used in special education and students with behavioral problems.

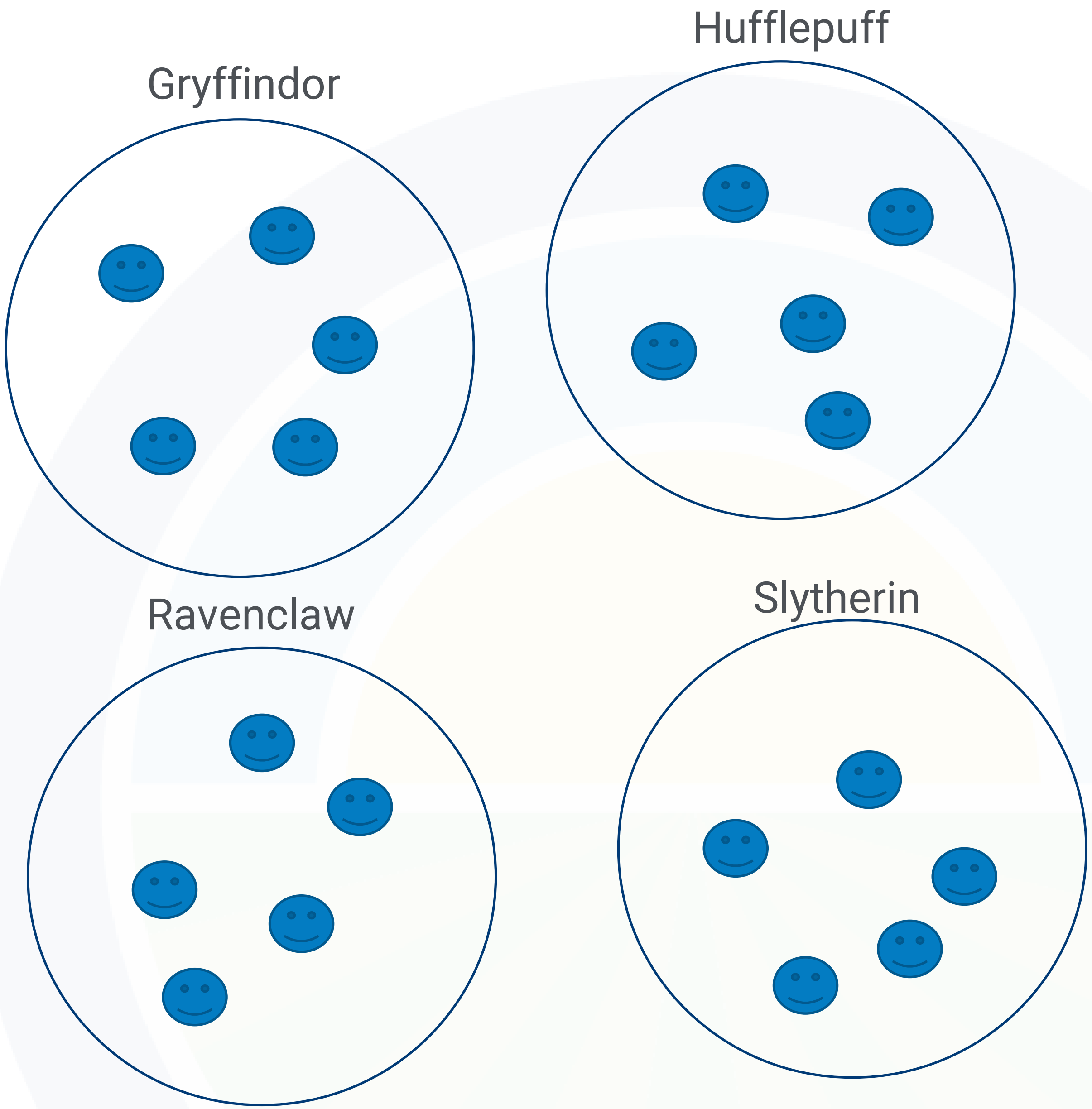


Quasi-experimental research designs

- Can tell you “attending our tutoring program likely causes an increase of 0.5 GPA points”
- Two groups of participants assigned by some non-random factor. One group gets the intervention, one does not

Quasi-experimental designs

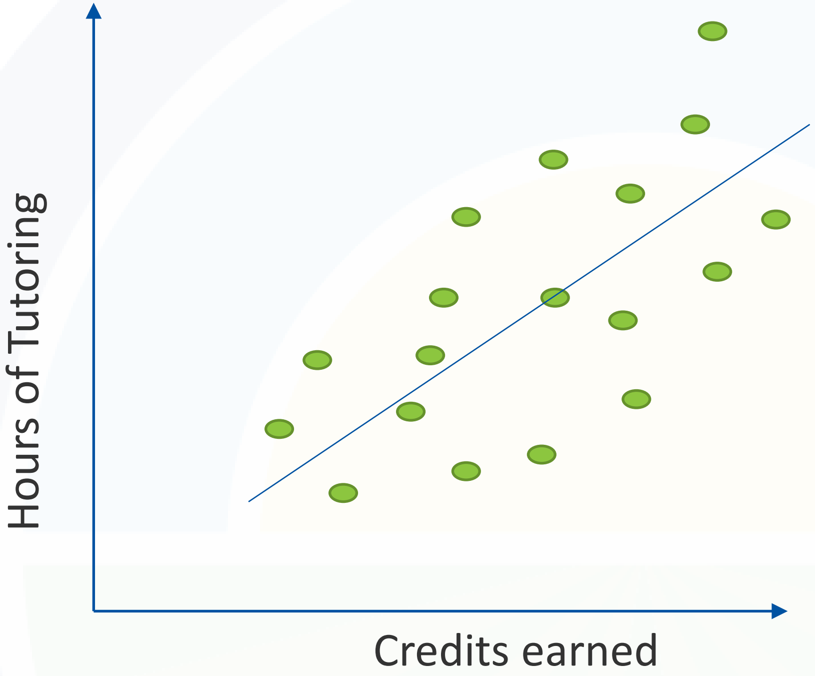
- Ways to assign students:
  - Residence Halls
  - Campus
  - Volunteering
  - Proximity
- Normally have some sort of statistical controls



Correlational Designs

- Can tell you “participating in our tutoring program is associated with a 0.5 GPA point increase”

- Use statistical methods to look at the effect of a program
- No comparison group
- Great for quick understandings of relationships



Questions?

Brain Break

How does this relate to what you proposed?

- Look at your evaluation plan and ask what type of evaluation did you propose?

What to think about?

Why do you need a comparison group?

Students who participated in the second semester tutoring program had GPAs that were 0.5 higher than their first semester

Were the classes easier second semester?

Did the students not ready already fail out?

Did the new grading policy cause higher GPAs?

Did the students who were struggling switch majors?

Did everyone's GPA go up?

How to create a comparison group

- Students who were eligible for the program, but didn’t apply
- Students who applied for the program, but there was not space
- Students who are demographically similar to those in the program
- Students who participate in a similar program for similar demographics

Why is this better?

Students who participated in the freshman tutoring program had GPAs that were 0.5 higher than similar freshman who did not participate

They took the same classes

They had the same prior achievement

They were impacted by the same changes to policy

They would have had the same ability to drop classes and switch majors

Their GPAs should have been influenced by the same factors

Less great ways to make comparison groups

- Using students at other campus
- Using students from a previous cohort
- Using students who are not similar

Why is this better?

Students who participated in the freshman tutoring program had GPAs that were 0.5 higher than similar freshman at a branch campus without the program

They took the same classes, but by different professors

They had the same prior achievement

There might be different policies impacting each group

They would have had the same ability to drop classes and switch majors

Their GPAs could be influenced by different factors

What if there is no obvious control group?

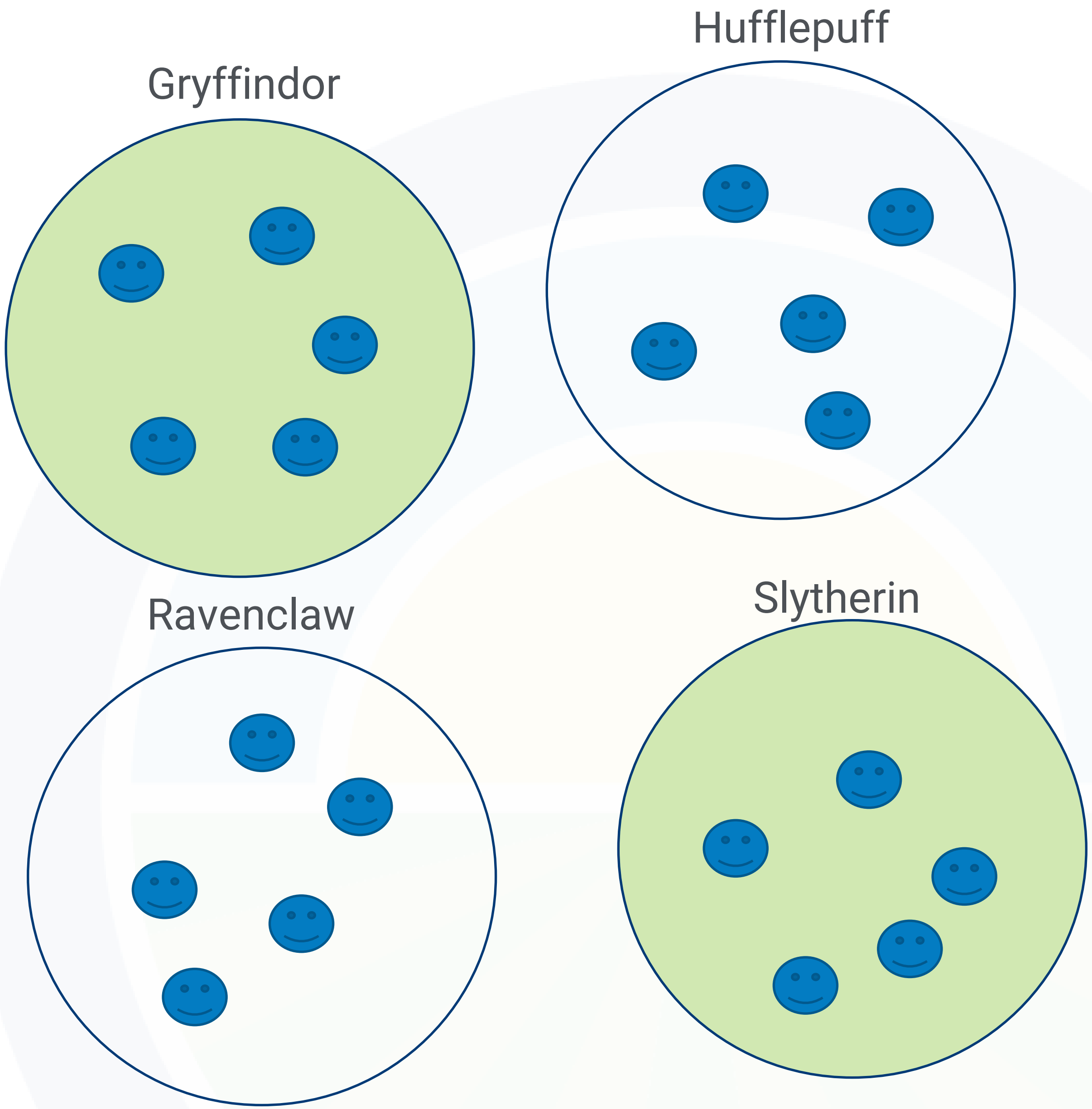
- You do not need to evaluate all of your students– is there a small group that might be comparable to others?
- You don’t need to do a treatment vs. nothing comparison, can you give half of your students one program and half a different? Or the program + something extra?
- Are there similar programs in the community? For example, state funded preschool and Head Start both serve low-income 4 year olds.

Things to worry about

- **Baseline equivalence:** Are the groups similar?
- **Power:** Do you have a big enough sample to see the impact?
- **Confounding variables:** What else could be causing this?

Things to worry about: Baseline equivalence

- Sometimes groups may differ on characteristics that are likely related to your outcome of interest. Let's say we are trying to increase GPA through a tutoring program:
  - What if honors students preferred Gryffindor and Slytherin?
  - What if Ravenclaw had both freshmen and sophomores?



Things to worry about: Power

- Power is the likelihood of detecting an effect when there actually is one.

- Power is largely comprised of sample size, effect size, and significance level.

- You can do a power analysis to determine the smallest sample size likely needed to find a positive effect if the program actually works.

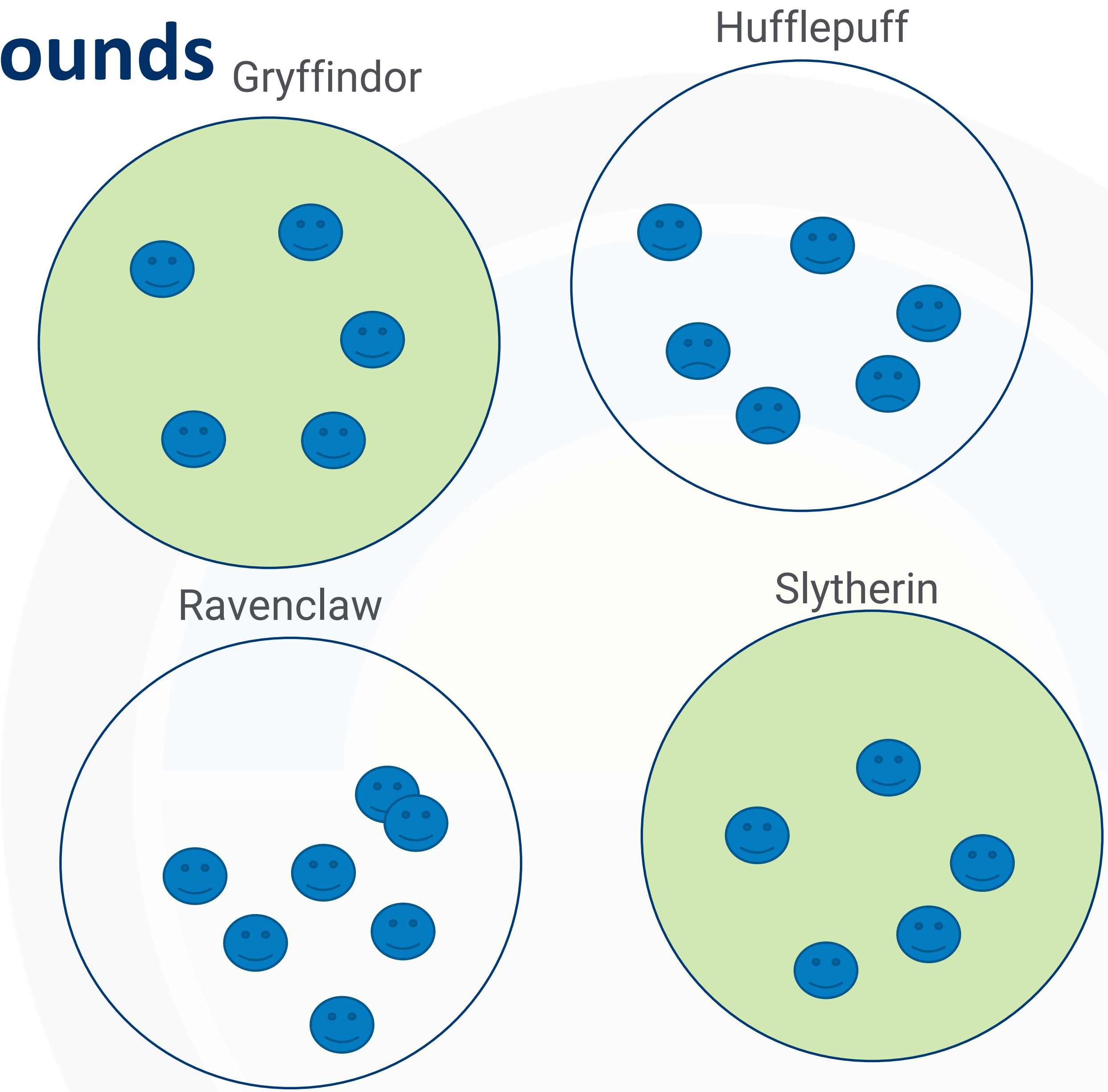
Things to worry about: Power

- Your program is effective, but the sample size is too small to determine an effect
  - This means you wasted resources doing the evaluation
  - You do not know if the program works or not

- Your sample sizes are so large that you find statistically significant differences, but they are so small to have practical meaning
  - This greatly increases costs

Things to worry about: Confounds

- Confounding variables are when there is an observed characteristic that could be responsible for the change, other than the intervention
  - Let’s say the yellow dots are non-athletes.



When to add in statistical controls (covariates)

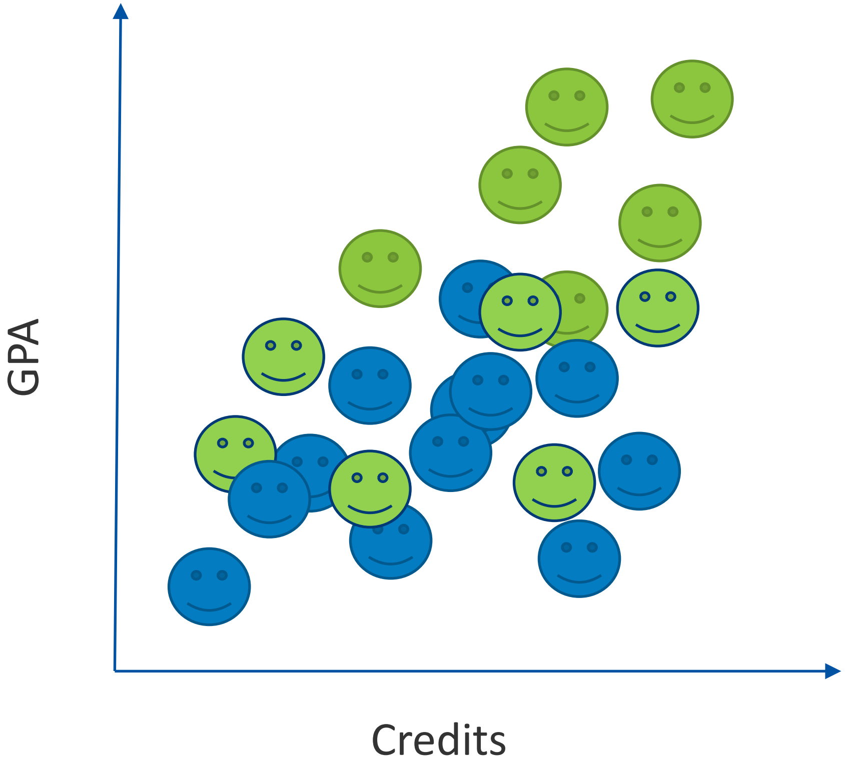
- You expect that there are observable traits that could cause the effect.

- Common controls:
    - Prior achievement
    - Demographic
    - Income/need
    - Language

- The way to think about this: how much of the effect can be explained by other factors?

- Include controls on places where there is not baseline equivalence

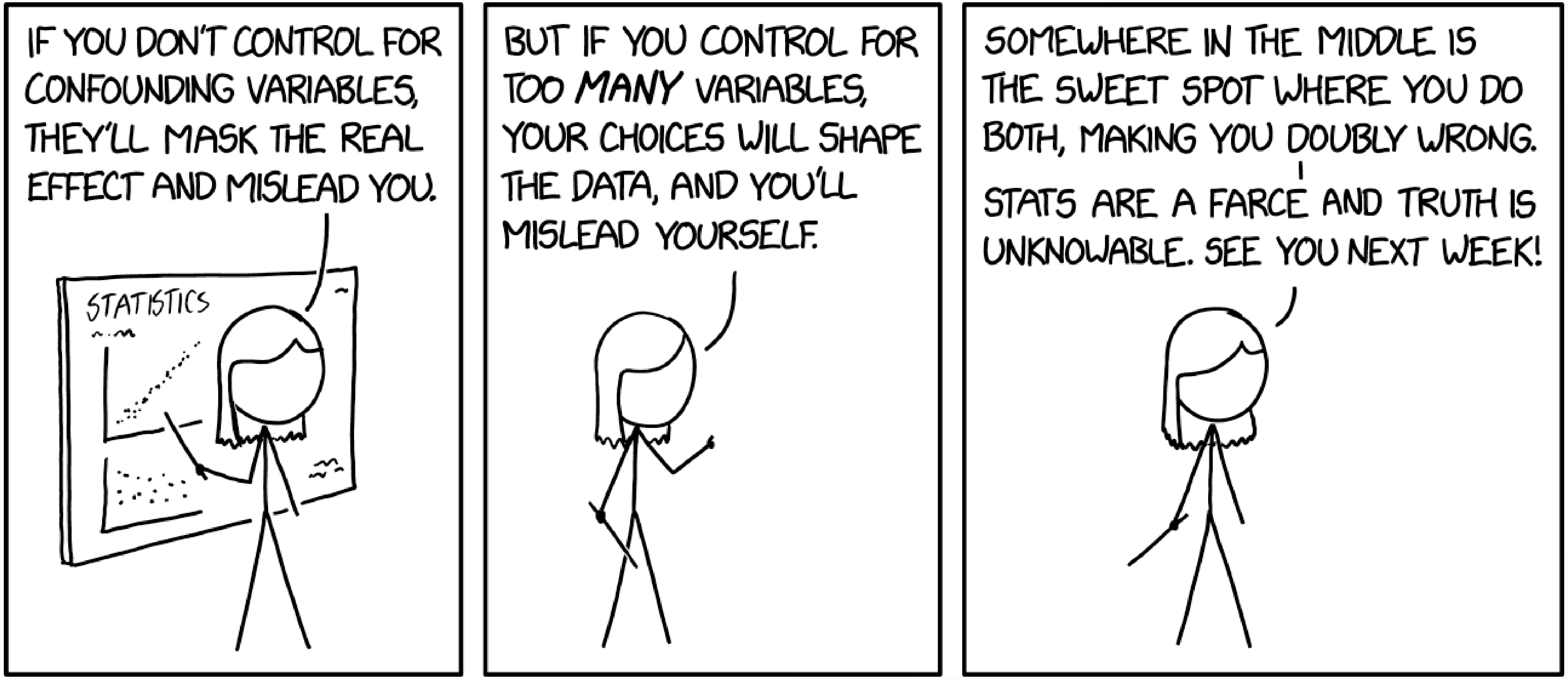
Things to help: Covariates



This program looks effective. But is it?

What if we control for SAT score?

So what have we learned?


IF YOU DON'T CONTROL FOR CONFOUNDING VARIABLES, THEY'LL MASK THE REAL EFFECT AND MISLEAD YOU.
STATISTICS
BUT IF YOU CONTROL FOR TOO *MANY* VARIABLES, YOUR CHOICES WILL SHAPE THE DATA, AND YOU'LL MISLEAD YOURSELF.
SOMEWHERE IN THE MIDDLE IS THE SWEET SPOT WHERE YOU DO BOTH, MAKING YOU DOUBLY WRONG.
STATS ARE A FARCE AND TRUTH IS UNKNOWABLE. SEE YOU NEXT WEEK!

Questions?

Brain Break

How does this relate to what you proposed?

- Look at your evaluation plan and ask:
    - How are you going to make a comparison group?
    - How will you determine baseline equivalence?
    - What types of statistical controls or covariates did you propose? Why?
    - Did you address power and sample size?

How does this relate to what you proposed?

- Find 2-3 colleagues, discuss:
  - What type of evaluation are you using? Why are you using it?
  - How are you going to create a comparison group?
  - What types of statistical controls are you using?

Group questions

- Do you have a plan for creating a control group? What is it?
- What types of covariates are you using?
- What concerns do you have?

So how to you actually do this?

So how do you add statistical controls?

- Almost always you use a form of **regression:**

$$Y=mX +b + \in$$

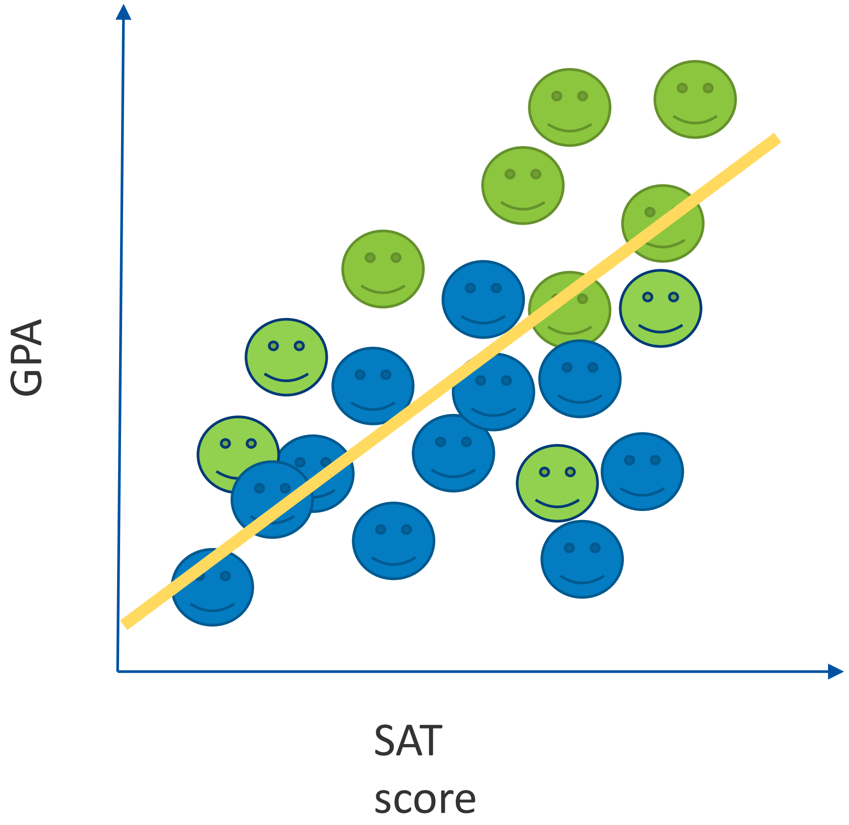
Regression example

- Let’s say we are trying to improve freshman GPA through a tutoring program. A basic model would be:

GPA= (amount of increase * SAT score) + minimum GPA + error

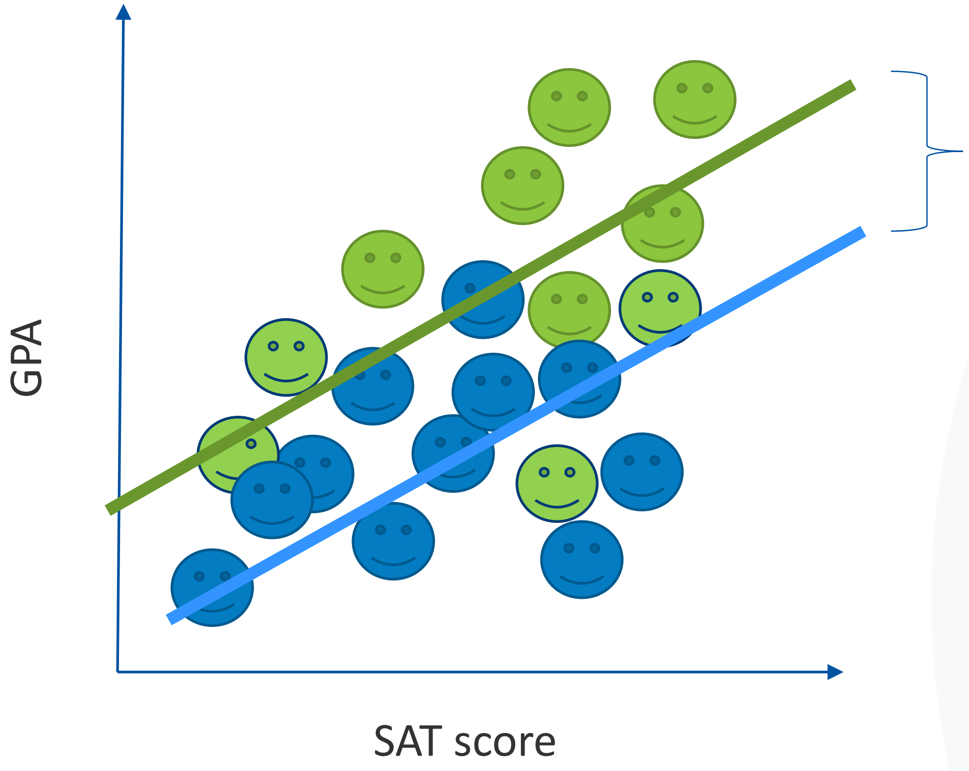
** But why SAT score?

So let's look at basic correlation relationship



$$Y = mX + b + \epsilon$$

So let’s look at basic correlation relationship

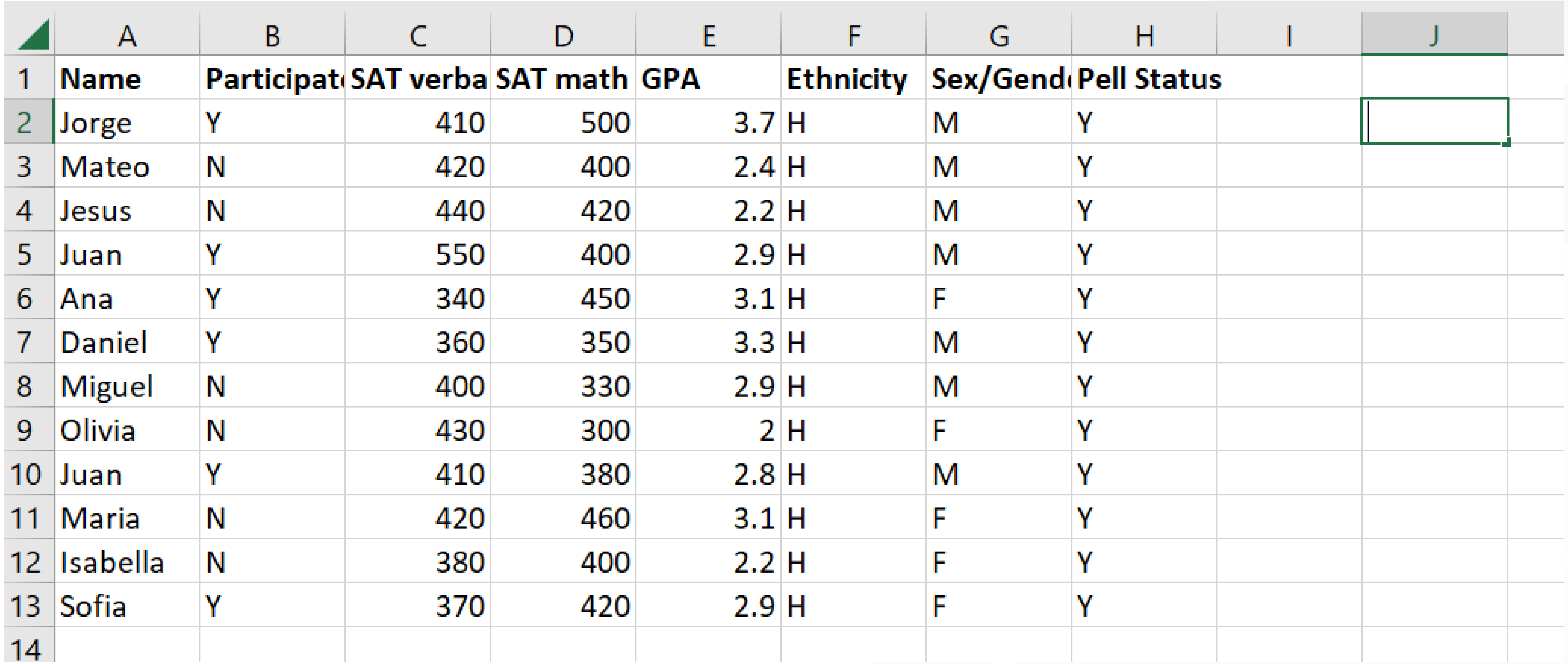


The difference in slopes is the effect of the program

So what types of controls to add?

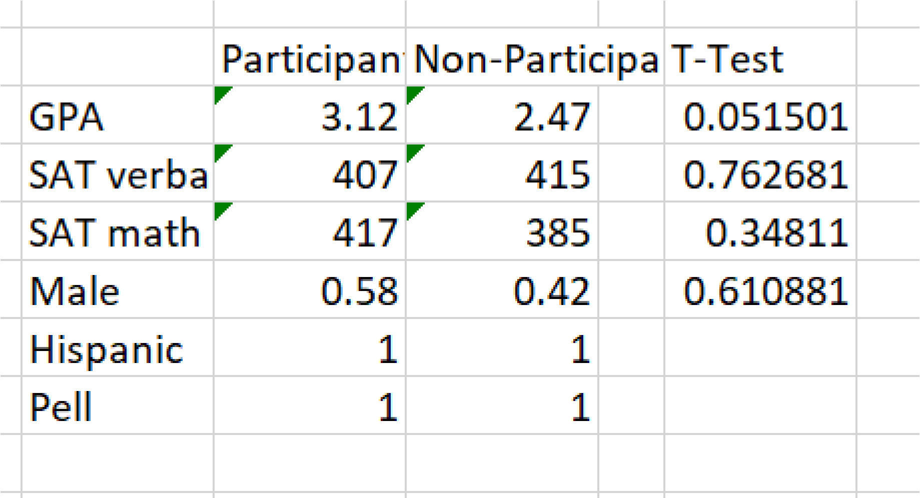
- Think about what types of things you can observe that you would expect to impact your outcome of interest.
- Some things to think about:
  - Prior achievement
  - Demographics (race, ethnicity*, language*, sex/gender)
  - Language status
  - Income
  - Learning disabilities

Example



	A	B	C	D	E	F	G	H	I	J
1	**Name**	**Participat**	**SAT verba**	**SAT math**	**GPA**	**Ethnicity**	**Sex/Gend**	**Pell Status**		
2	Jorge	Y	410	500	3.7	H	M	Y		
3	Mateo	N	420	400	2.4	H	M	Y		
4	Jesus	N	440	420	2.2	H	M	Y		
5	Juan	Y	550	400	2.9	H	M	Y		
6	Ana	Y	340	450	3.1	H	F	Y		
7	Daniel	Y	360	350	3.3	H	M	Y		
8	Miguel	N	400	330	2.9	H	M	Y		
9	Olivia	N	430	300	2	H	F	Y		
10	Juan	Y	410	380	2.8	H	M	Y		
11	Maria	N	420	460	3.1	H	F	Y		
12	Isabella	N	380	400	2.2	H	F	Y		
13	Sofia	Y	370	420	2.9	H	F	Y		
14										

Check for Baseline Equivalence

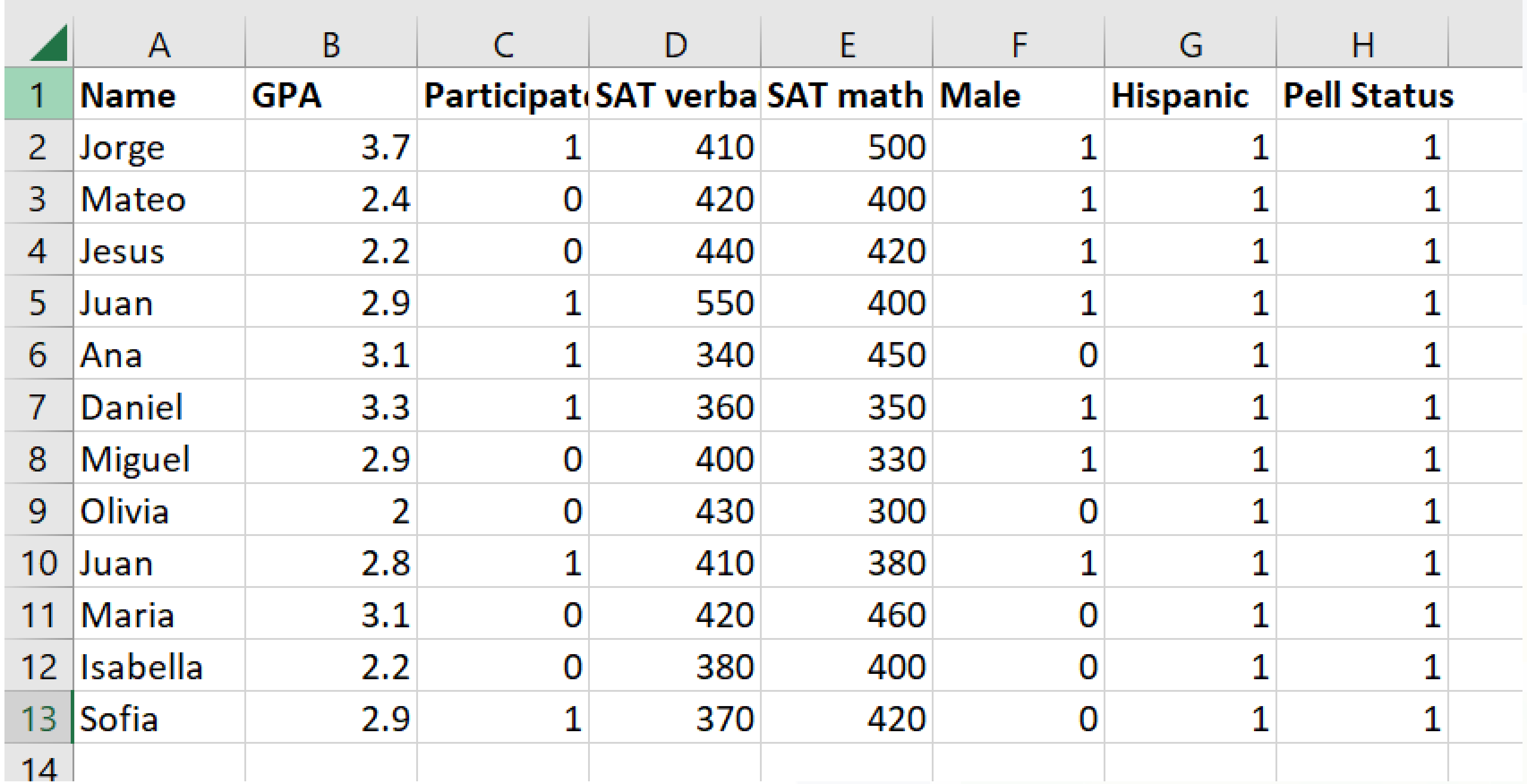


	Participan	Non-Participa		T-Test
GPA	3.12	2.47		0.051501
SAT verba	407	415		0.762681
SAT math	417	385		0.34811
Male	0.58	0.42		0.610881
Hispanic	1	1		
Pell	1	1		

=AVERAGE(D8:D13)

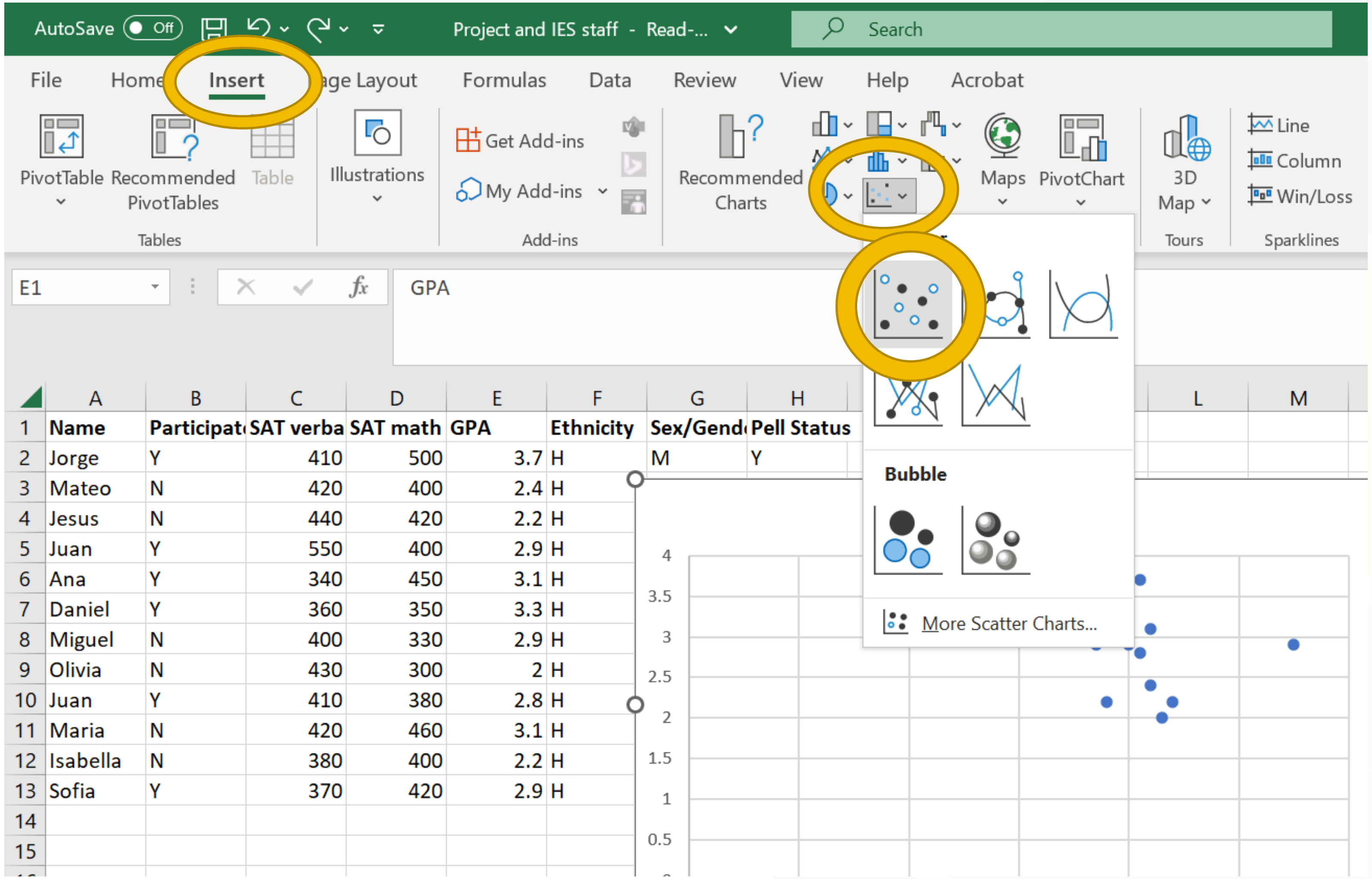
=TTEST(D2:D7,D8:D13,2,1)

Transform your data



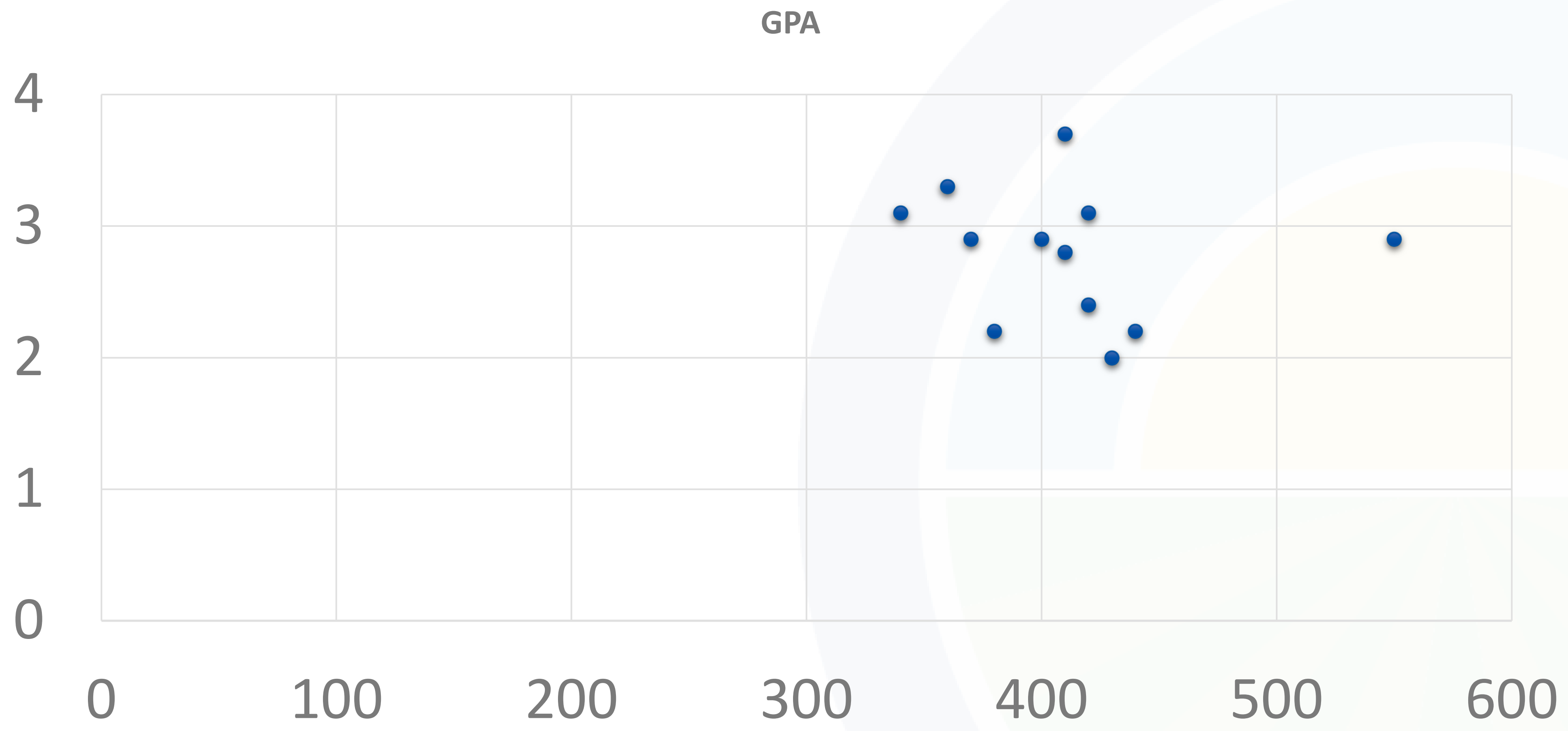
	A	B	C	D	E	F	G	H
1	Name	GPA	Participat	SAT verba	SAT math	Male	Hispanic	Pell Status
2	Jorge	3.7	1	410	500	1	1	1
3	Mateo	2.4	0	420	400	1	1	1
4	Jesus	2.2	0	440	420	1	1	1
5	Juan	2.9	1	550	400	1	1	1
6	Ana	3.1	1	340	450	0	1	1
7	Daniel	3.3	1	360	350	1	1	1
8	Miguel	2.9	0	400	330	1	1	1
9	Olivia	2	0	430	300	0	1	1
10	Juan	2.8	1	410	380	1	1	1
11	Maria	3.1	0	420	460	0	1	1
12	Isabella	2.2	0	380	400	0	1	1
13	Sofia	2.9	1	370	420	0	1	1
14								

First, look at the data to see if there is a relationship

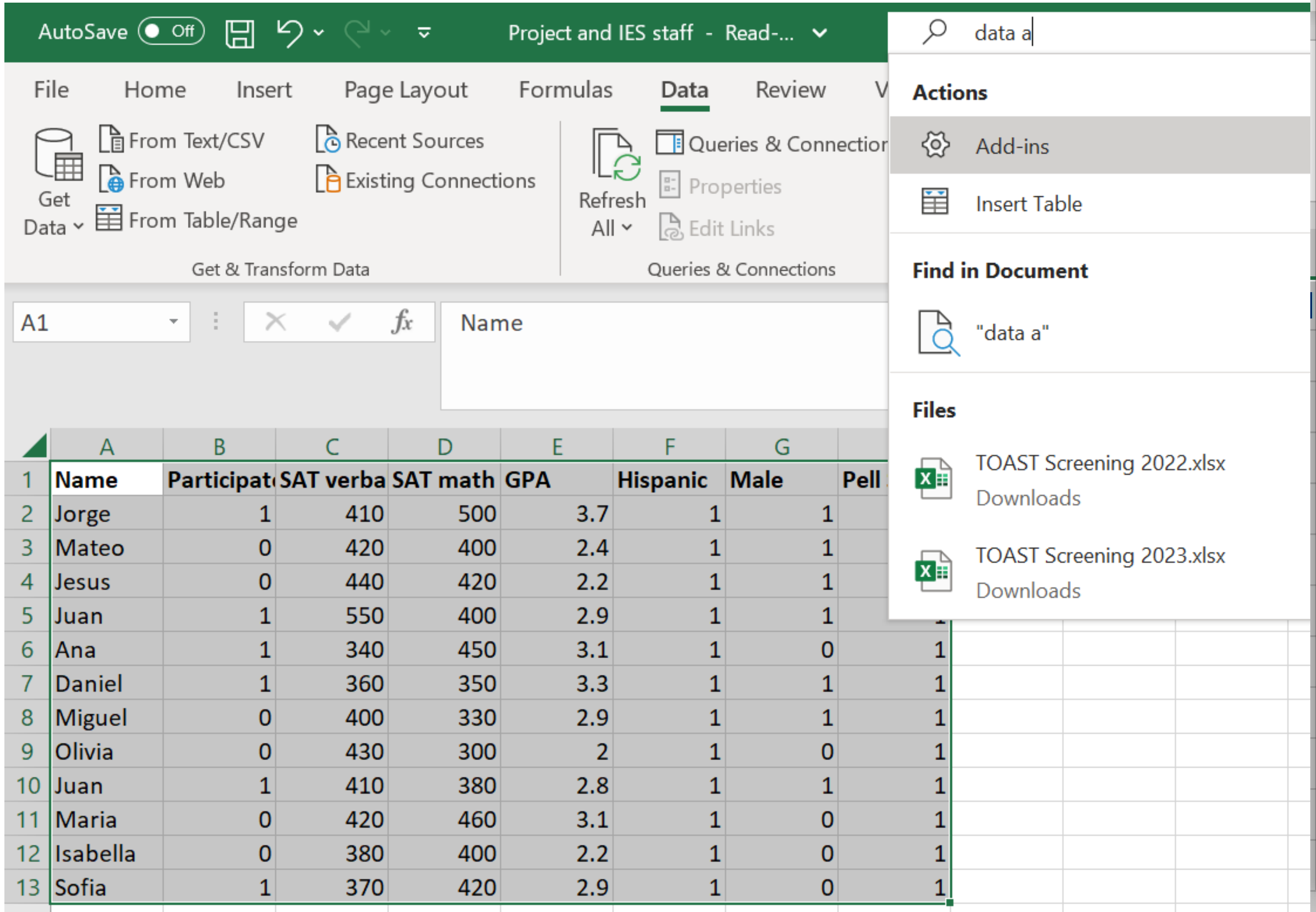


	A	B	C	D	E	F	G	H
1	Name	Participat	SAT verba	SAT math	GPA	Ethnicity	Sex/Gend	Pell Status
2	Jorge	Y	410	500	3.7	H	M	Y
3	Mateo	N	420	400	2.4	H		
4	Jesus	N	440	420	2.2	H		
5	Juan	Y	550	400	2.9	H		
6	Ana	Y	340	450	3.1	H		
7	Daniel	Y	360	350	3.3	H		
8	Miguel	N	400	330	2.9	H		
9	Olivia	N	430	300	2	H		
10	Juan	Y	410	380	2.8	H		
11	Maria	N	420	460	3.1	H		
12	Isabella	N	380	400	2.2	H		
13	Sofia	Y	370	420	2.9	H		
14								
15								

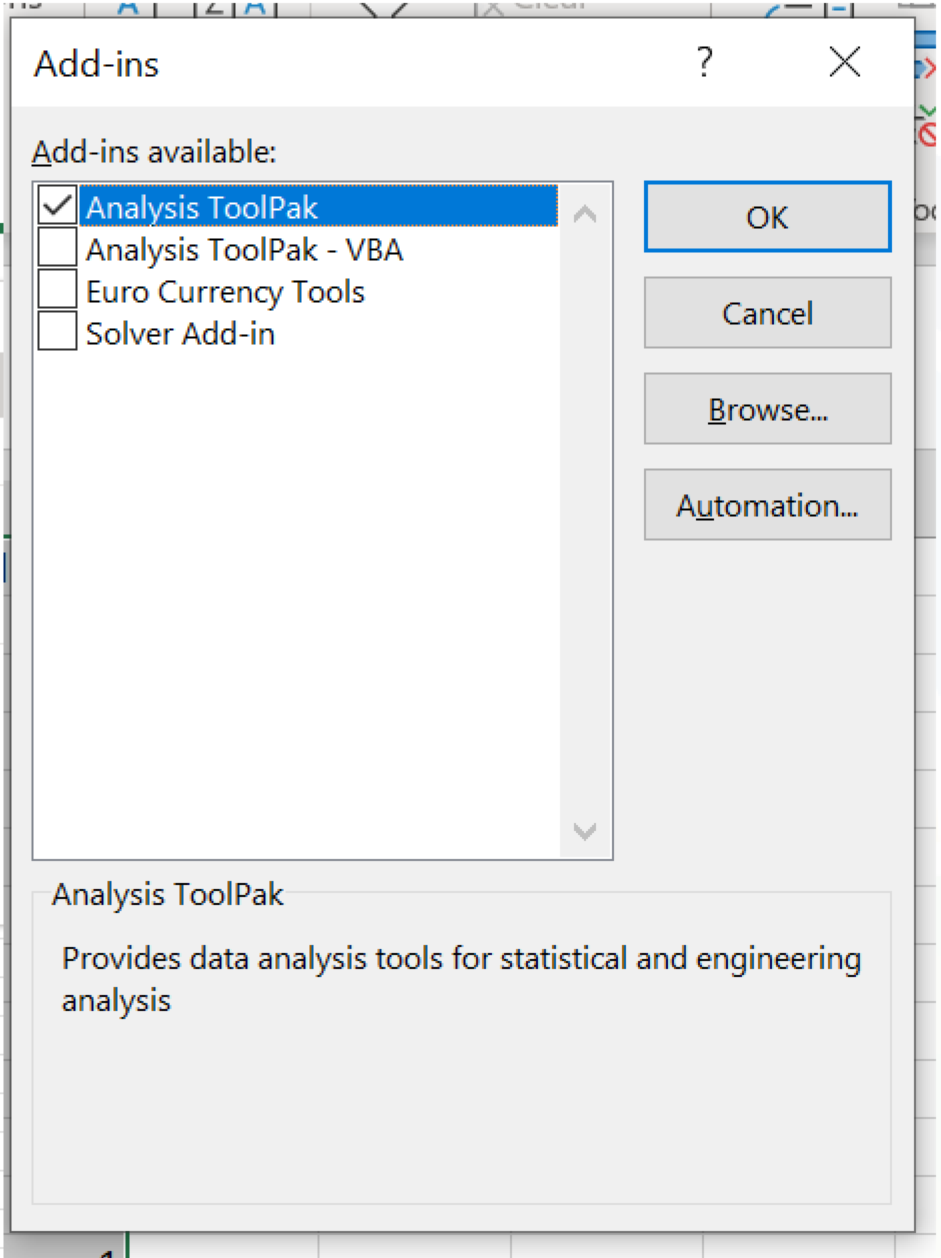
Make a graph to confirm it looks right



Get the Add-in



	A	B	C	D	E	F	G	
1	Name	Participat	SAT verba	SAT math	GPA	Hispanic	Male	Pell
2	Jorge	1	410	500	3.7	1	1	
3	Mateo	0	420	400	2.4	1	1	
4	Jesus	0	440	420	2.2	1	1	
5	Juan	1	550	400	2.9	1	1	1
6	Ana	1	340	450	3.1	1	0	1
7	Daniel	1	360	350	3.3	1	1	1
8	Miguel	0	400	330	2.9	1	1	1
9	Olivia	0	430	300	2	1	0	1
10	Juan	1	410	380	2.8	1	1	1
11	Maria	0	420	460	3.1	1	0	1
12	Isabella	0	380	400	2.2	1	0	1
13	Sofia	1	370	420	2.9	1	0	1



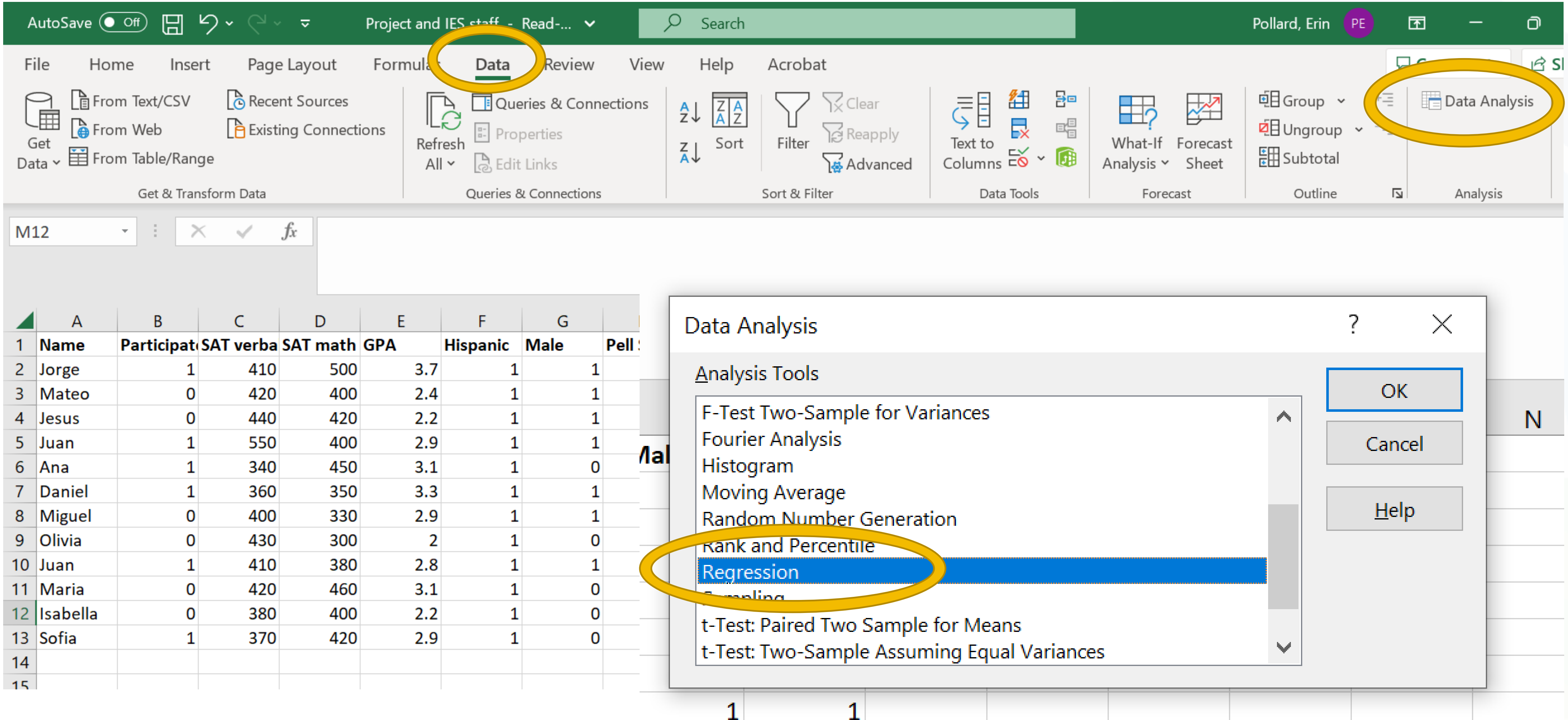
Add-ins available:

- ☑ Analysis ToolPak
- □ Analysis ToolPak - VBA
- □ Euro Currency Tools
- □ Solver Add-in

Analysis ToolPak

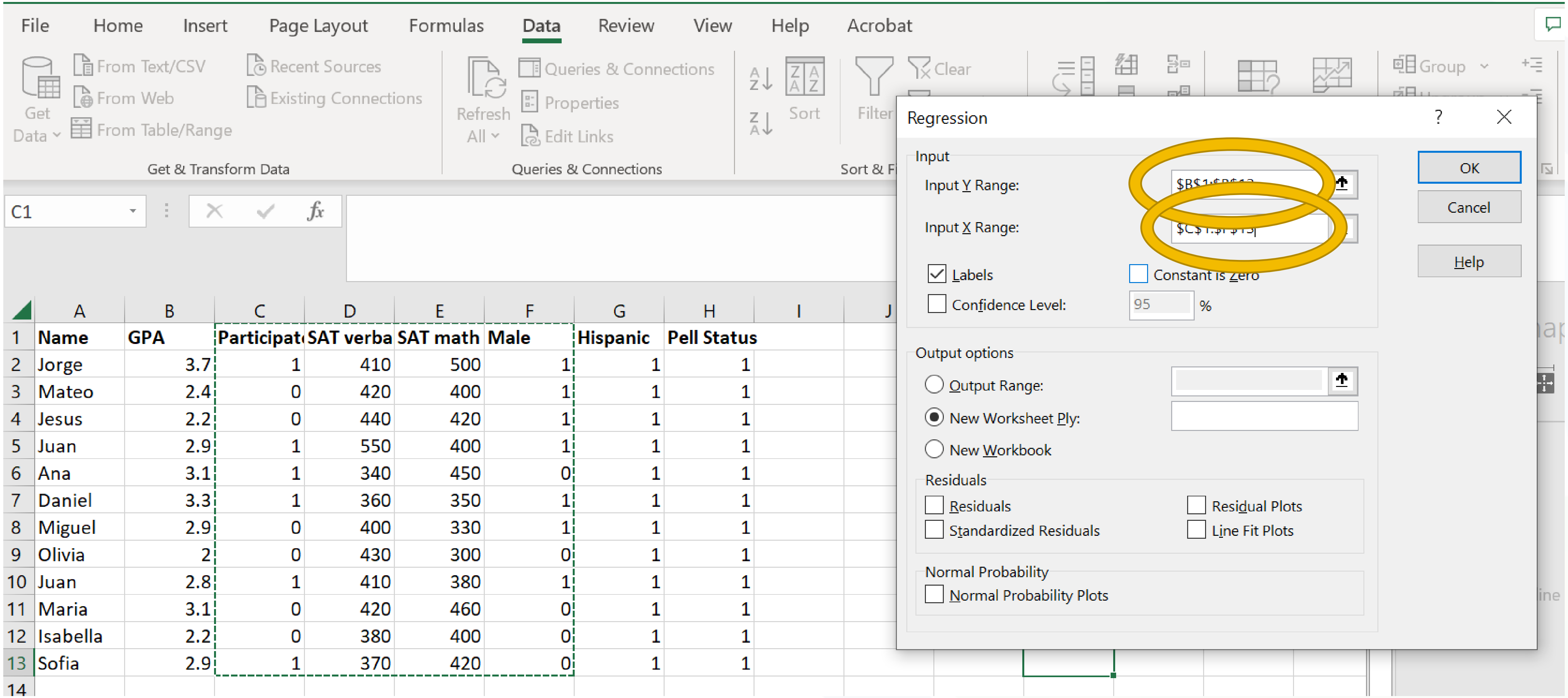
Provides data analysis tools for statistical and engineering analysis

Click on data analysis



	A	B	C	D	E	F	G	
1	Name	Participat	SAT verba	SAT math	GPA	Hispanic	Male	Pell
2	Jorge	1	410	500	3.7	1	1	
3	Mateo	0	420	400	2.4	1	1	
4	Jesus	0	440	420	2.2	1	1	
5	Juan	1	550	400	2.9	1	1	
6	Ana	1	340	450	3.1	1	0	
7	Daniel	1	360	350	3.3	1	1	
8	Miguel	0	400	330	2.9	1	1	
9	Olivia	0	430	300	2	1	0	
10	Juan	1	410	380	2.8	1	1	
11	Maria	0	420	460	3.1	1	0	
12	Isabella	0	380	400	2.2	1	0	
13	Sofia	1	370	420	2.9	1	0	
14								
15								

Create your model



	A	B	C	D	E	F	G	H
1	Name	GPA	Participat	SAT verba	SAT math	Male	Hispanic	Pell Status
2	Jorge	3.7	1	410	500	1	1	1
3	Mateo	2.4	0	420	400	1	1	1
4	Jesus	2.2	0	440	420	1	1	1
5	Juan	2.9	1	550	400	1	1	1
6	Ana	3.1	1	340	450	0	1	1
7	Daniel	3.3	1	360	350	1	1	1
8	Miguel	2.9	0	400	330	1	1	1
9	Olivia	2	0	430	300	0	1	1
10	Juan	2.8	1	410	380	1	1	1
11	Maria	3.1	0	420	460	0	1	1
12	Isabella	2.2	0	380	400	0	1	1
13	Sofia	2.9	1	370	420	0	1	1
14								

Regression

Input

Input Y Range: $B$1:$B$13

Input X Range: $C$1:$F$13

☑ Labels   □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals   □ Residual Plots

□ Standardized Residuals   □ Line Fit Plots

Normal Probability

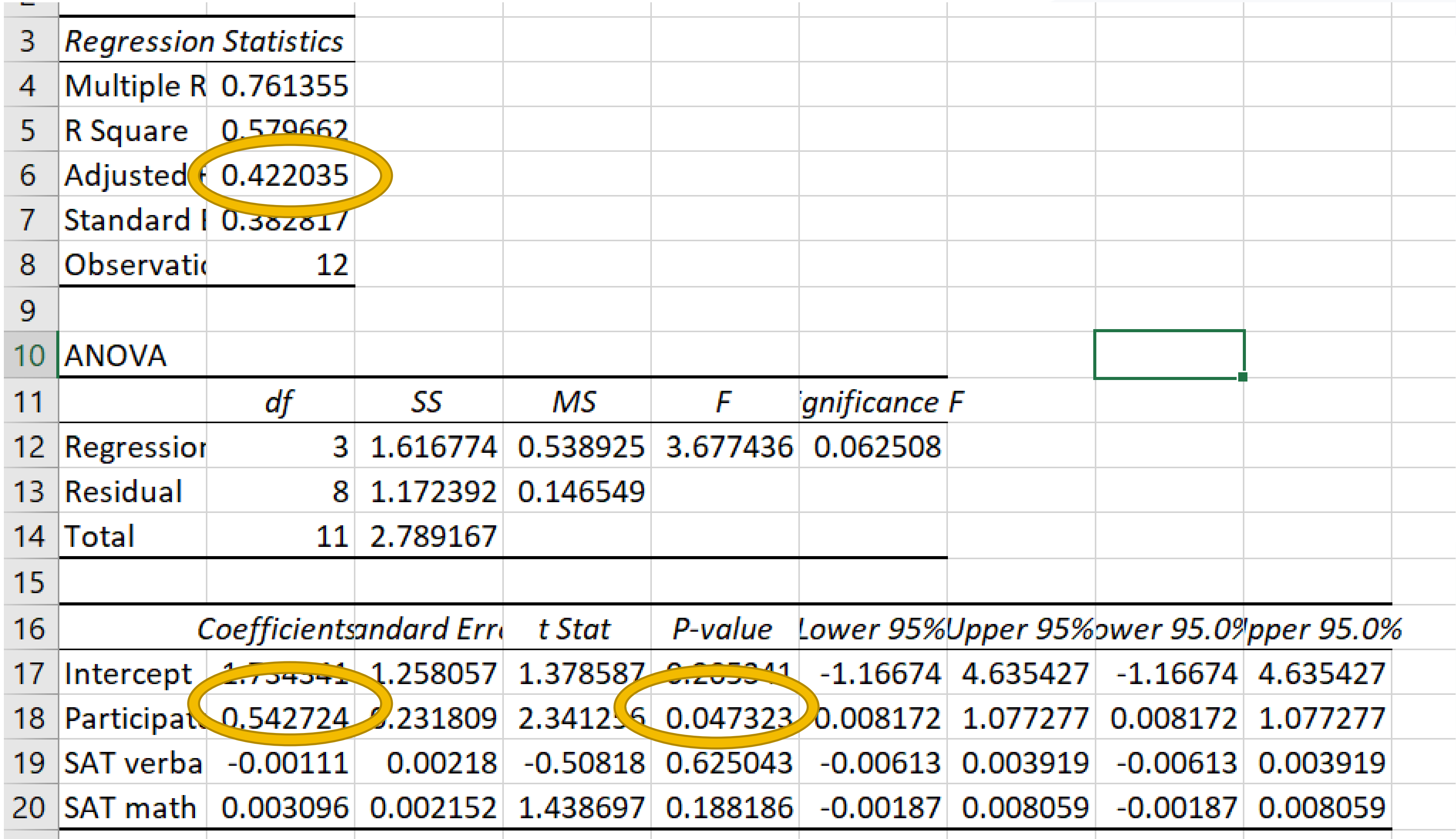
□ Normal Probability Plots

OK

Cancel

Help

Look at the results



3	*Regression Statistics*									
4	Multiple R	0.761355								
5	R Square	0.579662								
6	Adjusted	0.422035								
7	Standard	0.382817								
8	Observati	12								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*ignificance F*				
12	Regressior	3	1.616774	0.538925	3.677436	0.062508				
13	Residual	8	1.172392	0.146549						
14	Total	11	2.789167							
15										
16		*Coefficients*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*	
17	Intercept	1.734341	1.258057	1.378587	0.205341	-1.16674	4.635427	-1.16674	4.635427	
18	Participat	0.542724	0.231809	2.341256	0.047323	0.008172	1.077277	0.008172	1.077277	
19	SAT verba	-0.00111	0.00218	-0.50818	0.625043	-0.00613	0.003919	-0.00613	0.003919	
20	SAT math	0.003096	0.002152	1.438697	0.188186	-0.00187	0.008059	-0.00187	0.008059	

Alternate ways to do a regression

Now that I’ve done the stats, what do I do?

A program evaluation will tell you:

- What is the program
- How was it implemented
- Costs and resources necessary to implement the program
- Program participants
- **Did the program work?**
  - Did it work for everyone?
  - Are the effects the same for everyone?

This contextual data is often the most important part of the evaluation. Documenting the program is essential.

What do I want to know?

- What is the program?
- What do you do?
- How does the program work?
- Who did it serve?
- How did you study it?
- Who did you study?
- What did your evaluation find?
- Was it what you expected?

Length and details

- The goal is to be long enough to document everything people want to know, without adding unnecessary fluff.
- Some are 10 pages, others are over 300
- Add any surveys or extras in an appendix

What about sharing my findings?

What is required of you?

- You must:
  - Conduct an evaluation
  - Share the evaluation with your program officer
- You should:
  - Share your evaluation with your colleagues
  - Share your evaluation with the world at eric.ed.gov/?submit

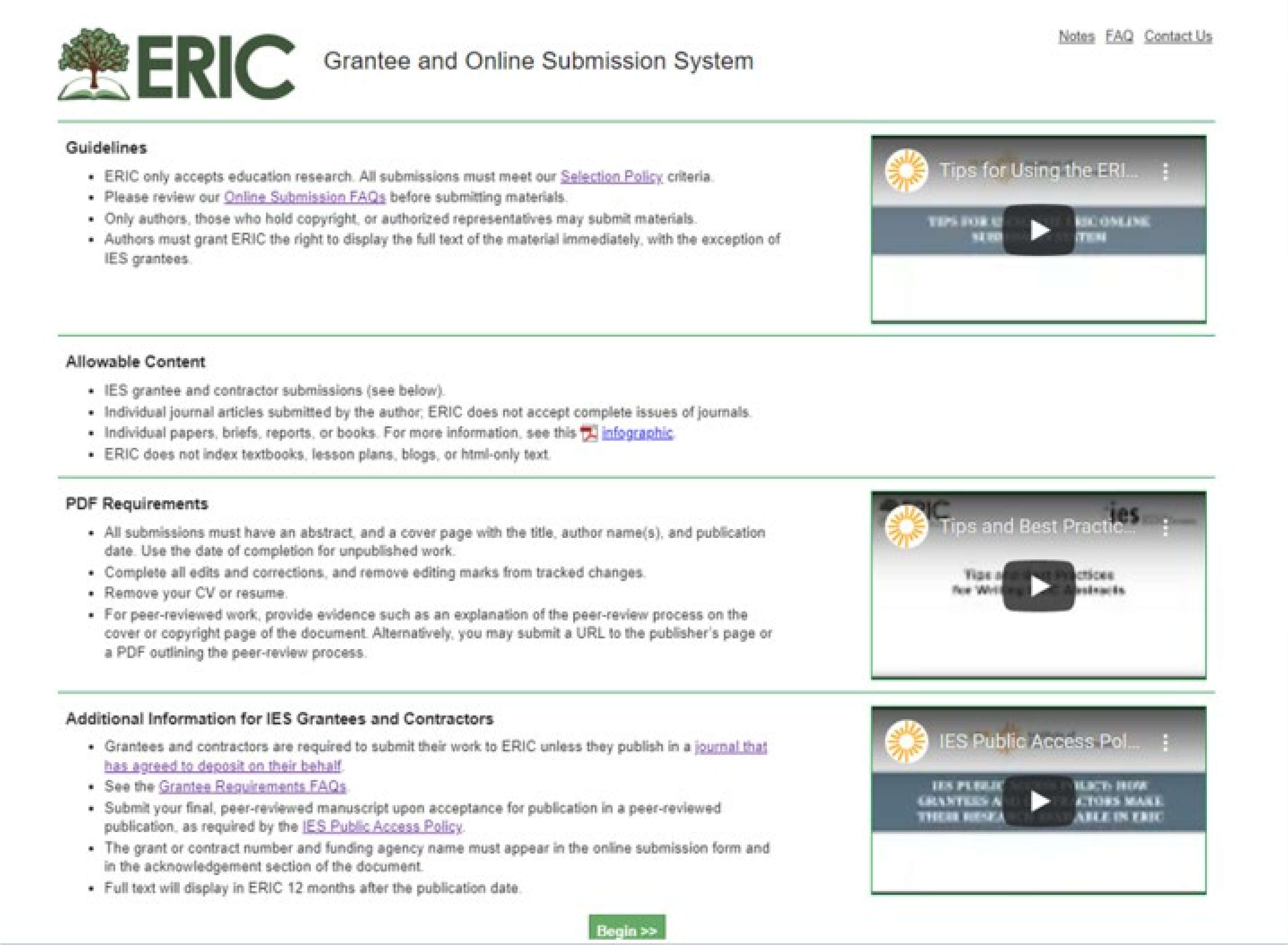
Share with your peers

- This is your network– you want to share what you know with them
- Learn from one another!
- Teach us– what worked? What didn’t?

Why share with the world?

- This is how we learn from each other
- It documents what was invested by the government to push for future funding
- People often worry that their work is not good enough to share, but what I hear from the field is that these are our most valuable products. People want to learn from you!

Share with the world



ERIC Grantee and Online Submission System

Guidelines

- ERIC only accepts education research. All submissions must meet our Selection Policy criteria.
- Please review our Online Submission FAQs before submitting materials.
- Only authors, those who hold copyright, or authorized representatives may submit materials.
- Authors must grant ERIC the right to display the full text of the material immediately, with the exception of IES grantees.

Allowable Content

- IES grantee and contractor submissions (see below).
- Individual journal articles submitted by the author; ERIC does not accept complete issues of journals.
- Individual papers, briefs, reports, or books. For more information, see this infographic.
- ERIC does not index textbooks, lesson plans, blogs, or html-only text.

PDF Requirements

- All submissions must have an abstract, and a cover page with the title, author name(s), and publication date. Use the date of completion for unpublished work.
- Complete all edits and corrections, and remove editing marks from tracked changes.
- Remove your CV or resume.
- For peer-reviewed work, provide evidence such as an explanation of the peer-review process on the cover or copyright page of the document. Alternatively, you may submit a URL to the publisher's page or a PDF outlining the peer-review process.

Additional Information for IES Grantees and Contractors

- Grantees and contractors are required to submit their work to ERIC unless they publish in a journal that has agreed to deposit on their behalf.
- See the Grantee Requirements FAQs.
- Submit your final, peer-reviewed manuscript upon acceptance for publication in a peer-reviewed publication, as required by the IES Public Access Policy.
- The grant or contract number and funding agency name must appear in the online submission form and in the acknowledgement section of the document.
- Full text will display in ERIC 12 months after the publication date.

To recap:

- Document and share your evaluations– we want to learn from each other!
- Follow what your evaluation plans
- Going forward, think about how to add rigor

Questions?

Thank you!

- Erin Pollard, What Works Clearinghouse and ERIC, Institute of Education Sciences: Erin.Pollard@ed.gov
- Katrina Ballard, HEP and CAMP Data and Evaluation SME, Office of Migrant Education: Katrina.Ballard@ed.gov