Concerns about (E)DM: why to look for *strong, causal* relationships

Joseph E. Beck, Doug Selent and Dovan Rai

Main point

- With large data sets, statistical hypothesis testing breaks down
  - Finds far more relationships than we care about (or that even exist!)
  - Can find relationships that are meaningless
- Better approaches
  - Tetrad for relationship mining
  - Using cutoff of absolute magnitude of effect rather than P-values

Let’s imagine a study

- We collect data on 20 variables about students in our study (# of columns)

Collect data on these 20 variables

- gender
  learn_rate
  prior_knowledge
  grit
  like_subject
  num_problems_solved
  correct
  time_in_tutor
  gaming
  off_task
- prior_exam_score
  tutor_version
  teacher_quality
  age
  amount_of_homework
  homework_rate
  num_hints
  parental_involvement
  SES
  pre_post_gain

Let’s see what this looks like

- Each row is the data obtained from one student
  - Typically each student contributes to more than one row, but keeping things straightforward

- (SPSS)

Run a study

- Collect data on 100 students

- Run a correlation analysis to find related variables
  - Correlation tests to see if there is a linear relationship between two variables

Correlations

- Statistical test between 2 variables
- Ranges frm -1 to 1
  - 1 perfect positive relationship
- Height / weight correlation at about 0.6
- Height / IQ correlation about 0.2
- Joe's rule of thumb: ignore (-0.2, 0.2)

Quick demo

- (SPSS)

- Show
  - Correlation table
  - Statistical significance (*, **)
    - Smaller→more certainty
  - Scatterplots

Results for 100 student study

- Find 66 relationships with $P<0.01$
  - Statistically powerful relationship

- Thoughts?

Results for 100 student study

- Find 66 relationships with $P<0.01$

- Thoughts?
  - More results than I want to write about
  - Or read

Tell our grad students to run a bigger study

- Collect data on 1000 students

N	# relationships P<0.01
100	66
1000	84

- Thoughts?

Hire some additional assistants

- Collect data on **10,000** students

N	# relationships P<0.01
100	66
1000	84
10,000	94

- Yep, more data lets us find more relationships

Really make assistants work…

- Collect data on **100,000** students

N	# relationships P<0.01
100	66
1000	84
10,000	94
100,000	103

- Thoughts?

Really make assistants work…

- Collect data on **100,000** students

N	# relationships P<0.01
100	66
1000	84
10,000	94
100,000	103

- What is the relationship between the amount of data and our ability to understand how the world works?

Really make assistants work…

- Collect data on **100,000** students

N	# relationships P<0.01
100	66
1000	84
10,000	94
100,000	103

- What if I told you there were only **29 actual** relationships in the data?

How could I know how many relationships?

- Generated the data synthetically
- Made up plausible model of how the world behaves
  - Was not thinking of pedagogical purposes or creating nightmare scenarios for statistics
- Let's take a look at it
  - (tetrad)

Seems to be a mismatch

- There are 29 relationships in the model (I counted)
  - But SPSS found from 66 to 103 relationships
- Why is SPSS finding so many more relationships?
  - 3 type of reasons

Reason 1: type I error

- Type I error: imaging there is a relationship there even when there isn't one due to random error
  - P<0.01 means a 1% chance of hallucinating a relationship

- 20 variables → $(20^2 - 20) / 2 = 180$ possible relationships
  - E(type I errors) = 180 * 0.01 = 1.8

Type I errors can matter

- Probably not in this case, since only 1.8 such errors
- But gets worse as C (# of columns) increases
  - 50 columns → 12.25 errors
  - 100 columns → 49.5 errors
- One disadvantage of aggregating information together

Reason 2: larger N → smaller cutoff for "significant" result

- Always remember what P-values mean
  - It is the probability the result is **nonzero**
  - Not big, not important, not meaningful

- More data provides more certainty that the result is not equal to 0

How data impact P-values

- Let's consider the relationship between amount of homework assigned and a student's *grit* (show in SPSS)

How data impact P-values

- Let's consider the relationship between amount of homework assigned and a student's *grit* (show in SPSS)

N	Correlation	P-value
100	-0.16	0.11
1000	-0.19	0.0000000041
10,000	-0.13	$3.9 * 10^{-39}$
100,000	-0.13	≈0

Correlation strength fairly stable

N	Correlation	P-value
100	-0.16	0.11
1000	-0.19	0.0000000041
10,000	-0.13	$3.9 * 10^{-39}$
100,000	-0.13	≈0

Correlation p-value strongly affected by data size

- Has relationship gotten any more important by collecting more data?

N	Correlation	P-value
100	-0.16	0.11
1000	-0.19	0.0000000041
10,000	-0.13	$3.9 * 10^{-39}$
100,000	-0.13	≈0

More data → find more relationships

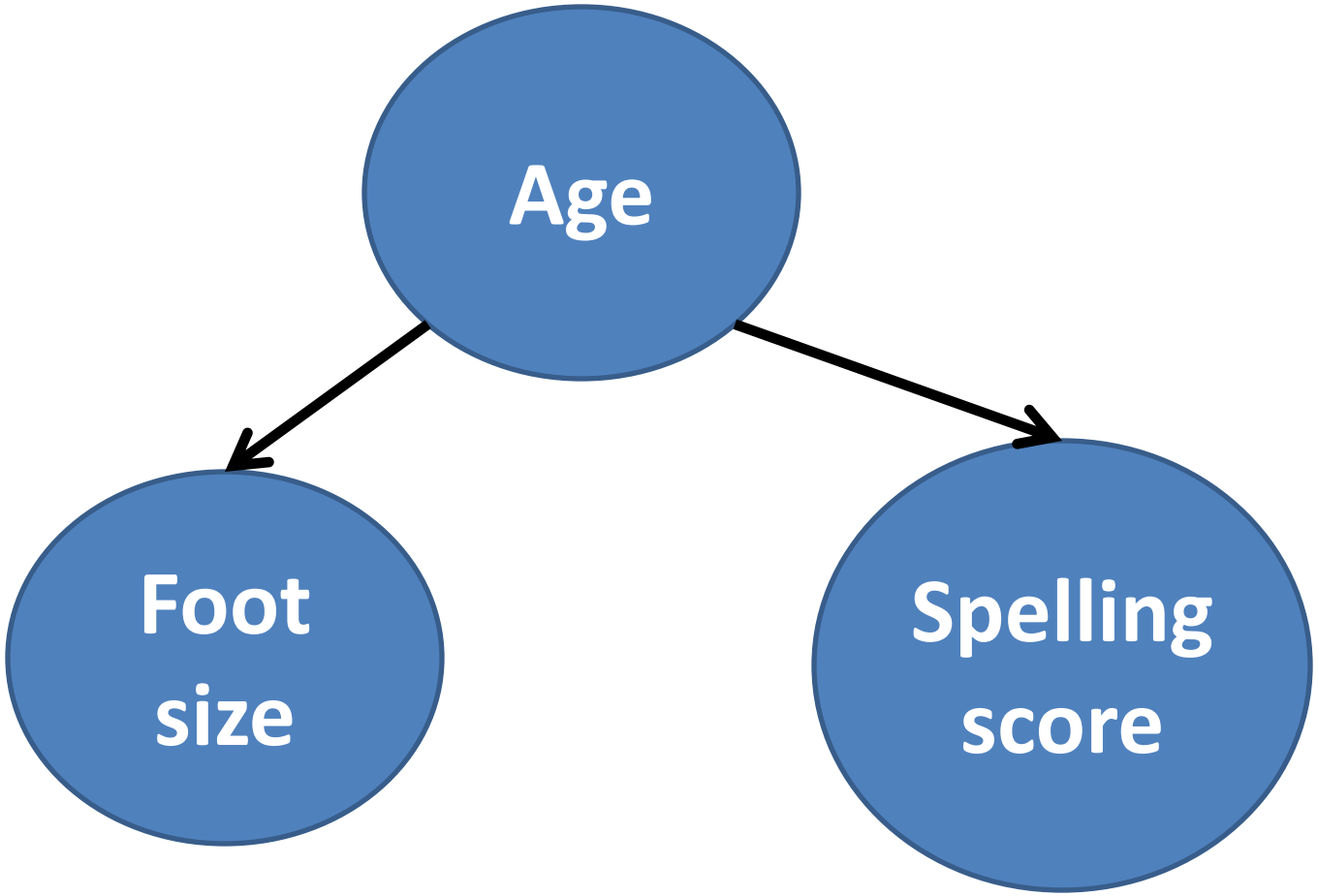
- Even if the strength of the relationship is marginal (height and IQ correlate at about 0.2)

N	Correlation	P-value
100	-0.16	0.11
1000	-0.19	0.0000000041
10,000	-0.13	$3.9 * 10^{-39}$
100,000	-0.13	≈0

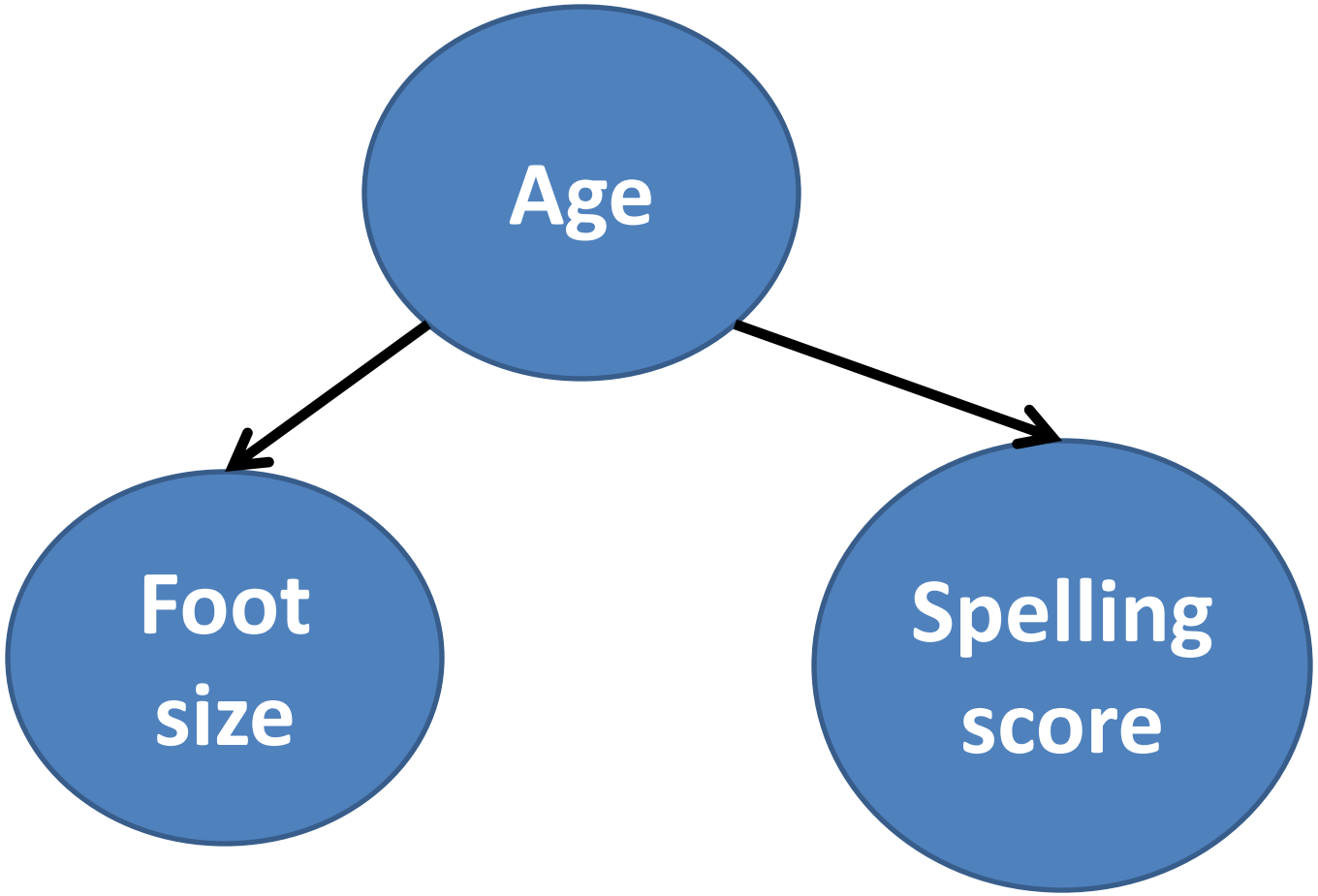
Reason 3: spurious relationships

- My favorite example: foot size and spelling ability are strongly correlated with each other for primary school students
  - Why?

Age is a common cause

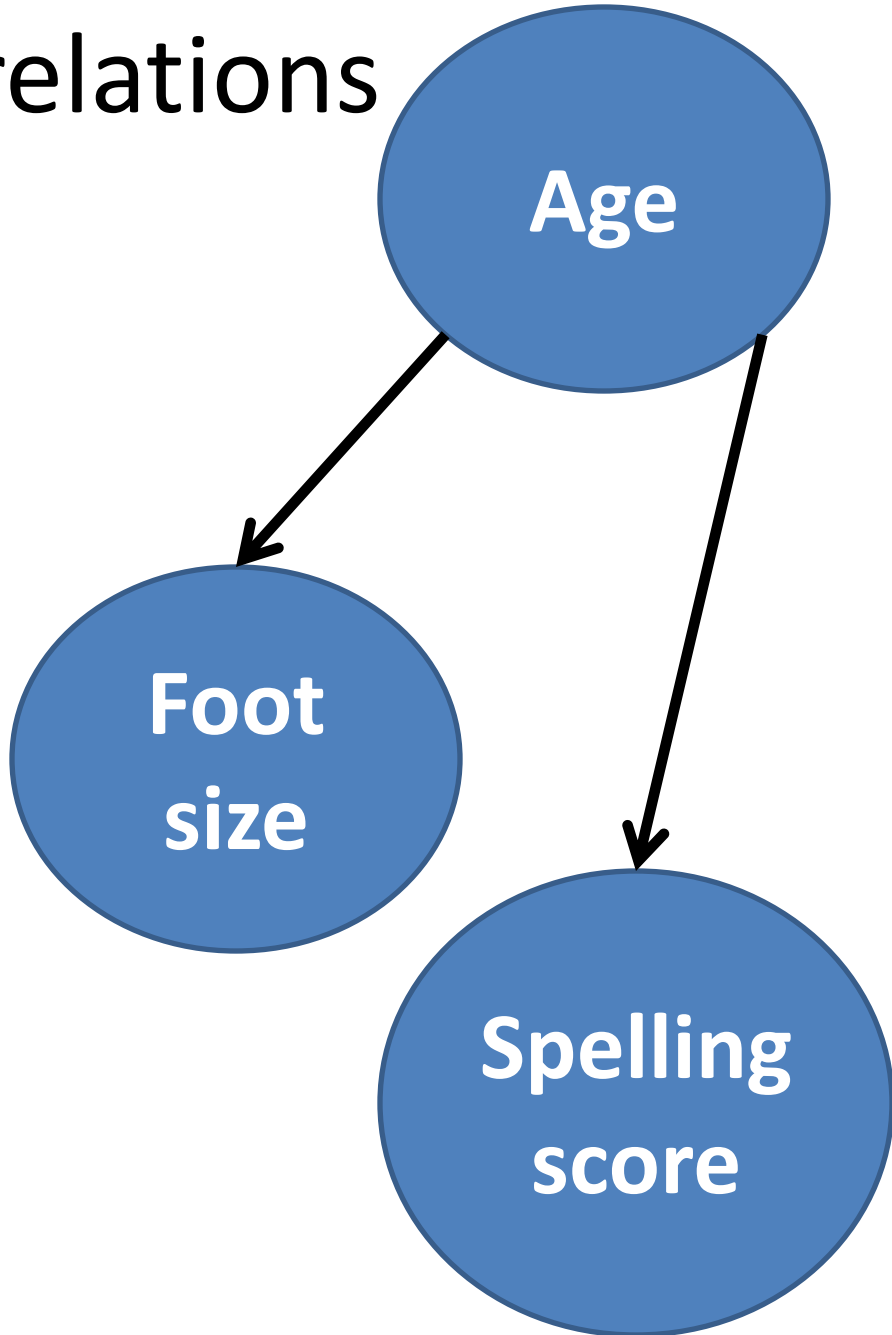


Do we care that foot size and spelling ability are correlated?

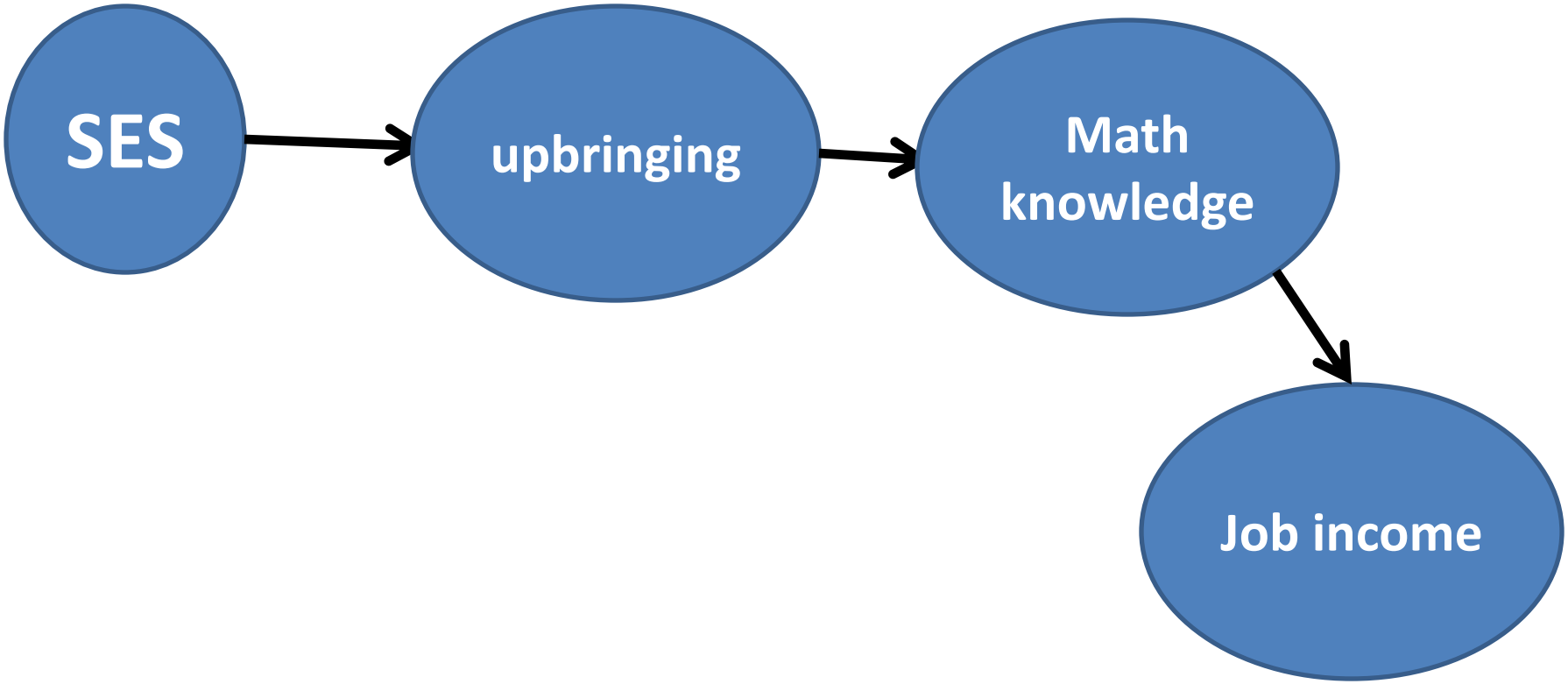


Partial correlations

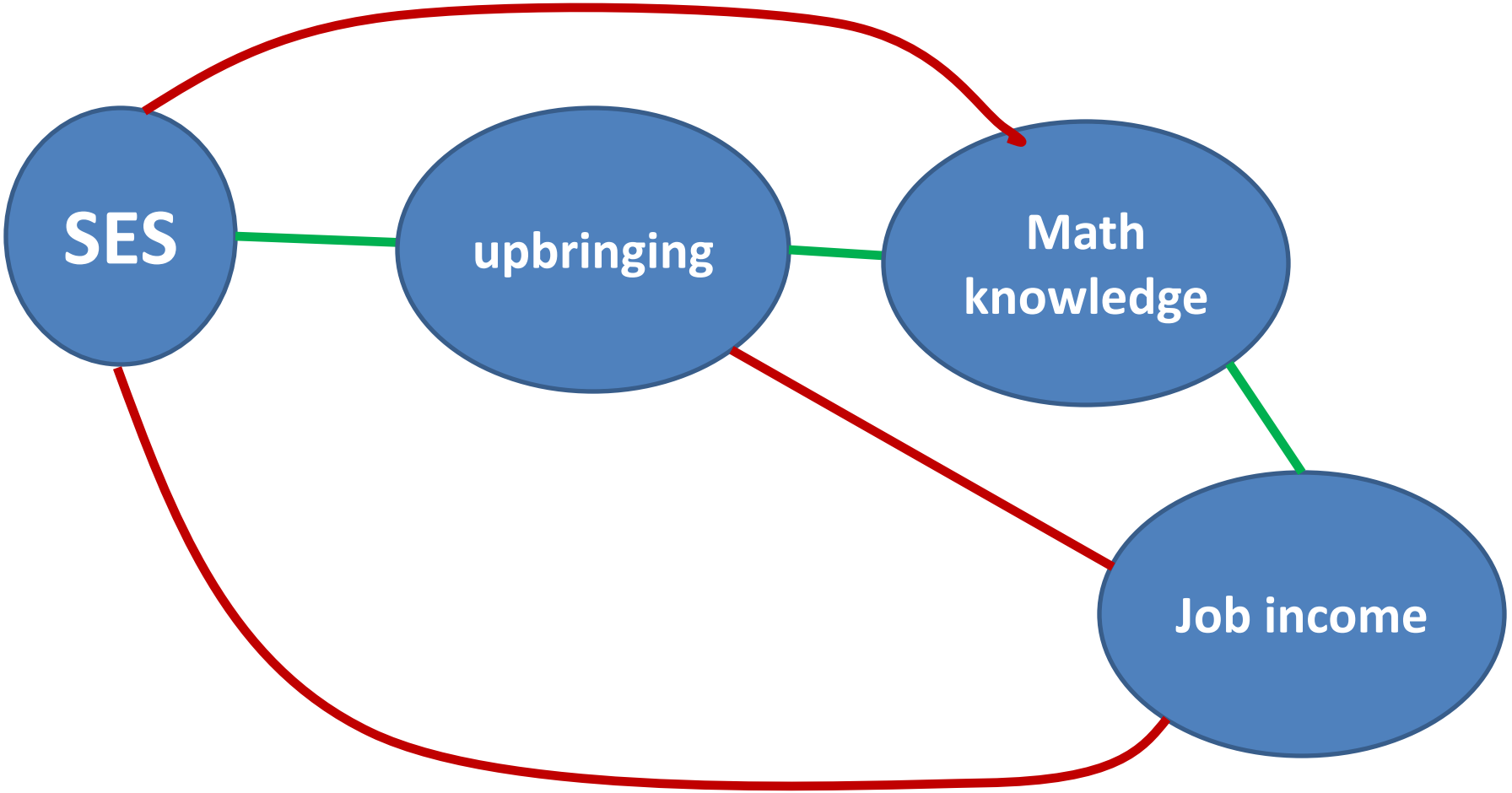
- Foot size and spelling are correlated
- Partial correlations control for impact of another variable and measure *direct relation*
- Partial correlation of foot size and spelling, partialing out age is ≈0



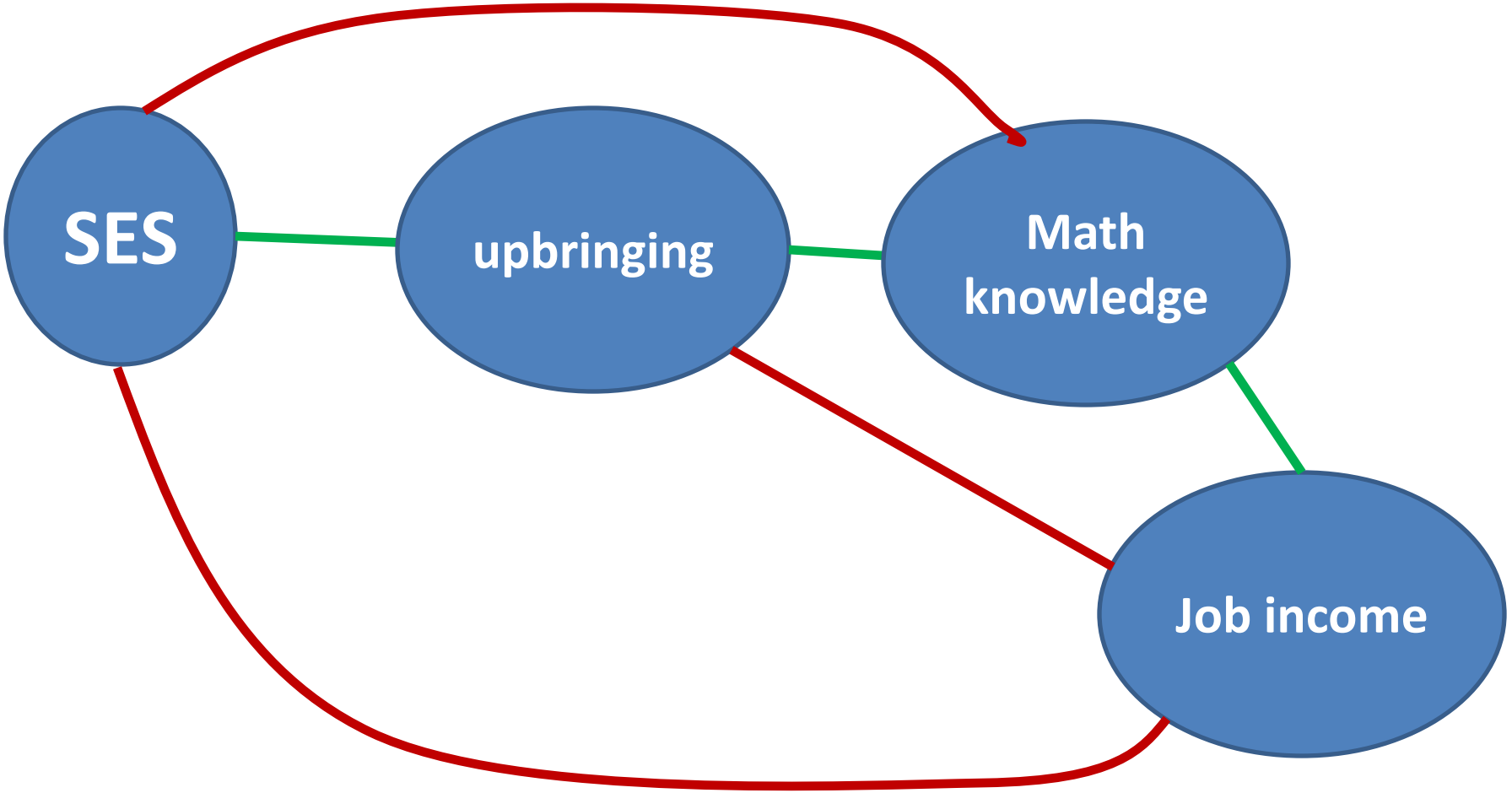
More generally



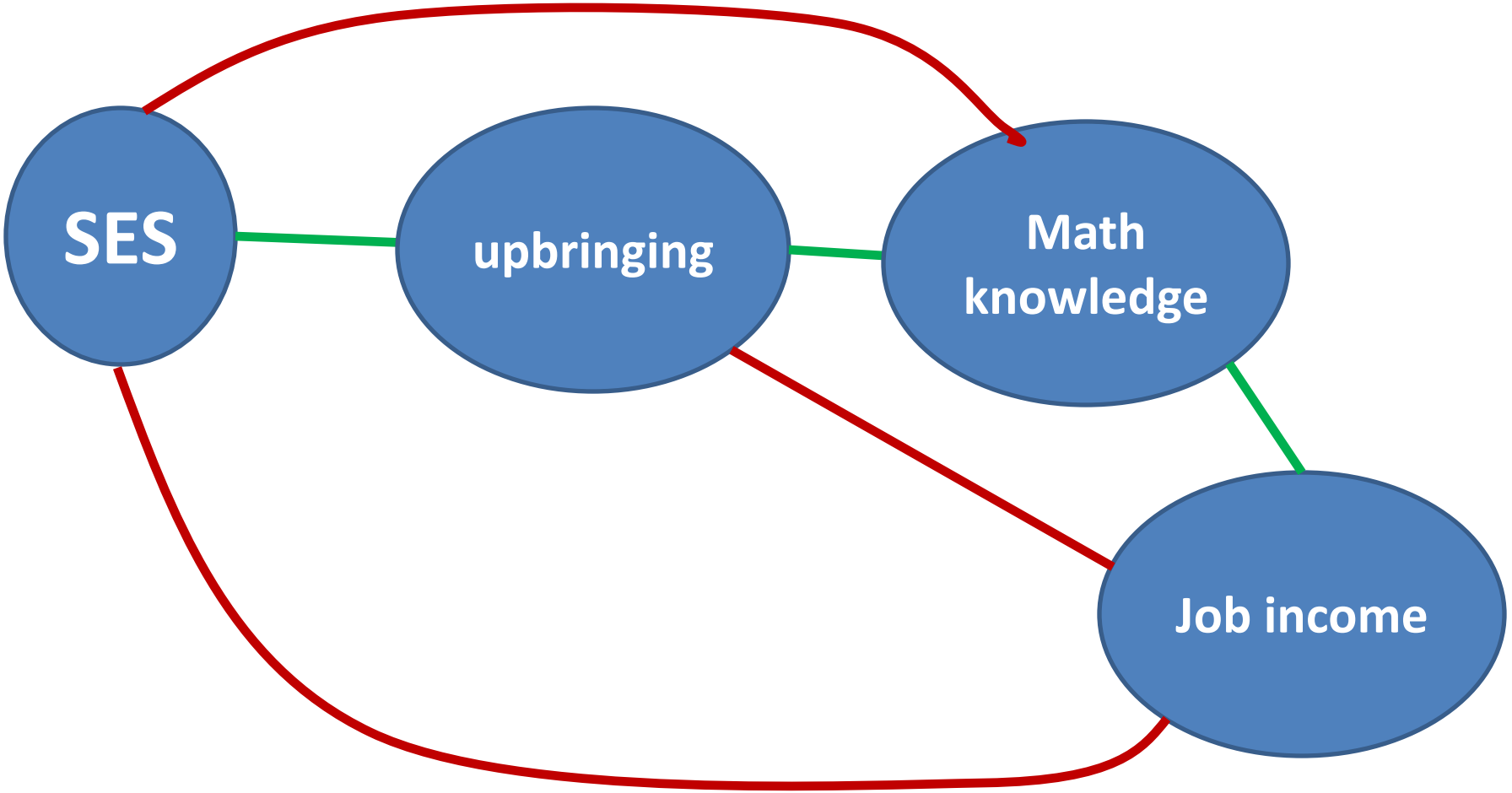
What statistical tests will find



Because those terms do correlate



Call these **spurious relationships**



An example from our data set

- SES (Socio Economic Status) and post test score correlate at 0.23
  - A ha! Wealthier students to better

- Is this relationship real, or like shoe size and spelling ability?
- What if we look at a third variable, amount of parental involvement in schooling?
  - Correlates at 0.49 with SES and 0.27 with test gain

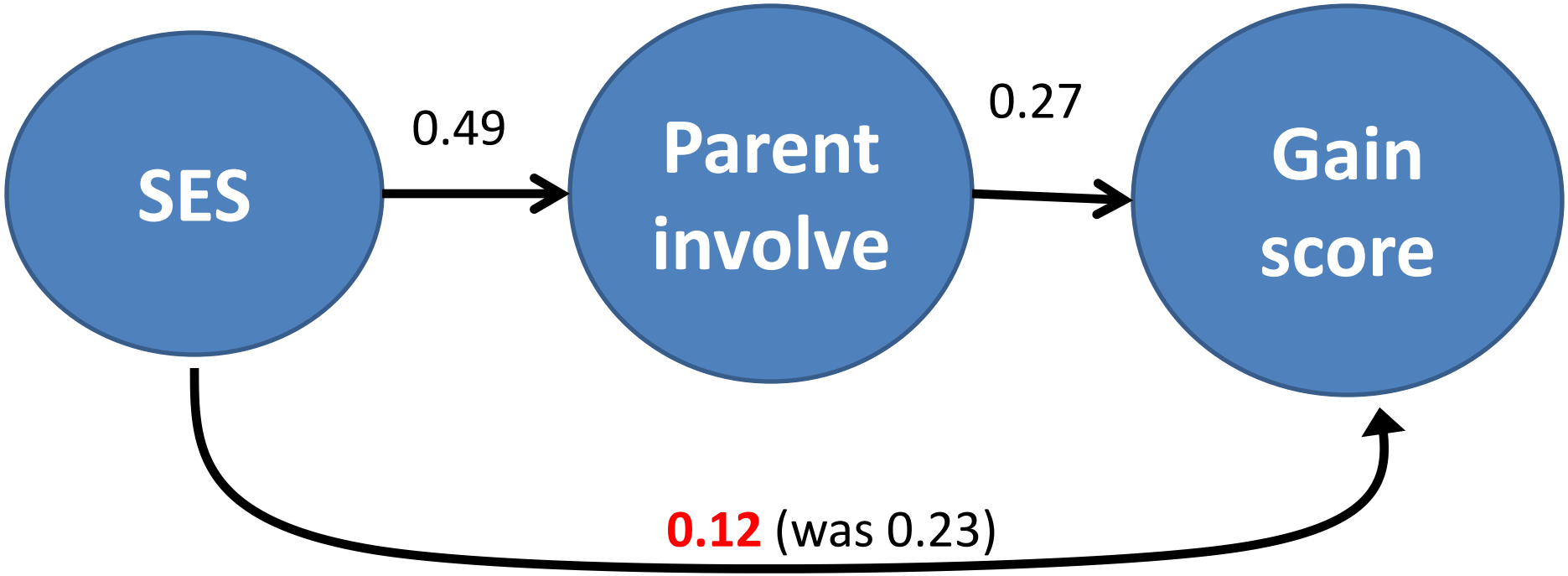
One possibility

- What if SES → parental involvement → test score gain

- Is there some way to test whether SES influences test score gain after accounting for parental involvement?
  - Yes. *Partial correlations*

What a **partial correlation** does

- Partial correlation SES, post test gain, partialing out parental involvement:



Neat!

- Partial correlations are a way to test *direct relationships*
  - i.e., A→B→C→D means A and D are associated, but not a direct relationship

- Partial correlation of shoe size, spelling ability, partialing out age ≈ 0

But…

- Irritating to keep running partial correlations
  - Lots of possible variables to consider for partialing
  - (SPSS)
  - Is one choice better than another?
- Weirdly, a partial correlation can cause a relationship to exist even when it doesn’t
- How to report it in a paper?

Wish list, a tool that

- Would test sets of partial correlations and discover direct relationships automatically

- Could display it in a easy to understand manner

Tetrad

- Designed as a causal modeling tool
  - Can sometimes infer causal relationships from observational data (really neat topic)
  - Let's forget about that aspect

- Constructs a graph, such that
  - A→B→C means: A influences B, B influences C, but there is no direct influence of A on C (even though A and C probably correlate with each other)

How well does it do…

- Used same simulated data as with SPSS
- Ran Tetrad *search* with it
  - Tries to recover a graph that represents relationships between variables
  - Returns a graph of the statistically reliable relationships *that are not spurious*

- (tetrad)

Number of relationships found (out of 29)

N	# relationships P<0.01	Tetrad
100	66	19
1000	84	25
10,000	94	27
100,000	103	28

Tetrad consistent number of relationships (except small data sets)

N	# relationships P<0.01	Tetrad
100	66	19
1000	84	25
10,000	94	27
100,000	103	28

Ability to zoom in

- We care about more than statistically reliable effects
  - Correlation of -0.008 is reliable with 100,000 data points – but who cares? (show in SPSS)
- Would like to focus on relationships with high magnitude
  - (tweaked tetrad demo)

Data from Wayang outpost

- We (well, Dovan Rai :-))tried to write an EDM conference paper on it
  - Was a mess
  - Very complex graph

- (tetrad demo)

Feedback

- Wayang folks were impressed :-)

- Extension to Tetrad developed by Doug Selent for a class project in my Graphical Models course

Why I care about this topic

- See a goal of science of discovering causal relationships about some domain
  - Do not care about incidental correlations (e.g. foot size and spelling scores)

- Which paper would you rather read?

Problem grows with bigger data sets

- More columns → many more effects to test and "discover"
  - Grows with $(C^2 – C) / 2$
- More rows → smaller and smaller effects can be detected
  - But doesn't make them any more meaningful!
- Concern about EDM being bogged down

More data → find more relationships

- Even if the strength of the relationship is marginal (height and IQ correlate at about 0.2)

N	Correlation	P-value
100	-0.16	0.11
1000	-0.19	0.0000000041
10,000	-0.13	$3.9 * 10^{-39}$
100,000	-0.13	≈0

Software

- Tetrad: available at http://www.phil.cmu.edu/projects/tetrad/
  - Free!
  - Google tetrad causal

- Modified Tetrad: email me (josephbeck@wpi.edu)
  - Experimental (not wrapped into main distribution)
  - Doug Selent's class project

software

- $P$$
- PSPP (freeware version of SPSS)

- SAS ($)
- R (freeware version of SAS)
  - Command line

Run a study

- Collect data on 100 students

- Run a correlation analysis to find related variables
  - Using correlation as common language, but many many ways to test a lot of relationships (e.g. ANOVA with interaction terms)