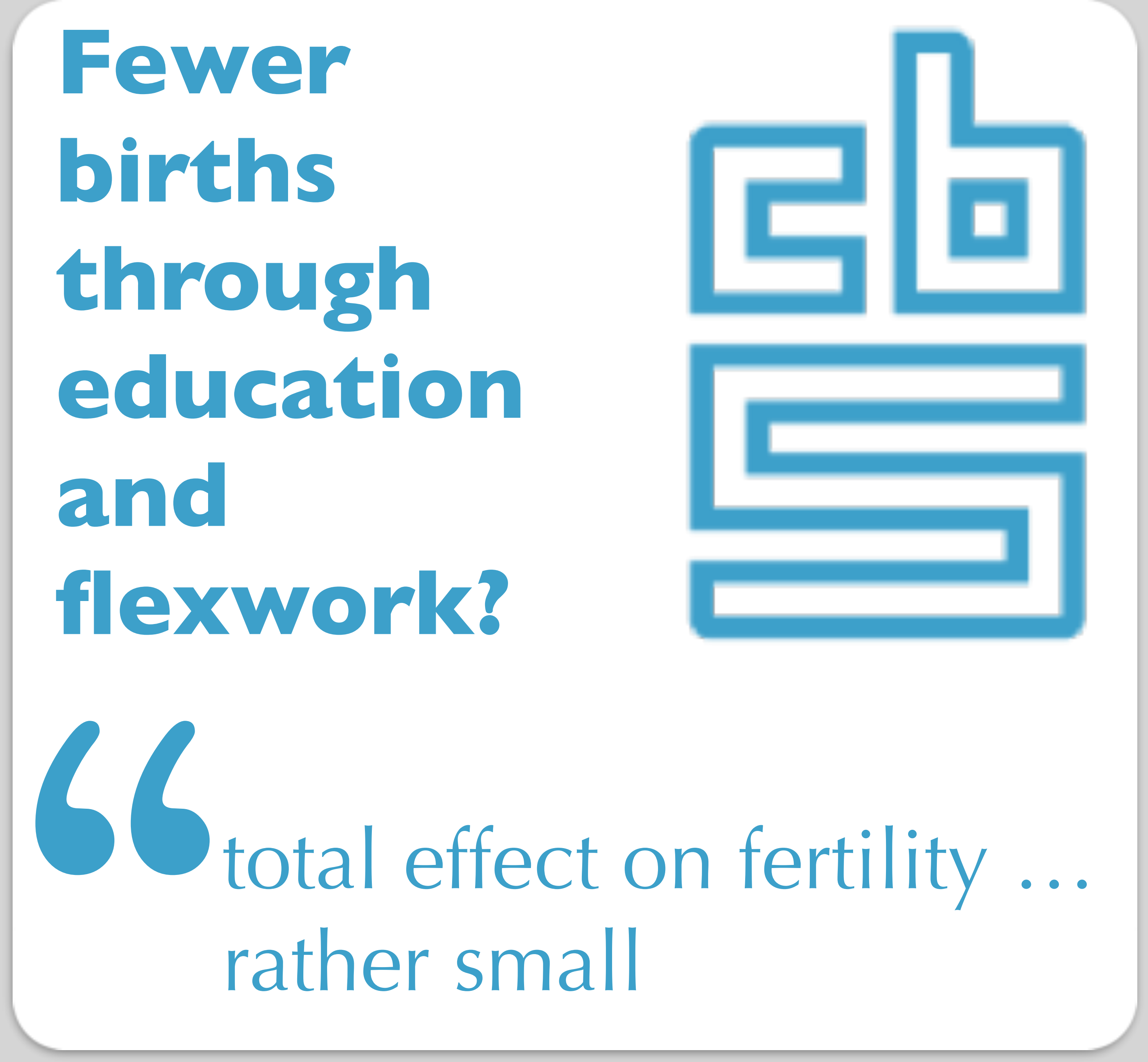
This mess we're in?

Or how simulation and prediction will advance the social sciences

gert stulp — gertstulp.com — g.stulp@rug.nl

How Well Are We Doing?

variables explain little

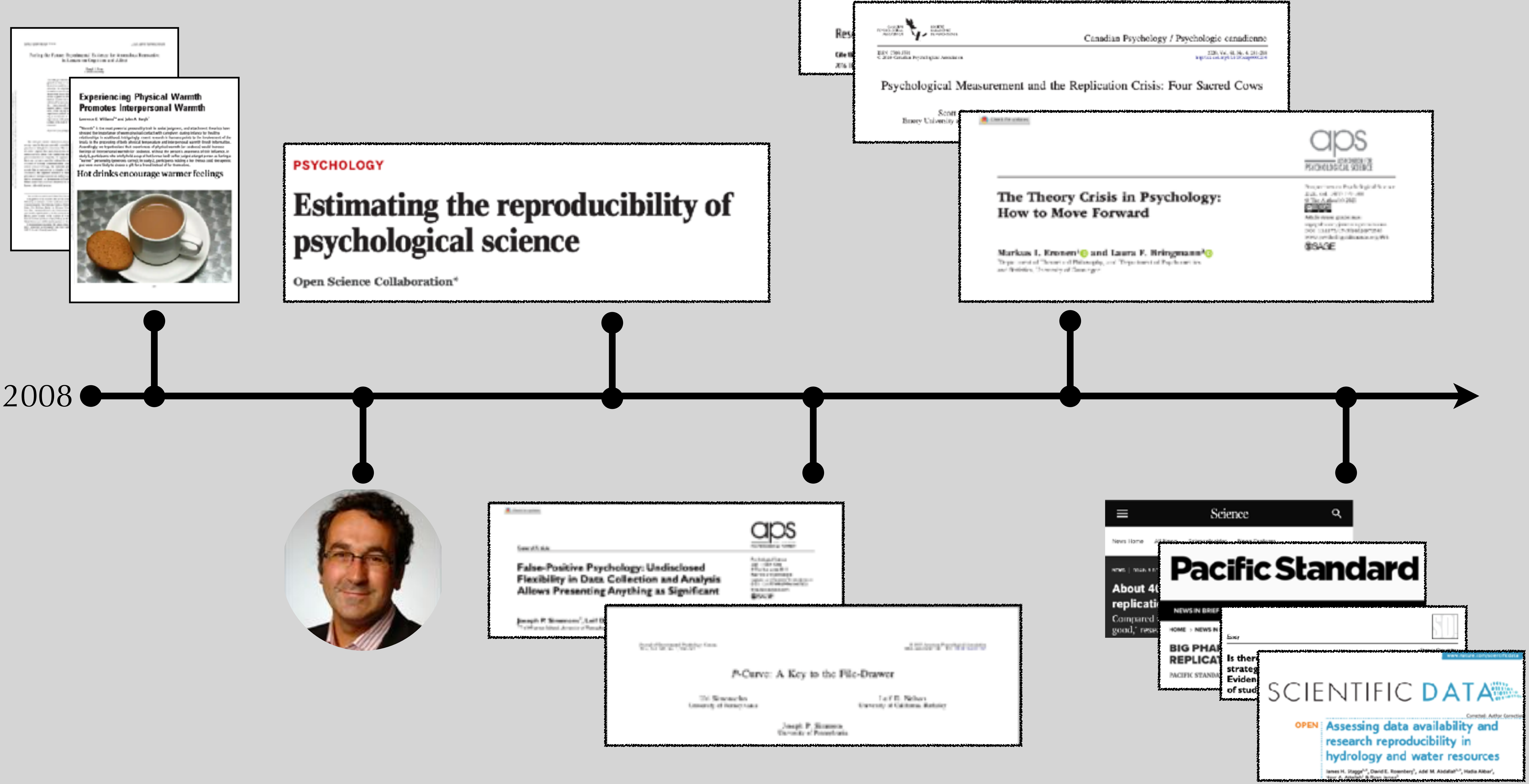


surprising patterns

incomparable results

non-replicable results

My Upbringing in Science



Replication (crisis) in Demography?

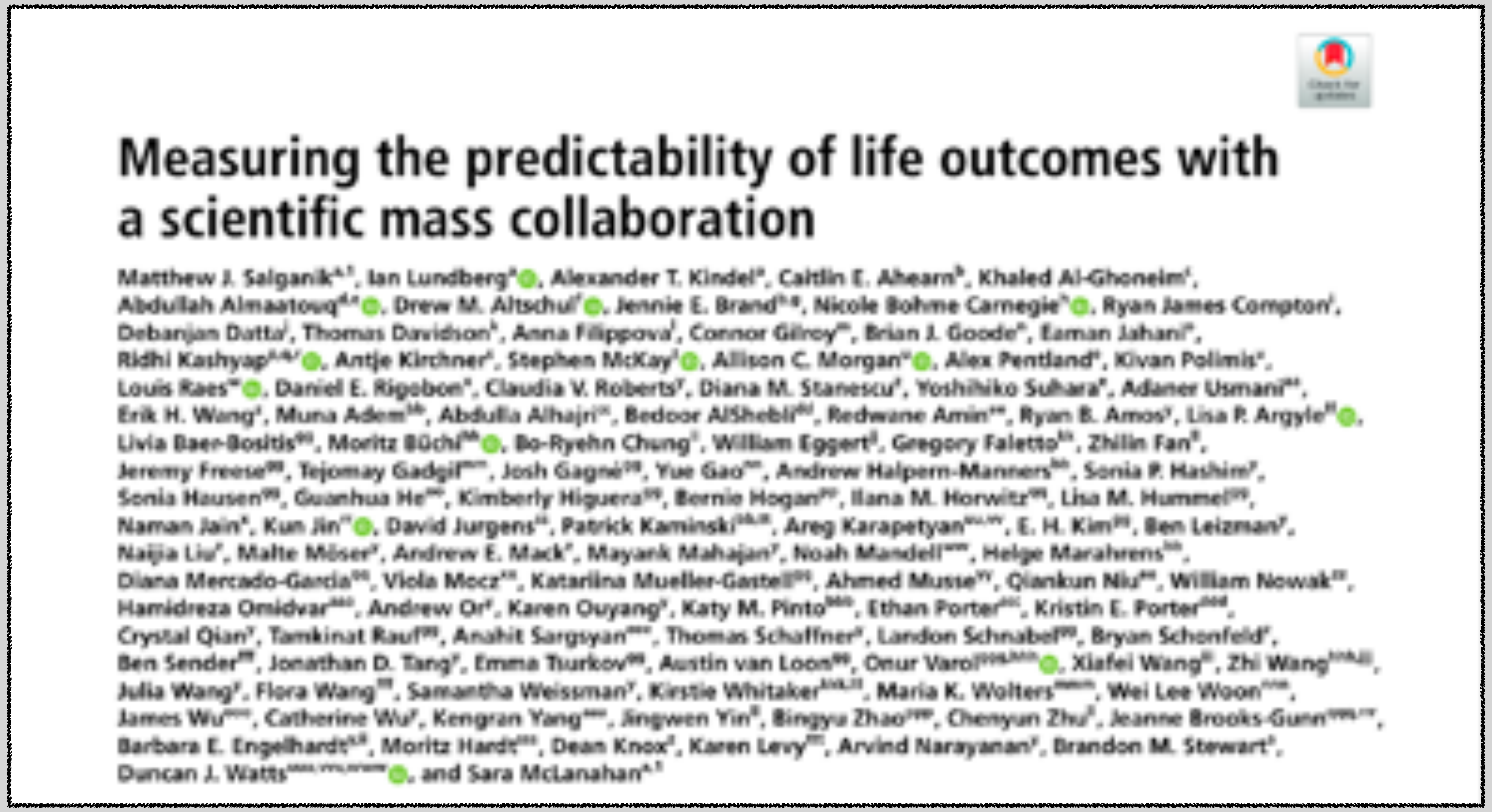
Reasons why not

- *Strong methods*
- *Strong focus on representative data*
- *Less measurement error*
- *Open data*
- *Large N*
- *Often descriptive*

Reasons why

- *Non-experimental*
- *Correlational, but little causal inference*
- *Large N, yet star gazing*
- *Controlling at will*
- *"Culture" as a get-out-of-jail-for-free card*

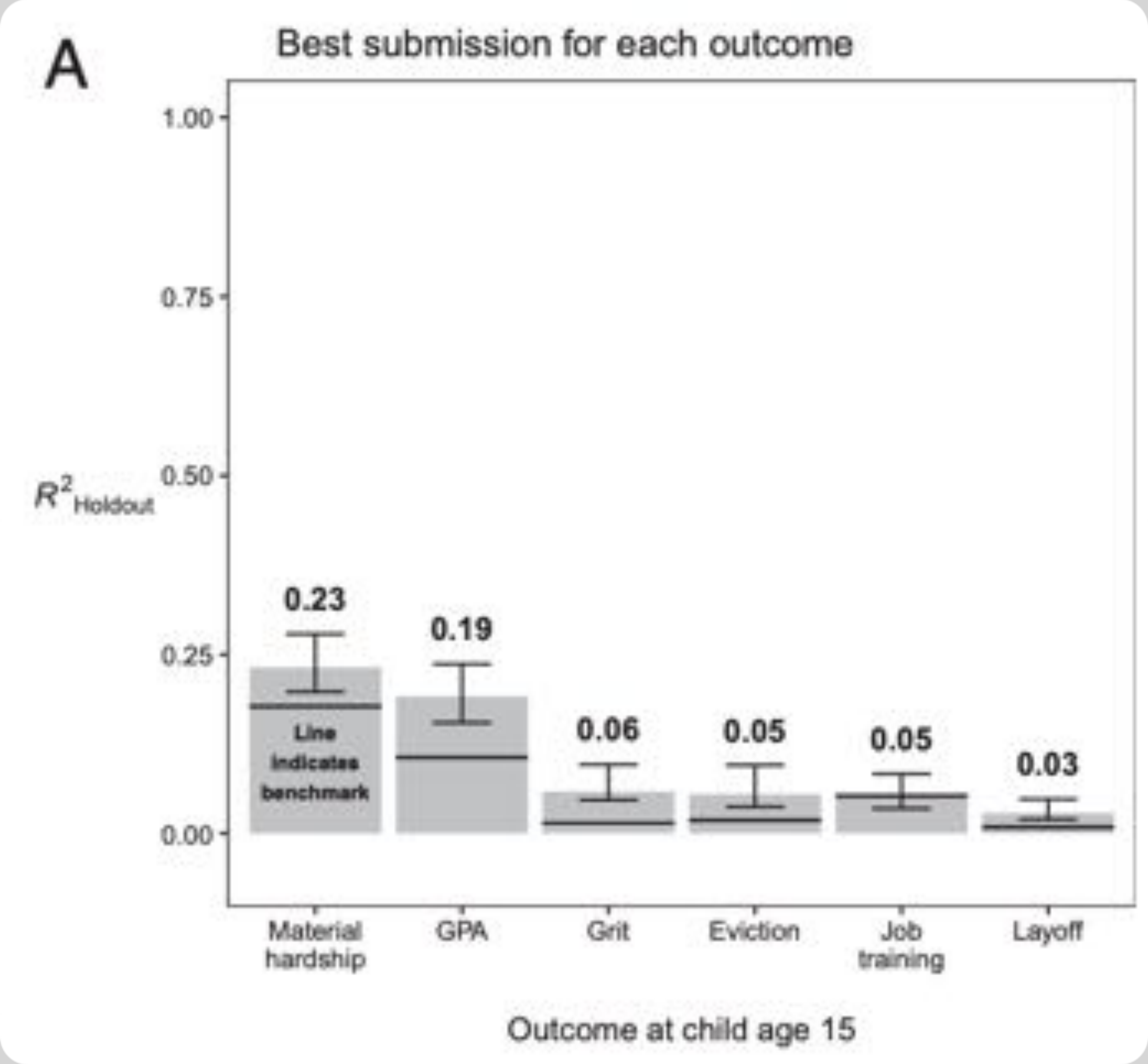
Predictability Crisis?



Measuring the predictability of life outcomes with a scientific mass collaboration

Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan

data challenge:
predicting life outcomes
based on ~6000 variables
by 160 teams
both theory- & data-driven



Predictability Crisis?

> “Social scientists studying the life course must find a way to reconcile a widespread belief that understanding has been generated by these data—as demonstrated by more than 750 published journal articles using the Fragile Families data with the fact that the very same data could not yield accurate predictions of these important outcomes.

How Well Are We Doing?

The Proposal

a shift towards **prediction** leads to a more reliable and useful social science

microsimulation can advance traditional statistical modelling

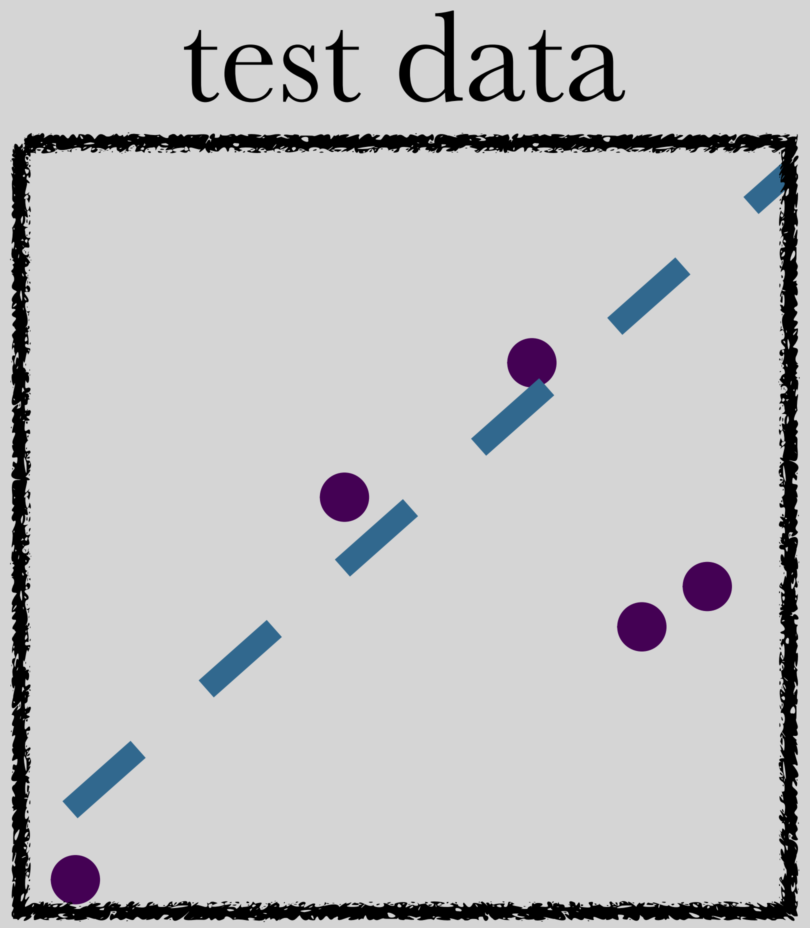
Take-Home Messages

a shift towards **prediction** leads to a more reliable and useful social science

Out-of-Sample Prediction

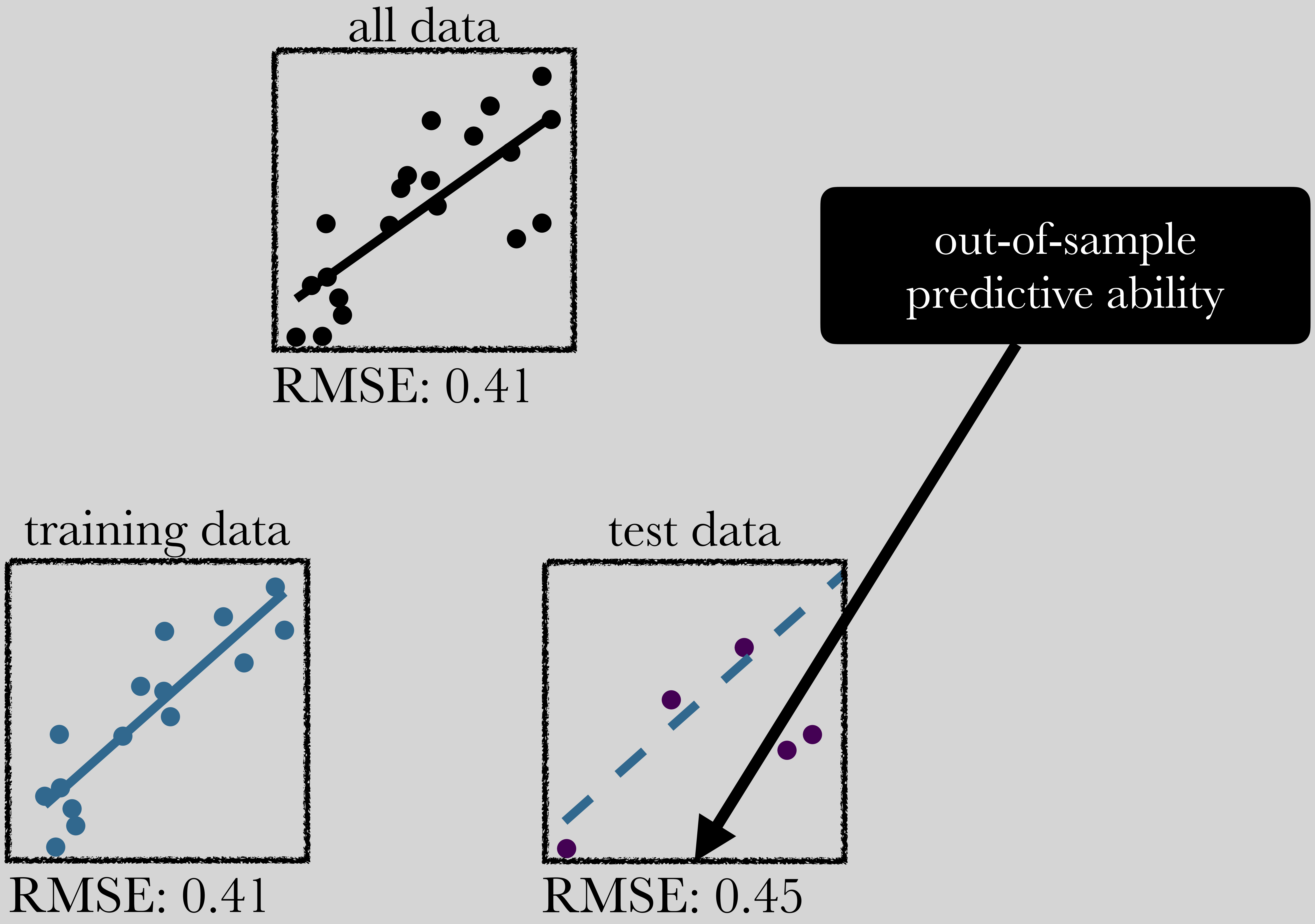
RMSE: 0.41

RMSE: 0.41

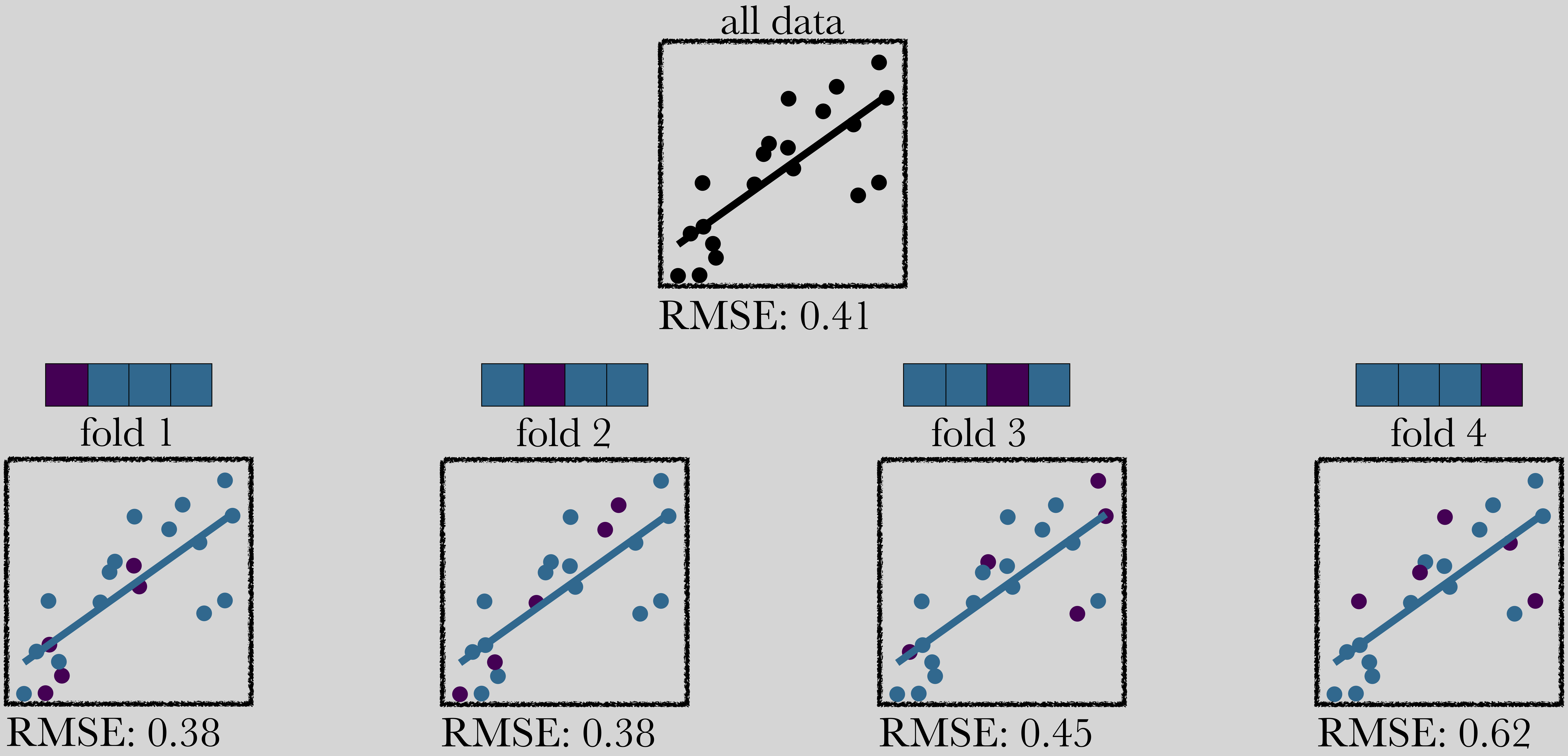


RMSE: 0.45

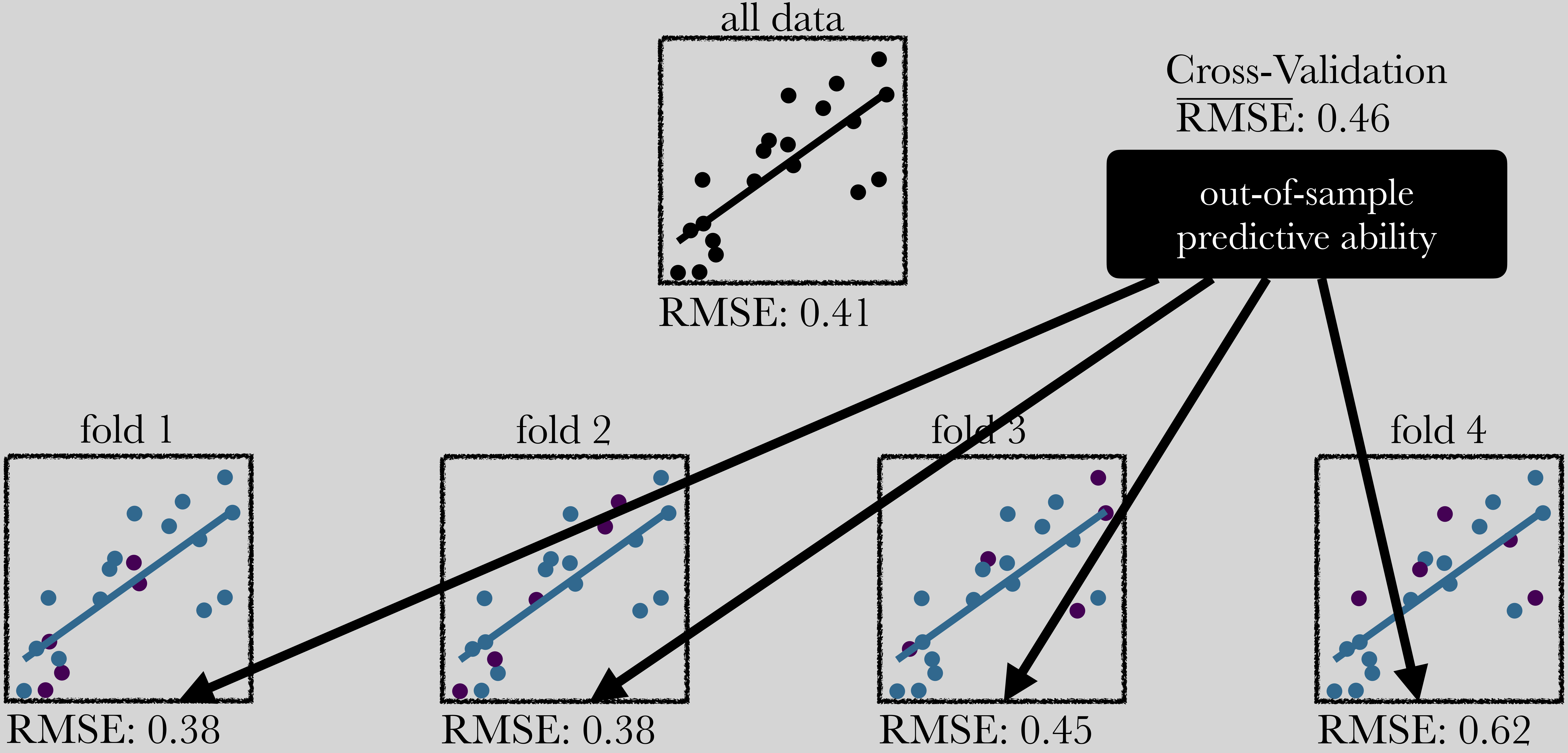
Out-of-Sample Prediction



Cross-Validation

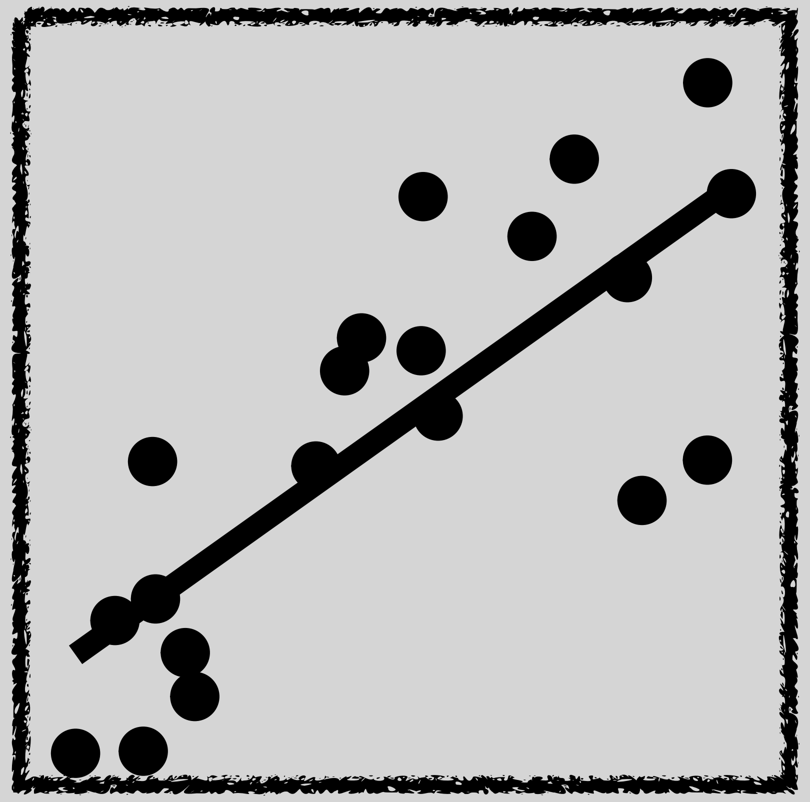


Out-of-Sample Prediction



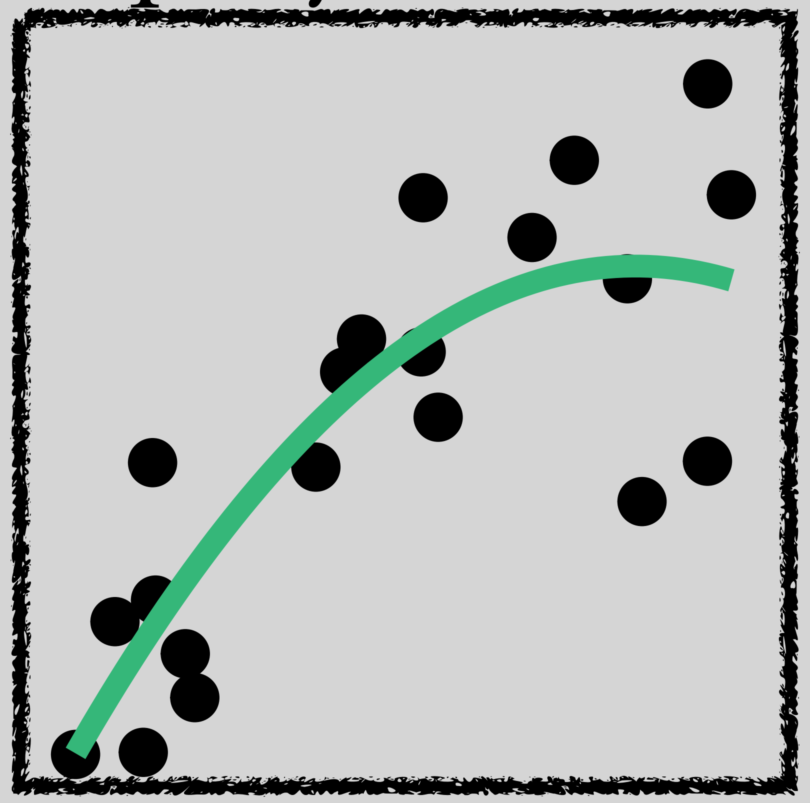
Out-of-Sample Prediction

all data



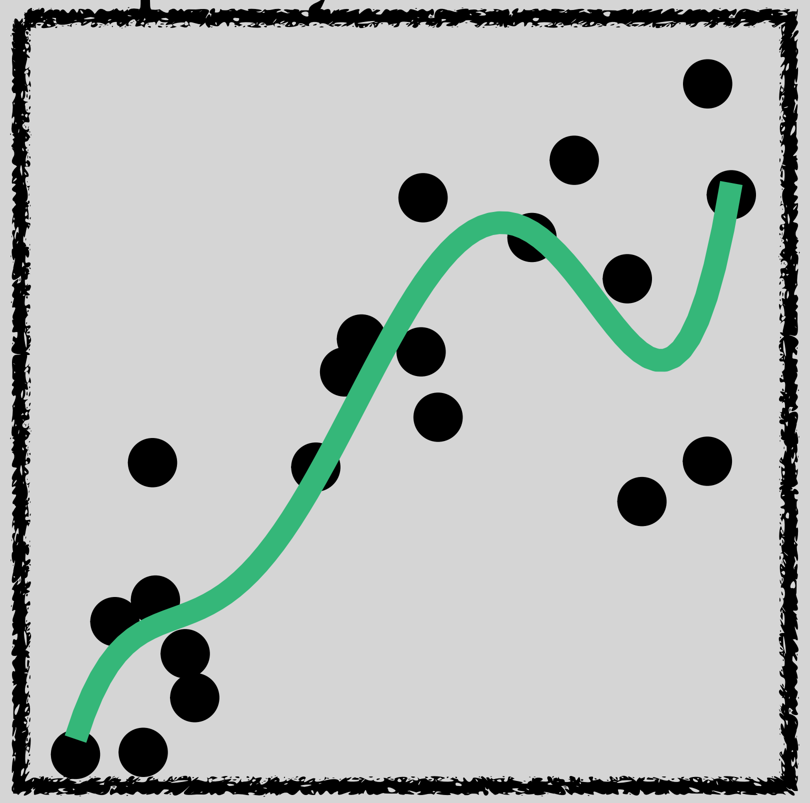
RMSE: 0.41
CV RMSE: 0.46

2 polynomials



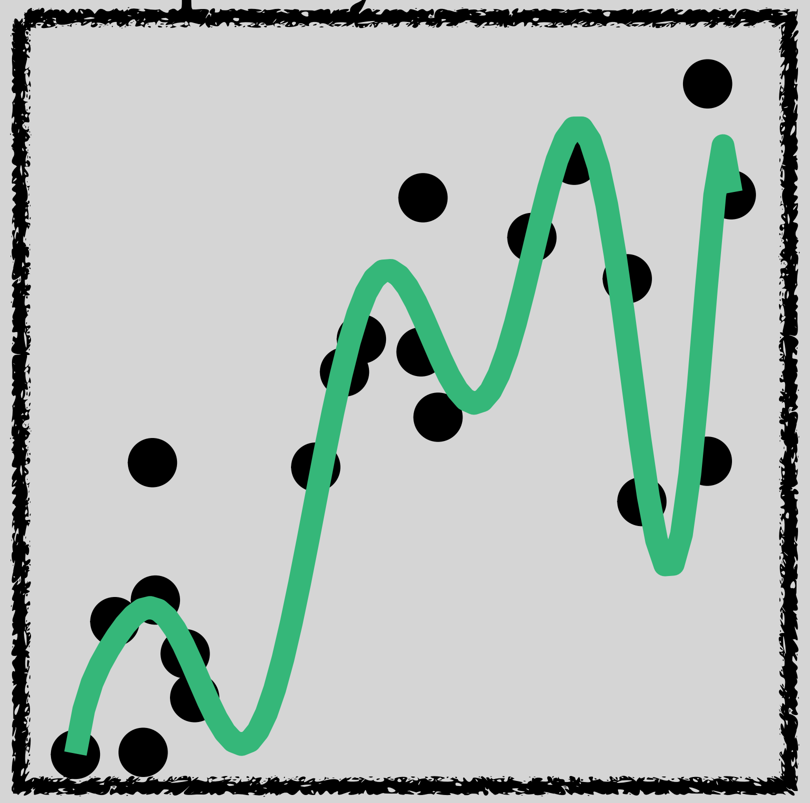
RMSE: 0.36

5 polynomials



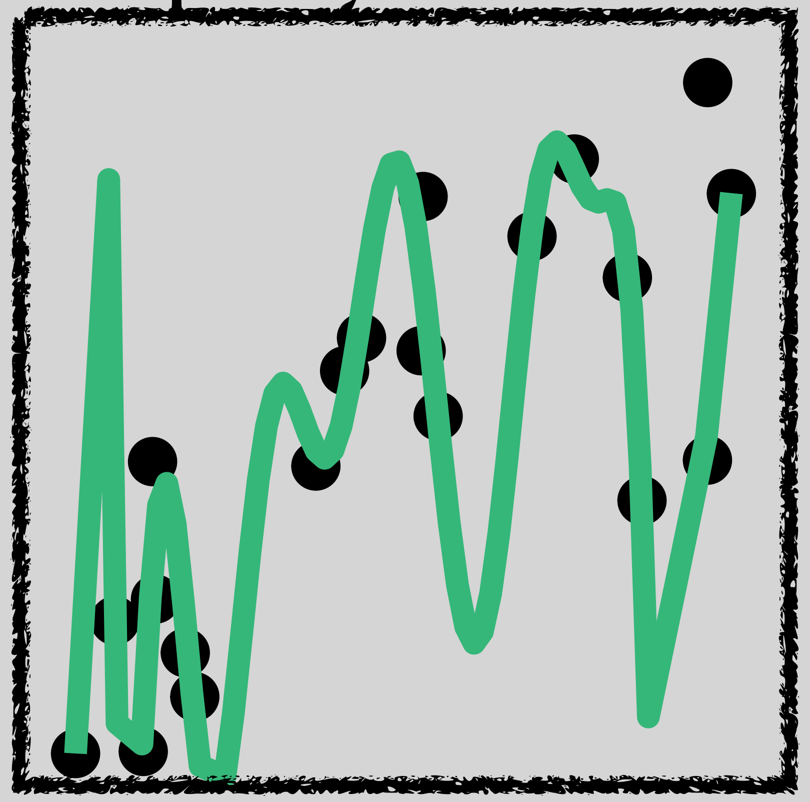
RMSE: 0.33

10 polynomials



RMSE: 0.28

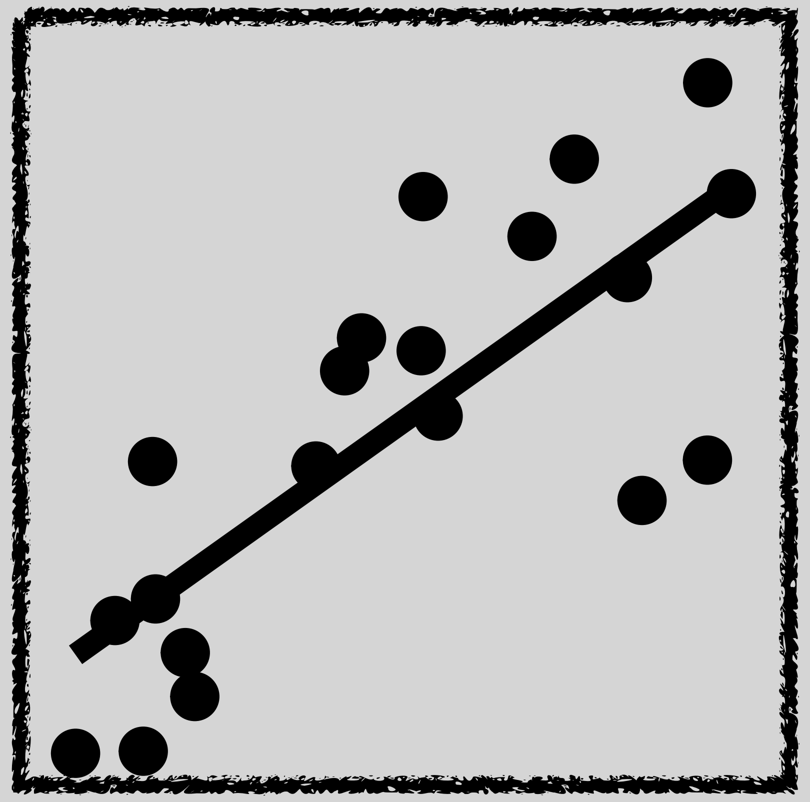
14 polynomials



RMSE: 0.22

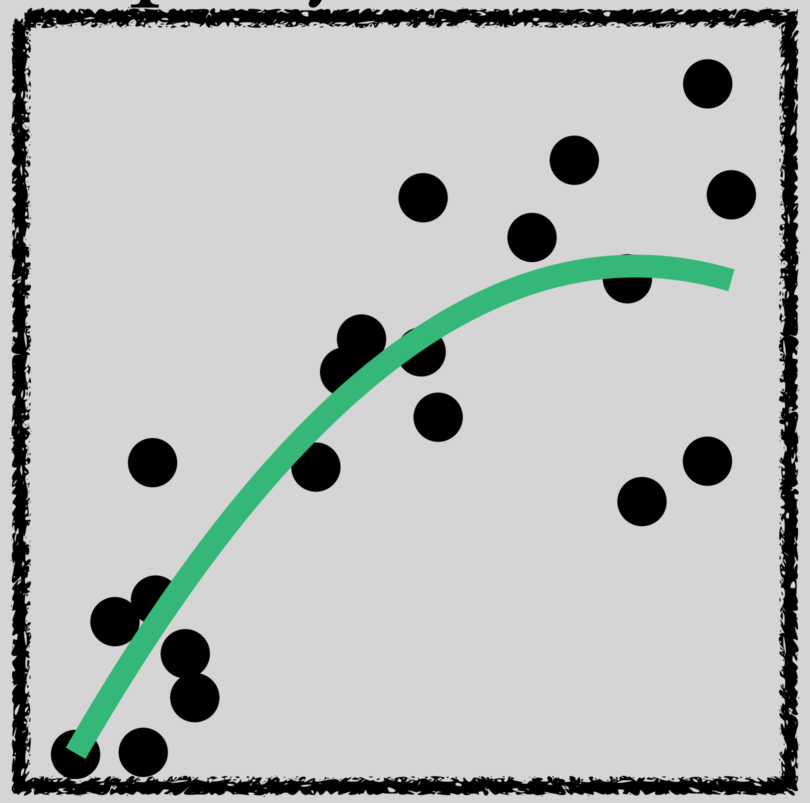
Out-of-Sample Prediction

all data



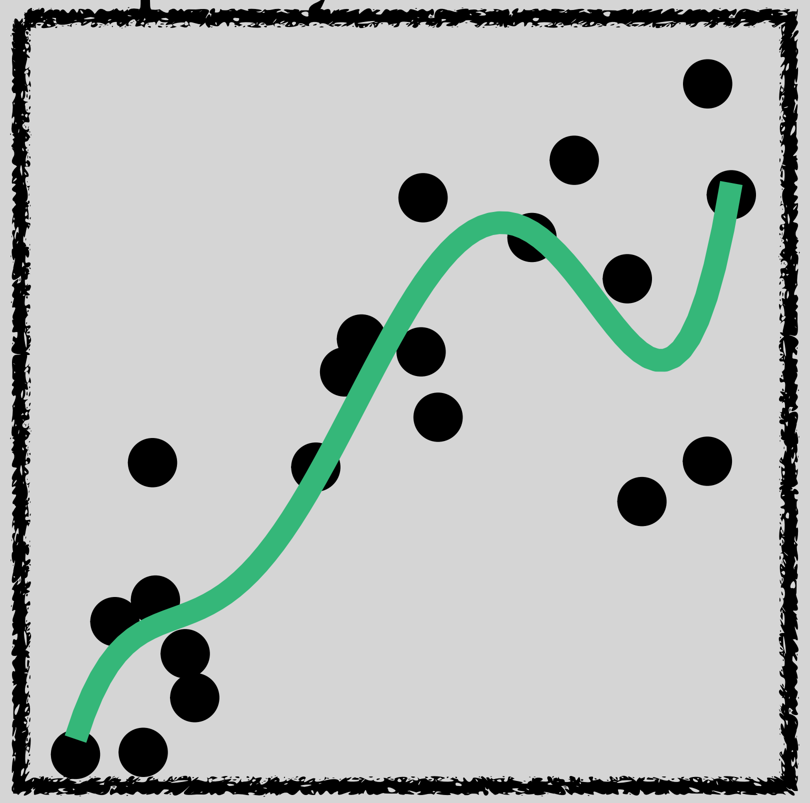
RMSE: 0.41
CV RMSE: 0.46

2 polynomials



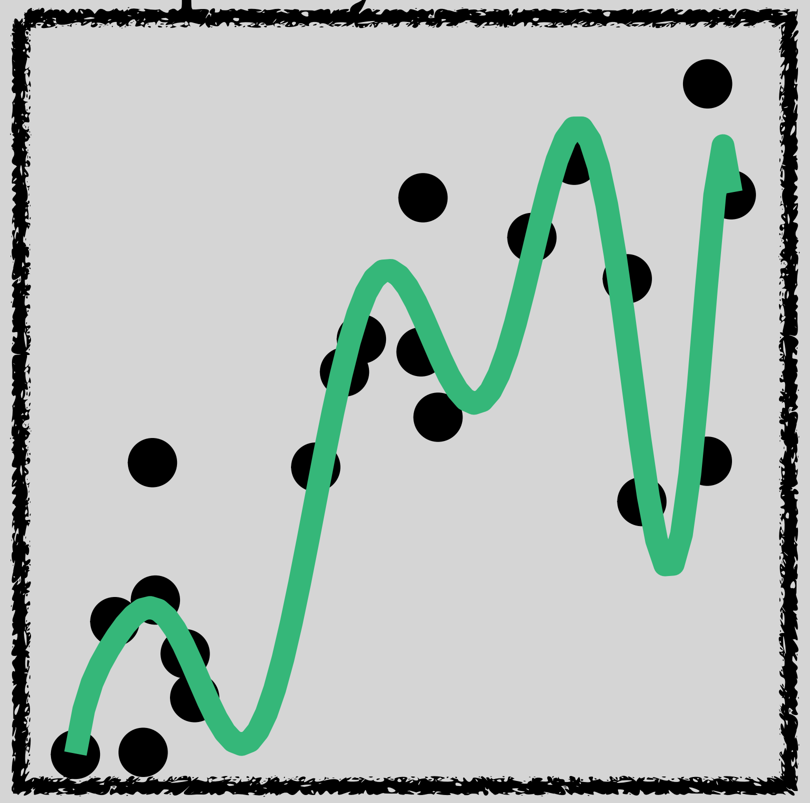
RMSE: 0.36
CV RMSE: 0.44

5 polynomials



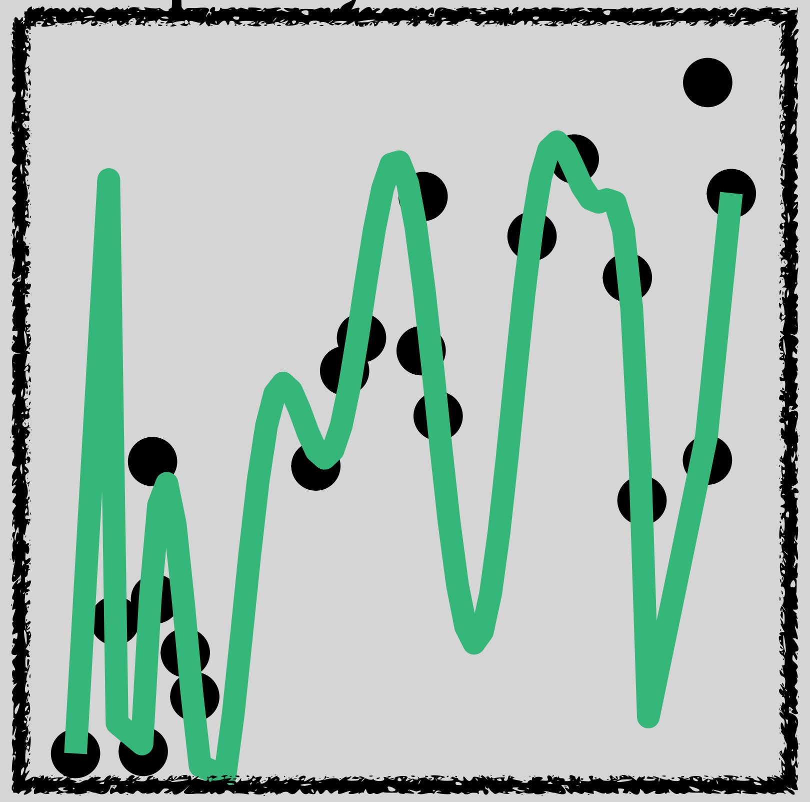
RMSE: 0.33
CV RMSE: 0.58

10 polynomials



RMSE: 0.28
CV RMSE: 6.4

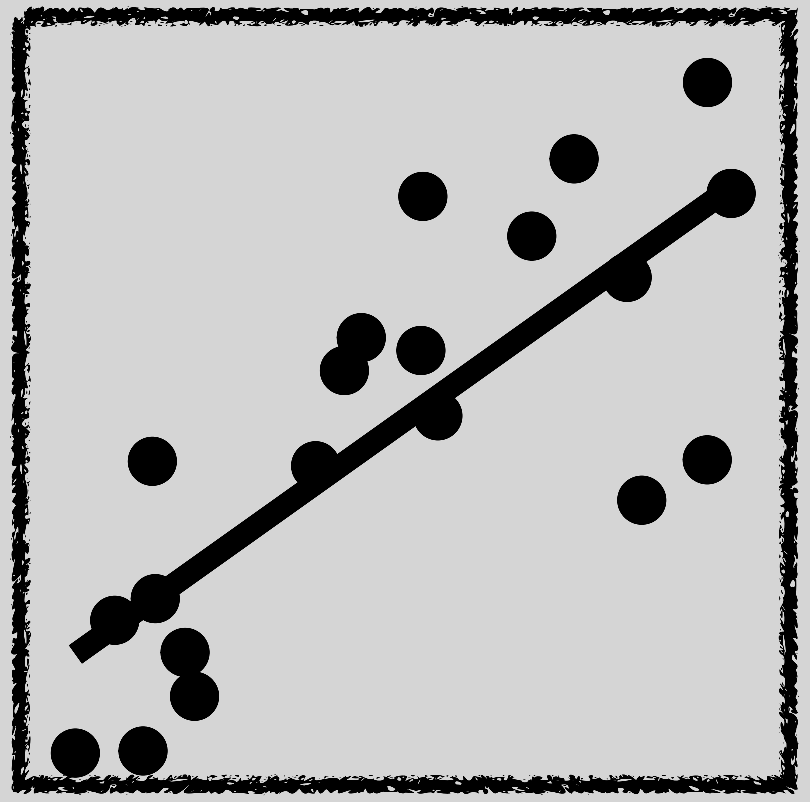
14 polynomials



RMSE: 0.22
CV RMSE: 1206

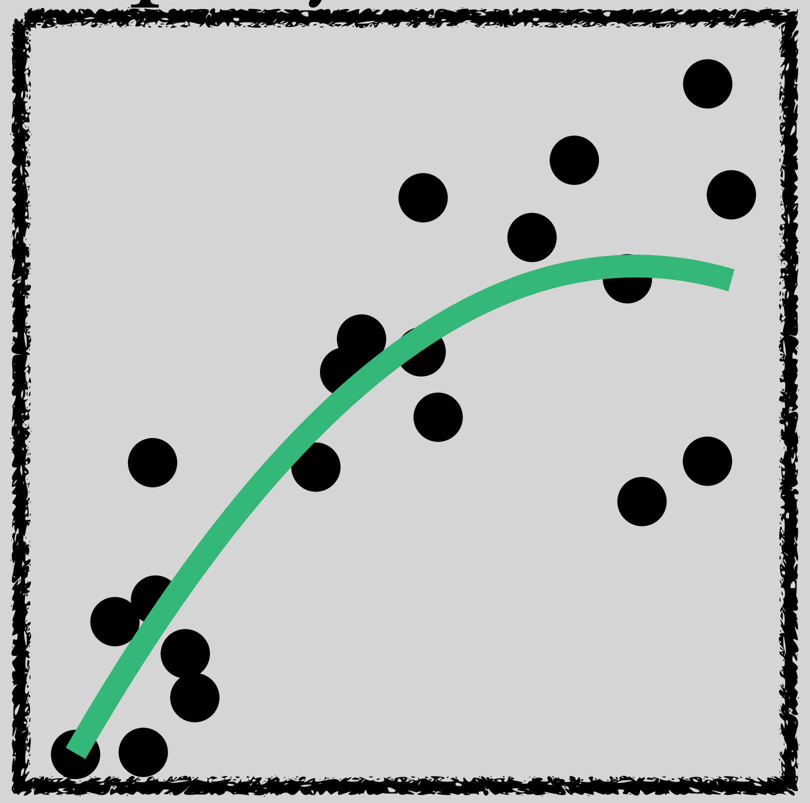
Out-of-Sample Prediction

all data



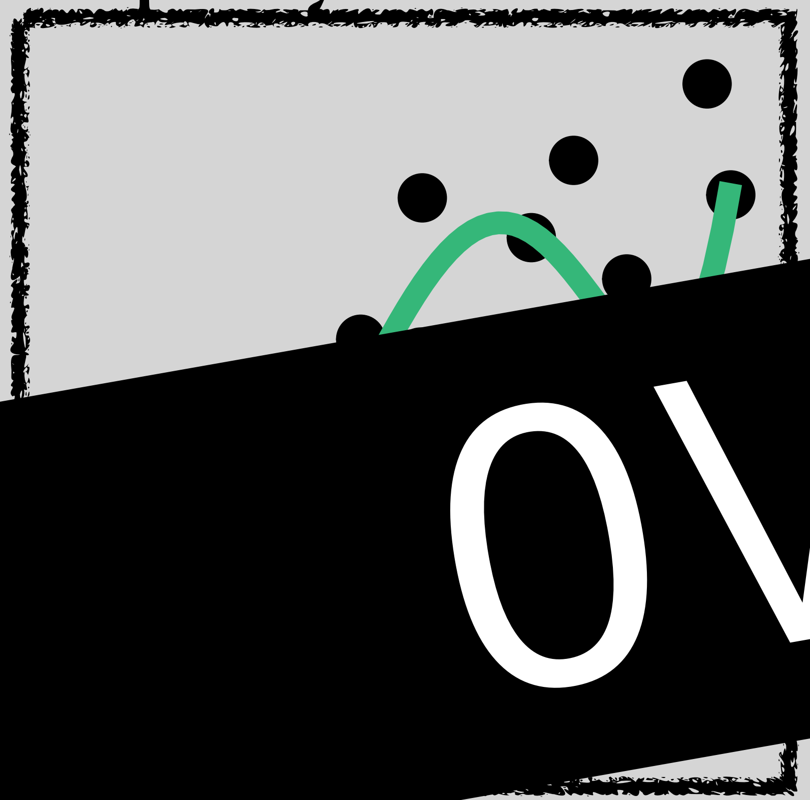
RMSE: 0.41
CV RMSE: 0.46

2 polynomials



RMSE: 0.36
CV RMSE: 0.44

5 polynomials



RMSE: 0.33
CV RMSE: 0.58

10 polynomials

RMSE: 0.28
CV RMSE: 6.4

RMSE: 0.22
CV RMSE: 1206

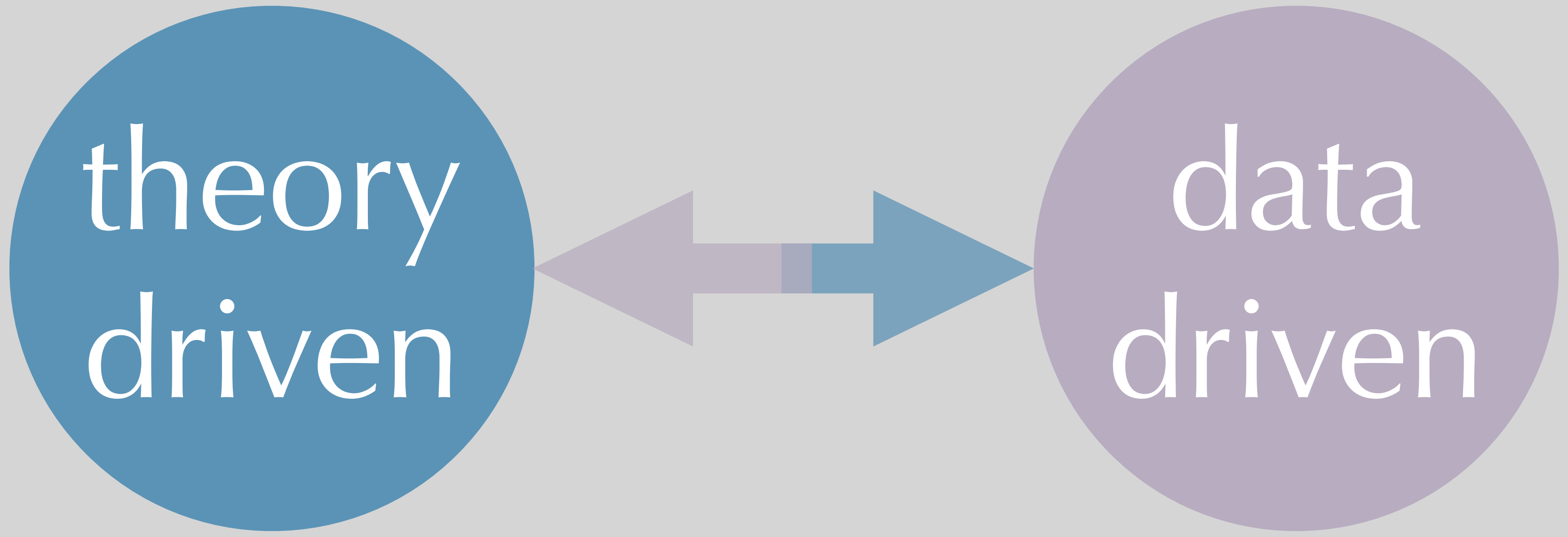
OVERFITTING

The Proposal

a shift towards **prediction** leads to a more reliable and useful social science

out-of-sample predictive ability:

clear measure of effect size



facilitates dialogue theory- and data-driven models

measure of distance theory and practice

out-of-sample predictive ability

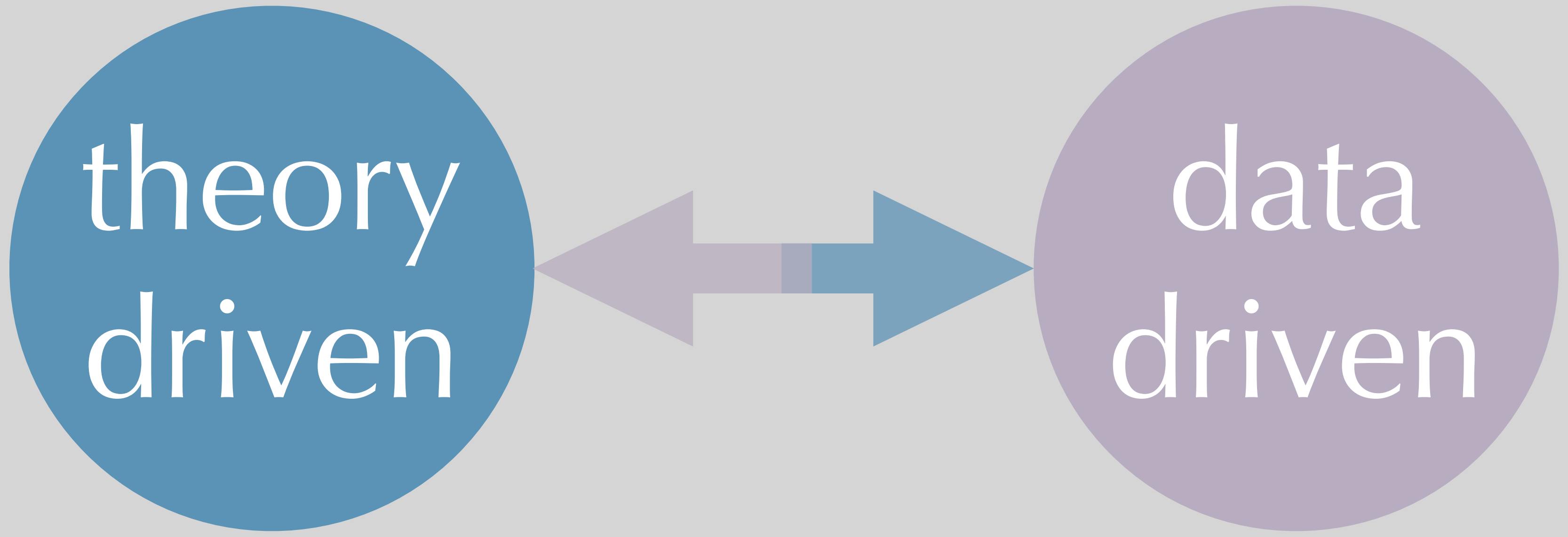
- *is easy(ier) to understand*
- *can be compared across analytical techniques*
- *can be compared across models*
- *is less gameable*

The Proposal

a shift towards **prediction** leads to a more reliable and useful social science

out-of-sample predictive ability:

clear measure of effect size



facilitates dialogue theory- and data-driven models

measure of distance theory and practice

theory driven ⟷ data driven

theory-driven *vs* data-driven

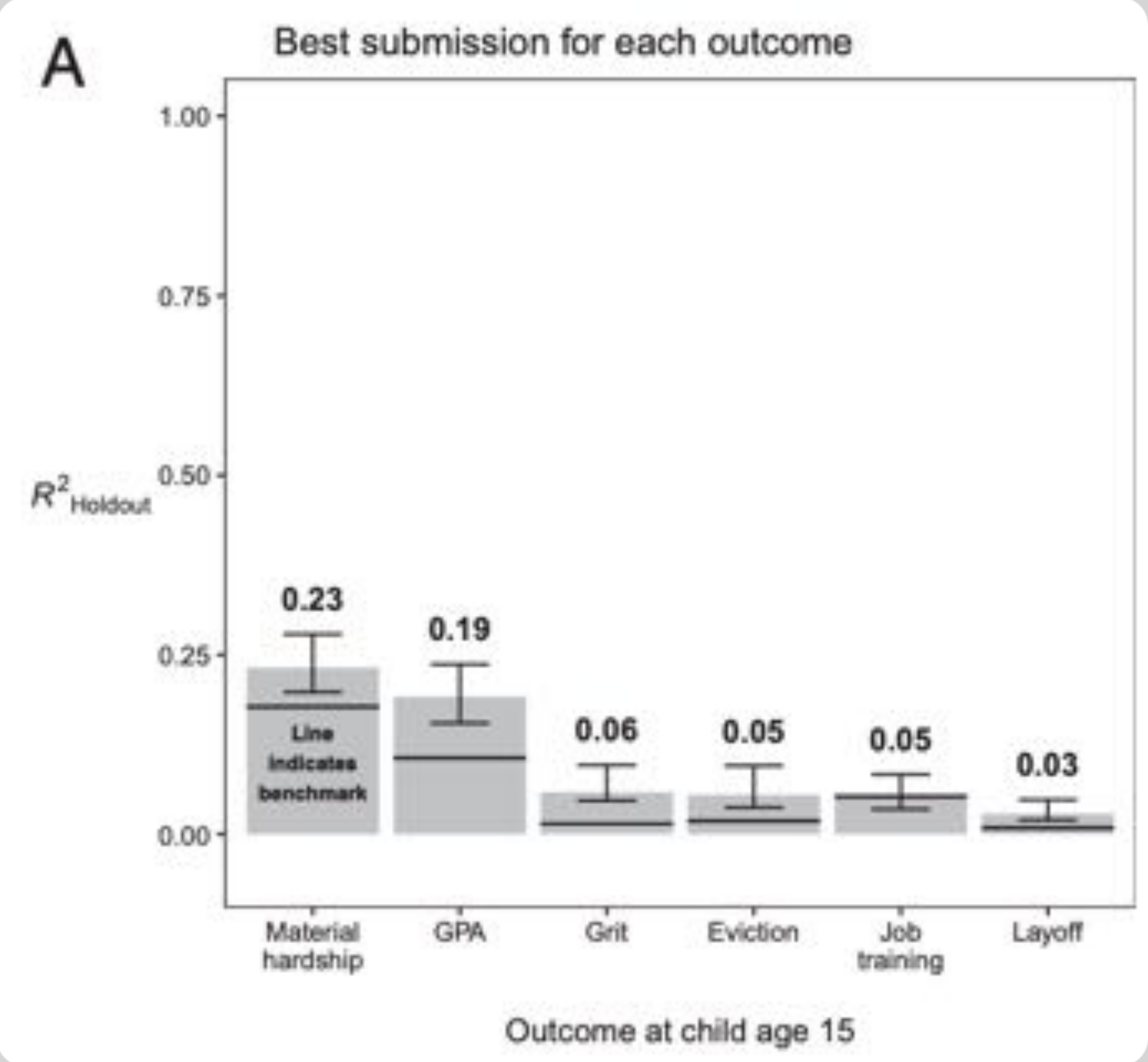
theory-driven	data-driven
focus on (causal) estimates	focus on predictive ability
support based on p-value	support based on prediction
limited number of variables (k)	k may be larger than n
NHST weird theory-testing	estimates less interpretable
long reign the linear model	computing intensive
pet variable problem	

Predictability Crisis?

Measuring the predictability of life outcomes with a scientific mass collaboration

Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan

data challenge:
predicting life outcomes
based on ~6000 variables
by 160 teams
both theory- & data-driven



data
challenge

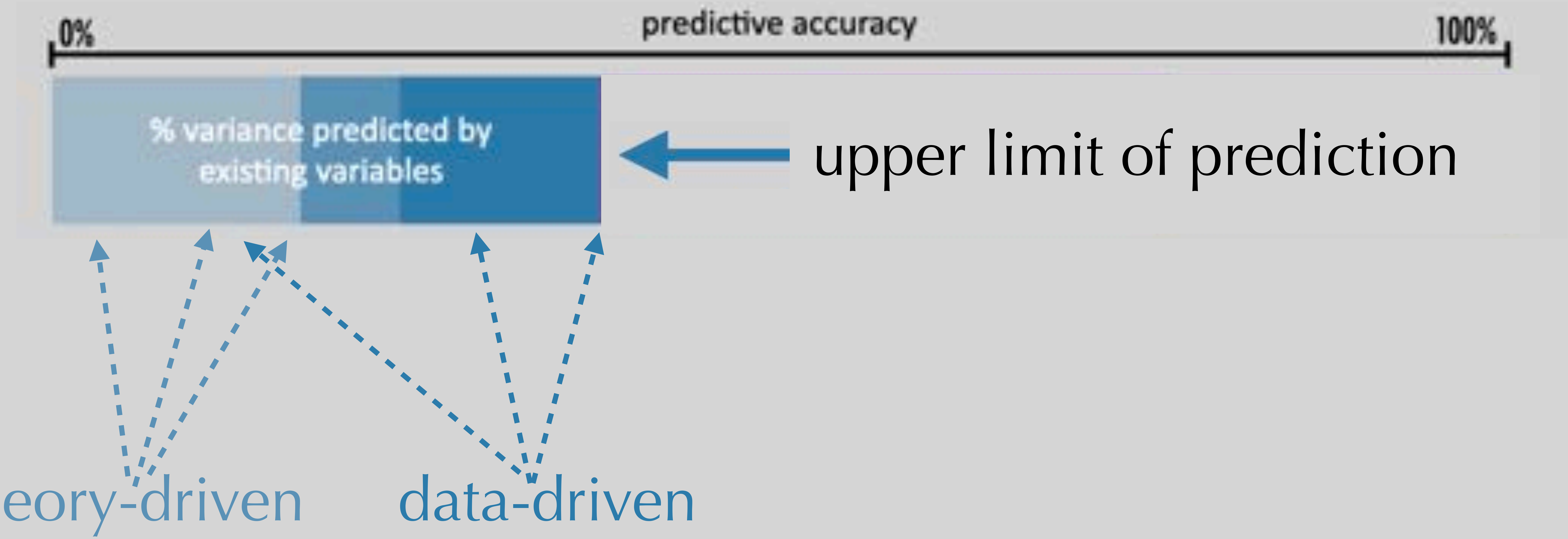
theory driven

data driven

theory- and data-driven teams
engage in common task
using common data
and common metric

Data Challenge

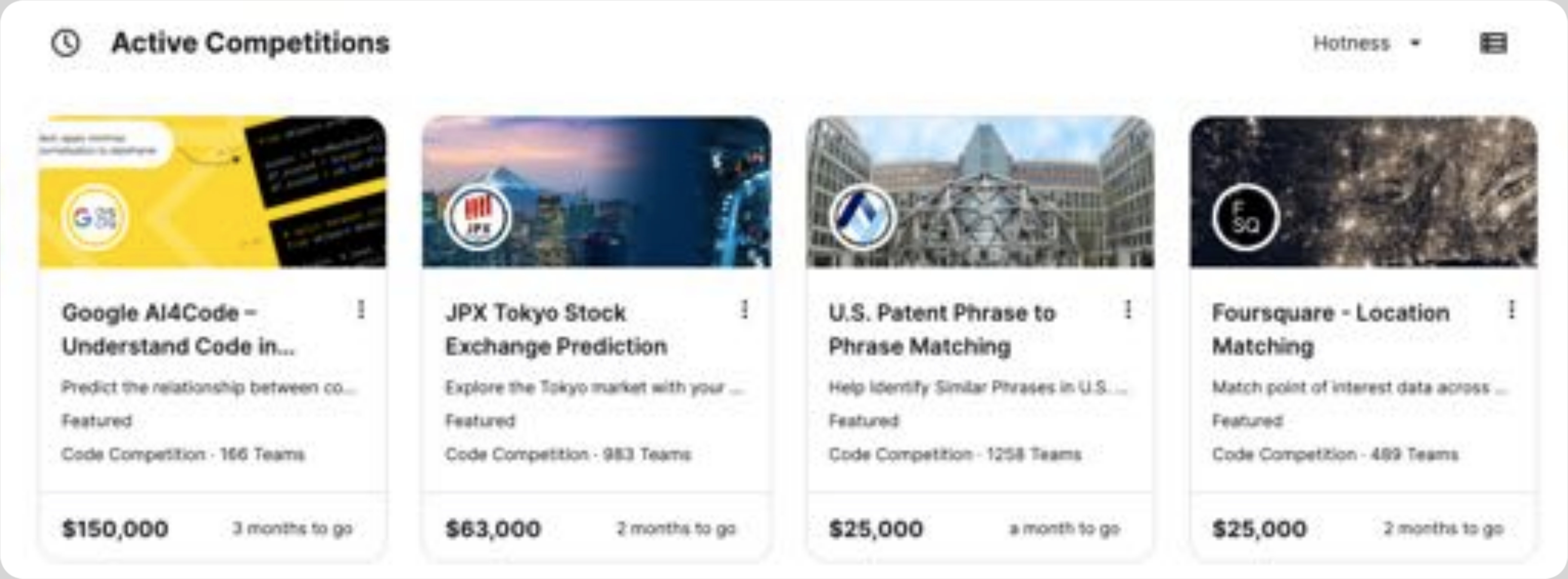
theory- and data-driven teams
engage in common task
using common data
and common metric



Prediction Benchmarks

> Progress usually comes from many small improvements; a change of 1% can be a reason to break out the champagne
>
> Liberman, 2012

kaggle



“secret sauce of data science

Donoho, 2015

FERTILITY PREDICTION CHALLENGE

March-August 2024

University of Groningen, Netherlands

0.54*

Is the current best [known to us] F1-score of a classifier that predicts who is going to have a child in the next three years

CAN YOU BEAT THIS SCORE?

- Do you want to contribute to research on fertility behavior and the methodology of using prediction in social sciences?
- Are you interested in working with unique registry-based datasets, including a social network for the entire Dutch population?
- Are you looking for an engaging practical task for your machine learning course or workshop?
- Or are you simply curious about the challenge and want to learn more about its design and prizes?

Sign up here to receive an update when the registration for the challenge opens and details are available

Contacts:
Gert Stulp g.stulp@rug.nl
Elizaveta Sivak e.sivak@rug.nl

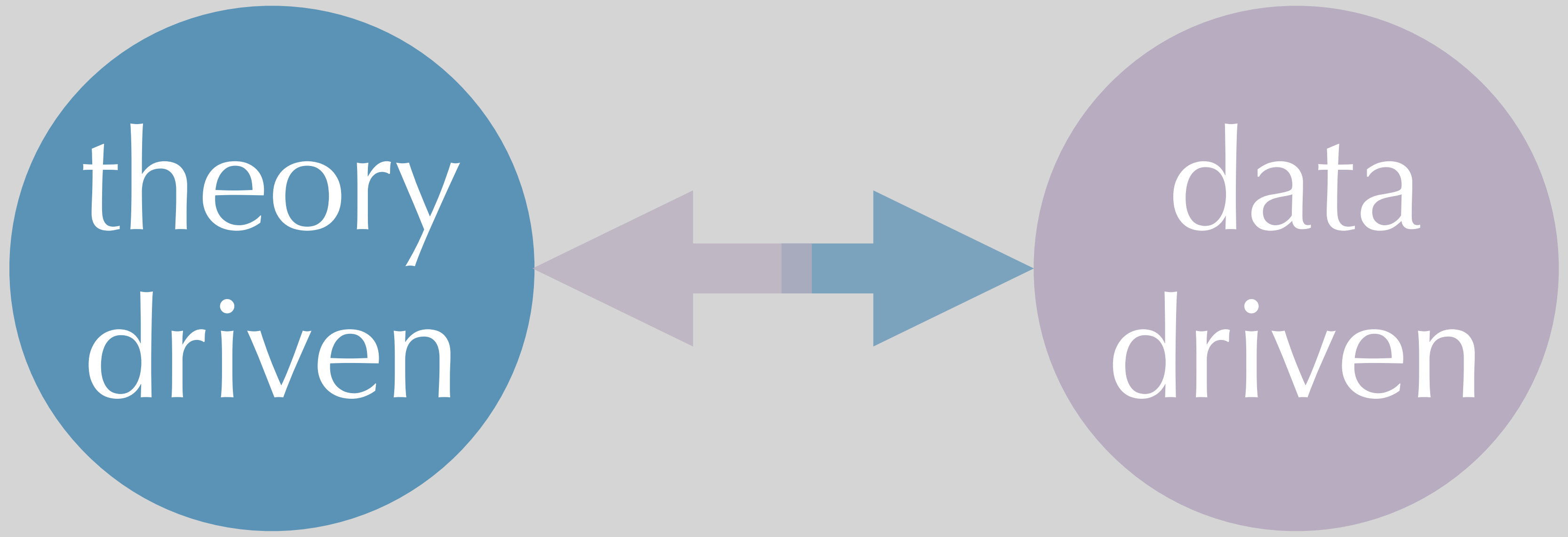
* This result was obtained by the STL Trio Titans team at the data challenge at the SICSS-ODISSEI summer school in June 2023.

Take-Home Messages

a shift towards **prediction** leads to a more reliable and useful social science

out-of-sample predictive ability:

clear measure of effect size



facilitates dialogue theory- and data-driven models

measure of distance theory and practice

out-of-sample predictive ability
is a measure of how useful
our theory is in the real world

Article

Journal of peace

The perils of policy by p-value: Predicting civil conflicts

Michael D Ward
Department of Political Science, Duke University
Brian D Greenhill
Department of Political Science, University of Washington
Kristin M Bakke
Department of Political Science, University College London

aps
ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
SAGE

So Useful as a Good Theory? The Practicality Crisis in (Social) Psychological Theory

Elliot T. Berkman and Sylas M. Wilson
Department of Psychology and Center for Translational Neuroscience, University of Oregon

out-of-sample predictive ability
is a measure of how useful
our theory is in the real world

PNAS

CrossMark

Why significant variables aren't automatically good predictors

Adeline Lo[a], Herman Chernoff[b,1], Tian Zheng[c], and Shaw-Hwa Lo[c,1]

[a]Department of Political Science, University of California, San Diego, La Jolla, CA 92093; [b]Department of Statistics, Harvard University, Cambridge, MA 02138; and [c]Department of Statistics, Columbia University, New York, NY 10027

Contributed by Herman Chernoff, September 17, 2015 (sent for review December 15, 2014)

Thus far, genome-wide association studies (GWAS) have been dis-

From the scientist's point of view there are two basic problems,

out-of-sample predictive ability
is a measure of how useful
our theory is in the real world

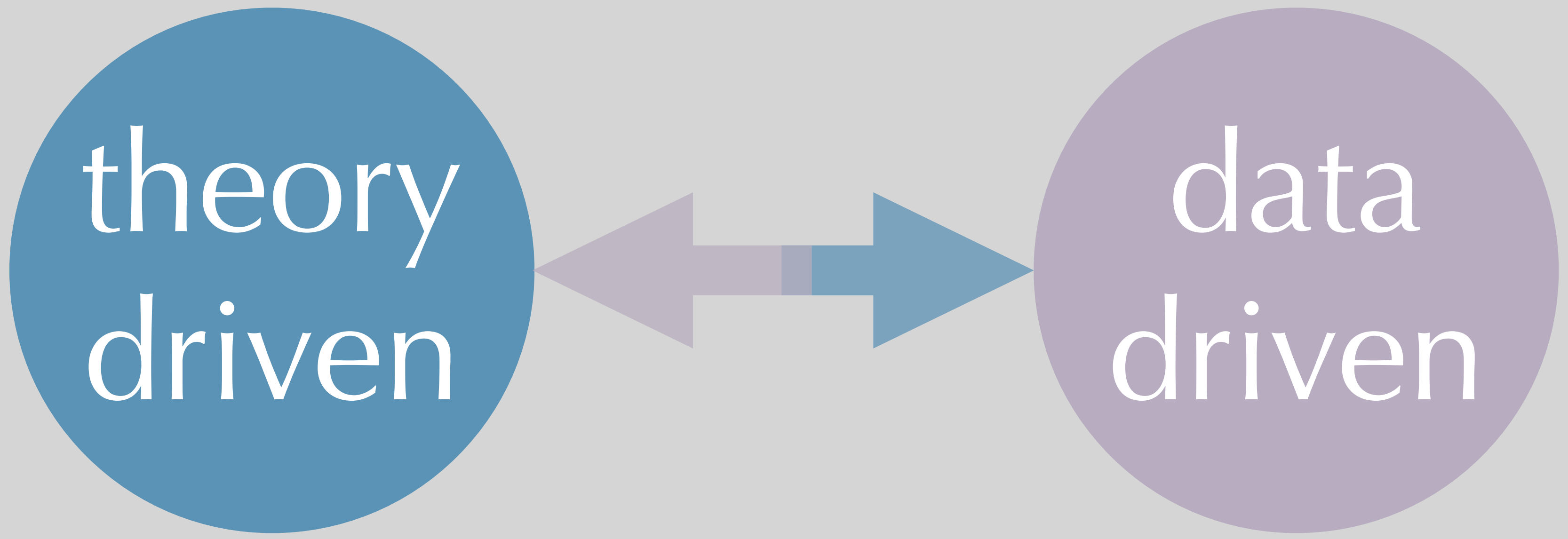
“Social scientists studying the life course must find a way to reconcile a widespread belief that understanding has been generated by these data—as demonstrated by more than 750 published journal articles using the Fragile Families data with the fact that the very same data could not yield accurate predictions of these important outcomes.

Take-Home Messages

a shift towards **prediction** leads to a more reliable and useful social science

out-of-sample predictive ability:

clear measure of effect size



facilitates dialogue theory- and data-driven models

measure of distance theory and practice

No Panacea

Patterns

CellPress
OPEN ACCESS

Article

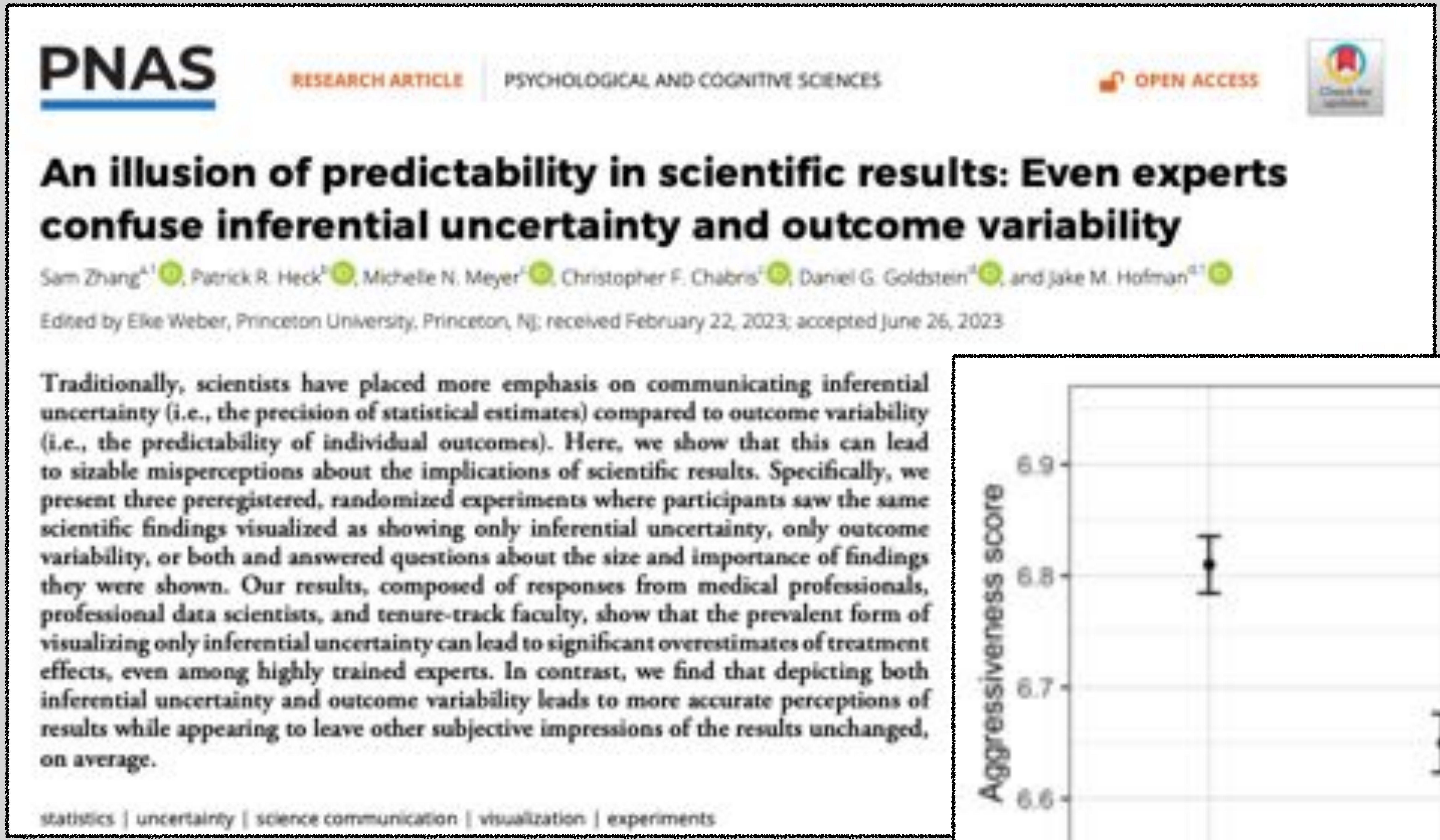
Leakage and the reproducibility crisis in machine-learning-based science

Sayash Kapoor[1,2,*] and Arvind Narayanan[1]
[1]Department of Computer Science and Center for Information Technology Policy, Princeton University, Princeton, NJ 08540, USA
[2]Lead contact
*Correspondence: sayashk@princeton.edu
https://doi.org/10.1016/j.patter.2023.100804

THE BIGGER PICTURE Machine learning (ML) is widely used across dozens of scientific fields. However, a common issue called "data leakage" can lead to errors in data analysis. We surveyed a variety of research that uses ML and found that data leakage affects at least 294 studies across 17 fields, leading to overoptimistic findings. We classified these errors into eight different types. We propose a solution: model info sheets that can be used to identify and prevent each of these eight types of leakage. We also tested the reproducibility of ML in a specific field: predicting civil wars, where complex ML models were thought to outperform traditional statistical models. Interestingly, when we corrected for data leakage, the supposed superiority of ML models disappeared: they did not perform any better than older methods. Our work serves as a cautionary note against taking results in ML-based science at face value.

1 2 3 4 5 **Development/Pre-production:** Data science output has been rolled out/validated across multiple domains/problems

But Much Needed



PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

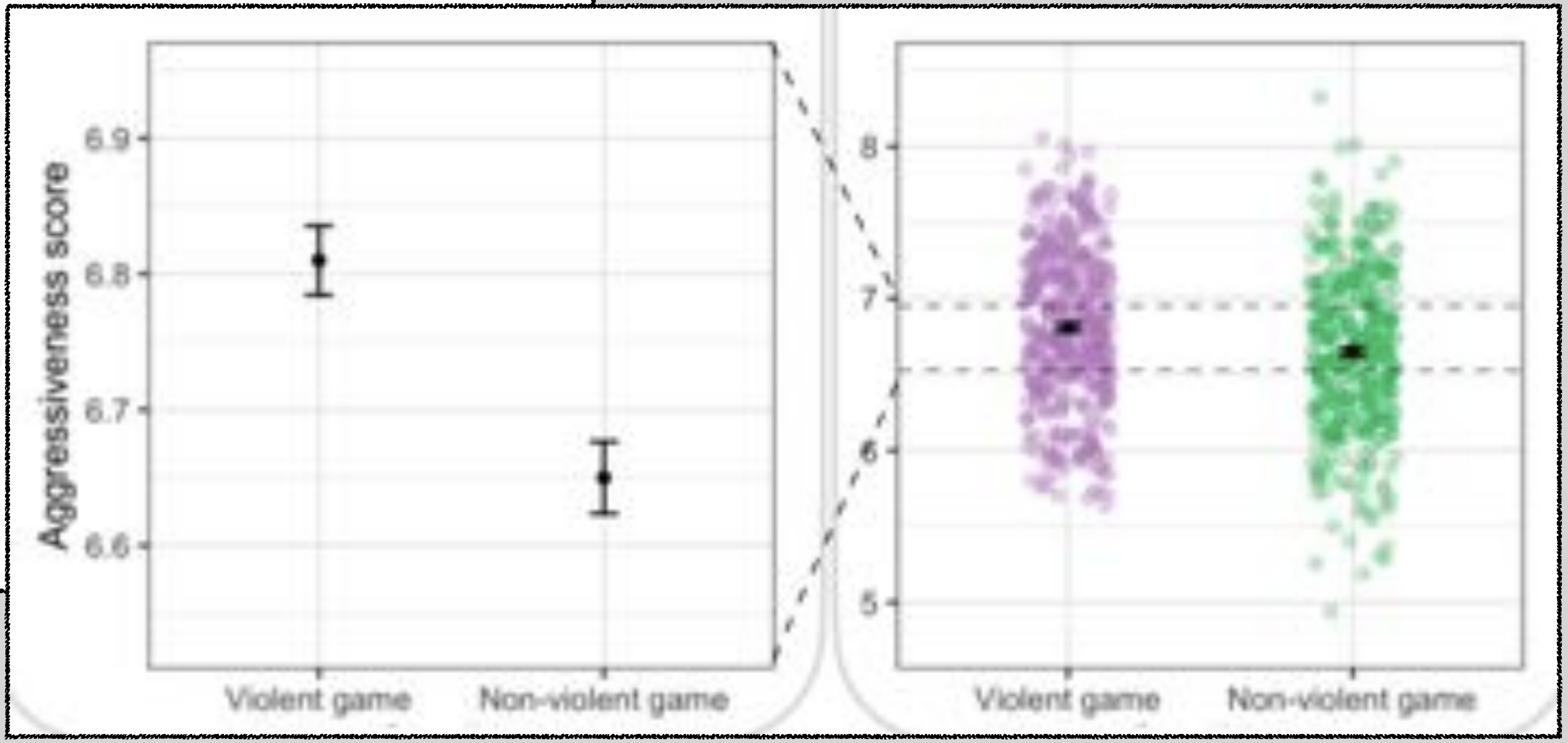
An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability

Sam Zhang[a,1], Patrick R. Heck[b], Michelle N. Meyer[c], Christopher F. Chabris[c], Daniel G. Goldstein[d], and Jake M. Hofman[d,1]

Edited by Elke Weber, Princeton University, Princeton, NJ; received February 22, 2023; accepted June 26, 2023

Traditionally, scientists have placed more emphasis on communicating inferential uncertainty (i.e., the precision of statistical estimates) compared to outcome variability (i.e., the predictability of individual outcomes). Here, we show that this can lead to sizable misperceptions about the implications of scientific results. Specifically, we present three preregistered, randomized experiments where participants saw the same scientific findings visualized as showing only inferential uncertainty, only outcome variability, or both and answered questions about the size and importance of findings they were shown. Our results, composed of responses from medical professionals, professional data scientists, and tenure-track faculty, show that the prevalent form of visualizing only inferential uncertainty can lead to significant overestimates of treatment effects, even among highly trained experts. In contrast, we find that depicting both inferential uncertainty and outcome variability leads to more accurate perceptions of results while appearing to leave other subjective impressions of the results unchanged, on average.

statistics | uncertainty | science communication | visualization | experiments



The Proposal

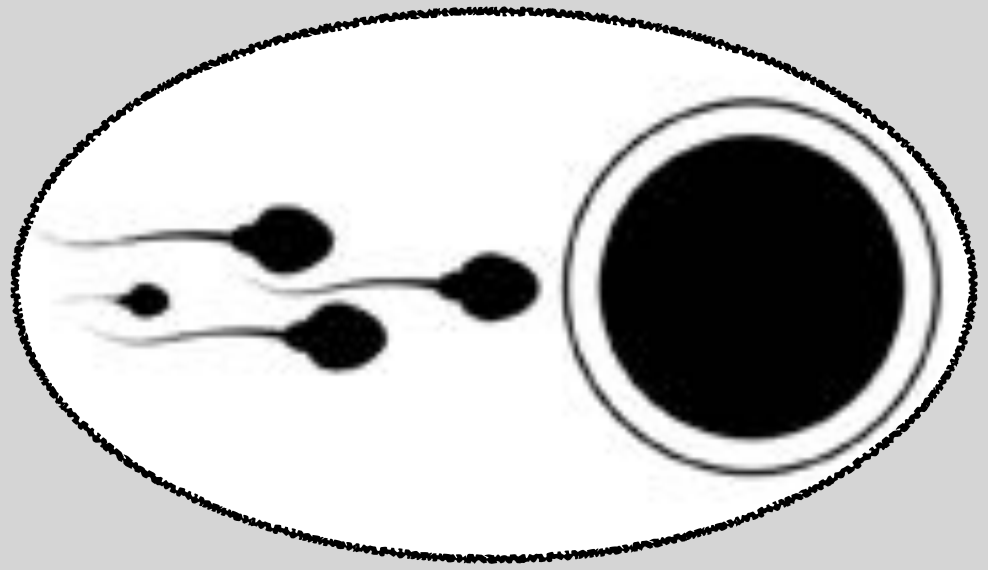
a shift towards **prediction** leads to a more reliable and useful social science

microsimulation can advance traditional statistical modelling

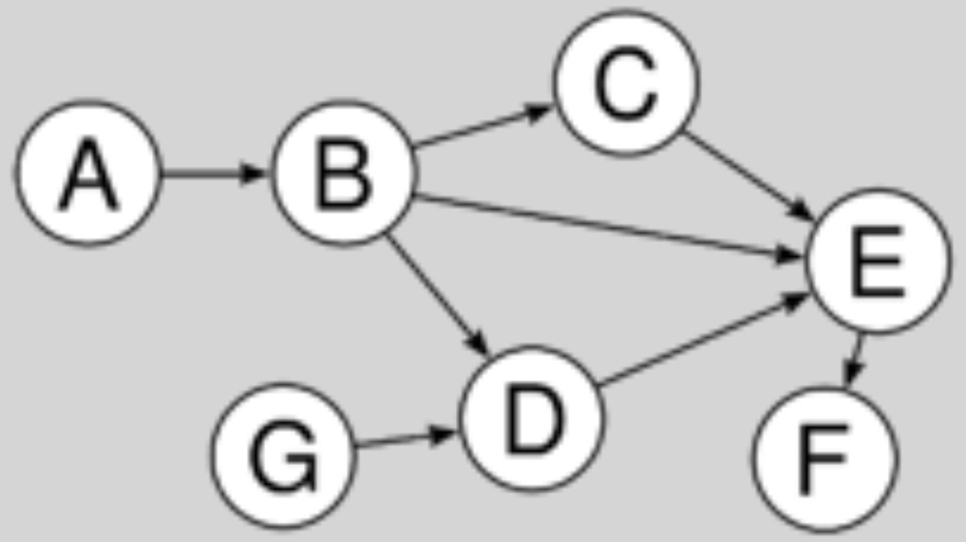
Take-Home Messages

microsimulation can advance traditional statistical modelling

microsimulation can:



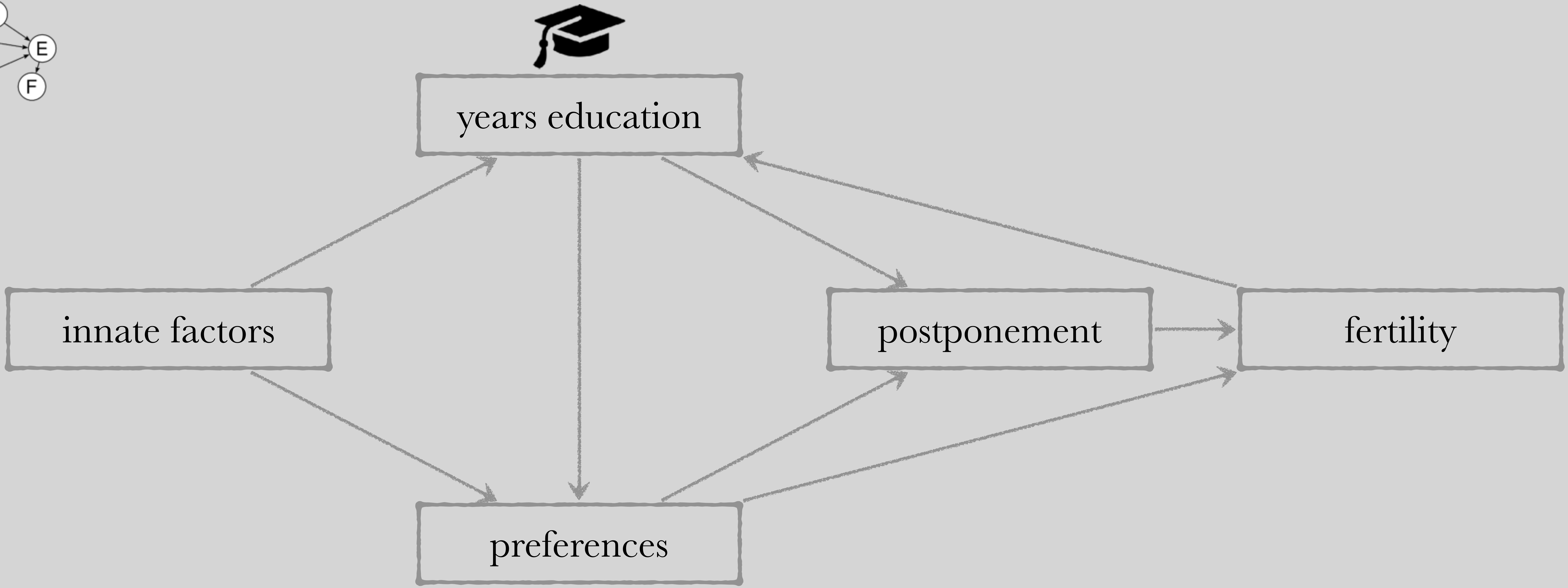
include biological information

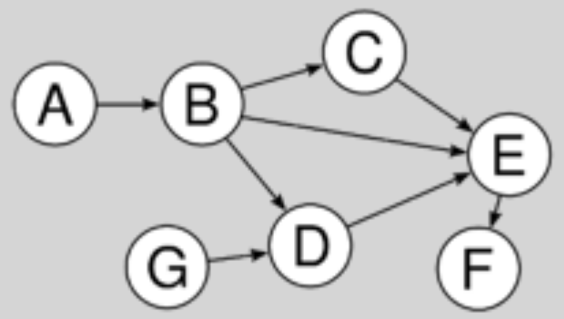


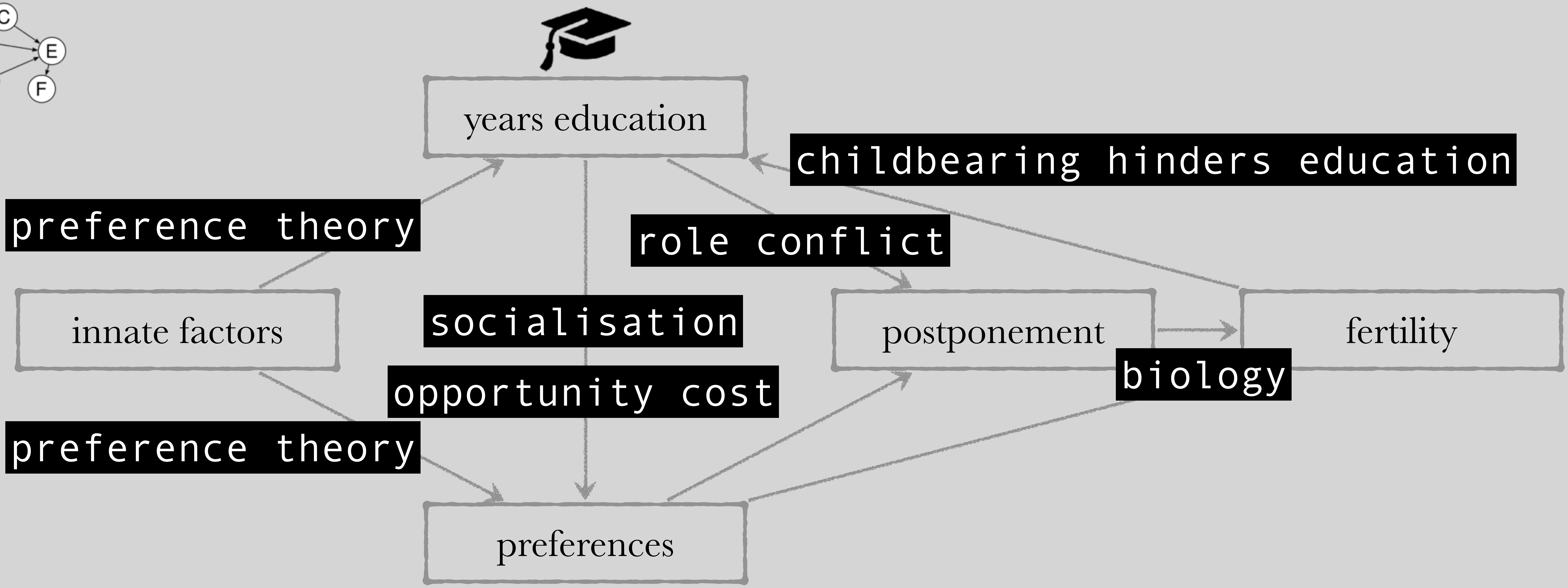
test (causal) mechanisms

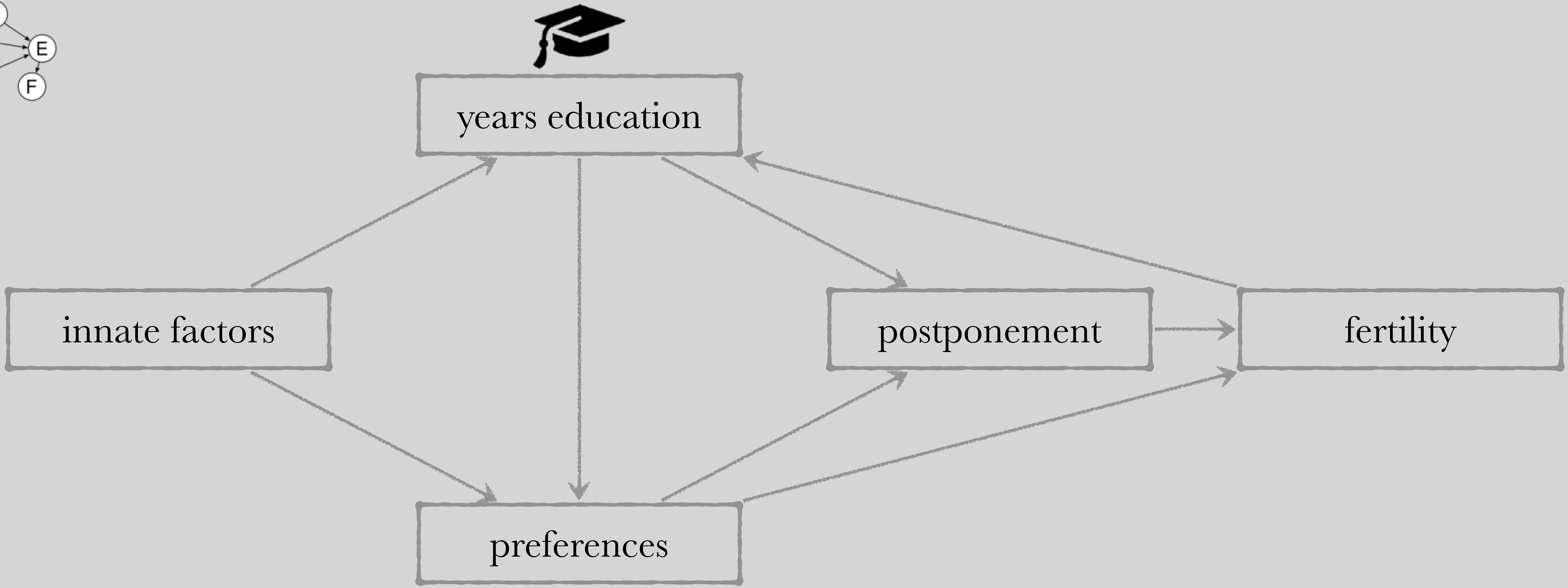
quantify unpredictability


50
0
% childfree
25%
7%
17%
24%
38%
I
II
III
IV
V
level of education
Generations & Gender Programme


years education
innate factors
postponement
fertility
preferences

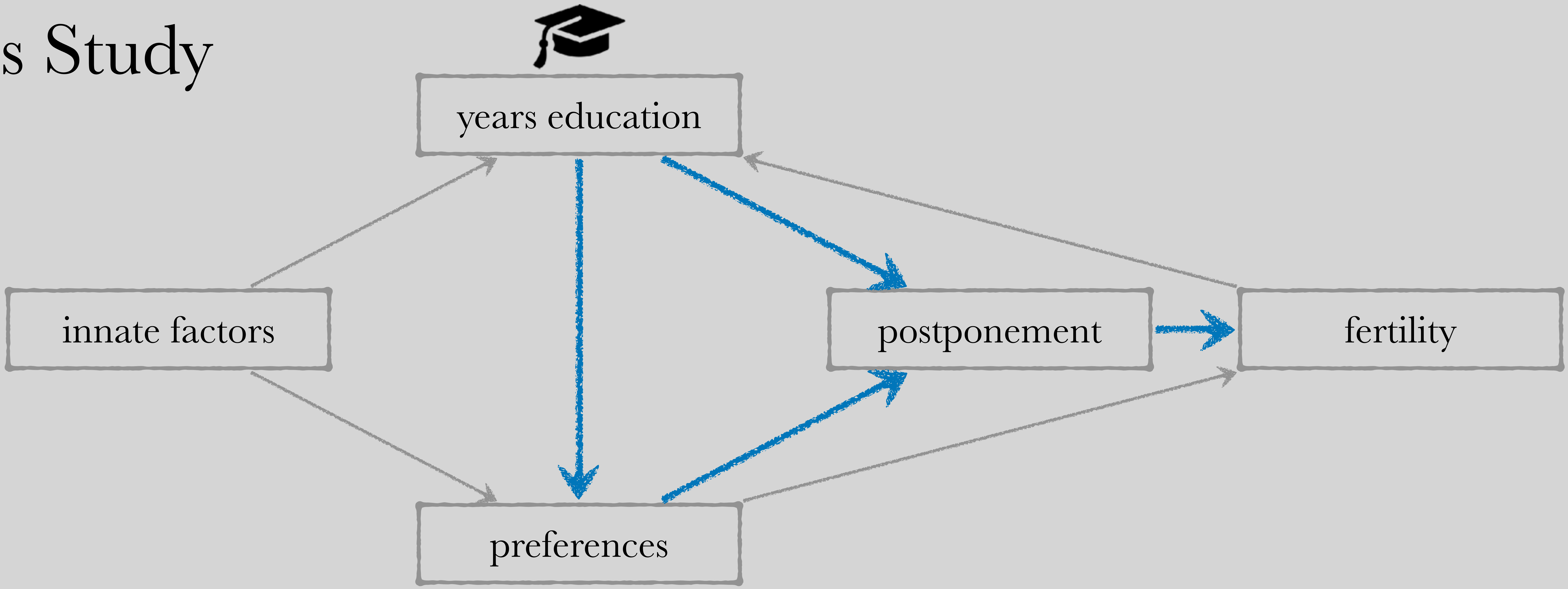

A
B
C
D
E
F
G


years education
childbearing hinders education
preference theory
role conflict
innate factors
socialisation
postponement
fertility
biology
opportunity cost
preference theory
preferences



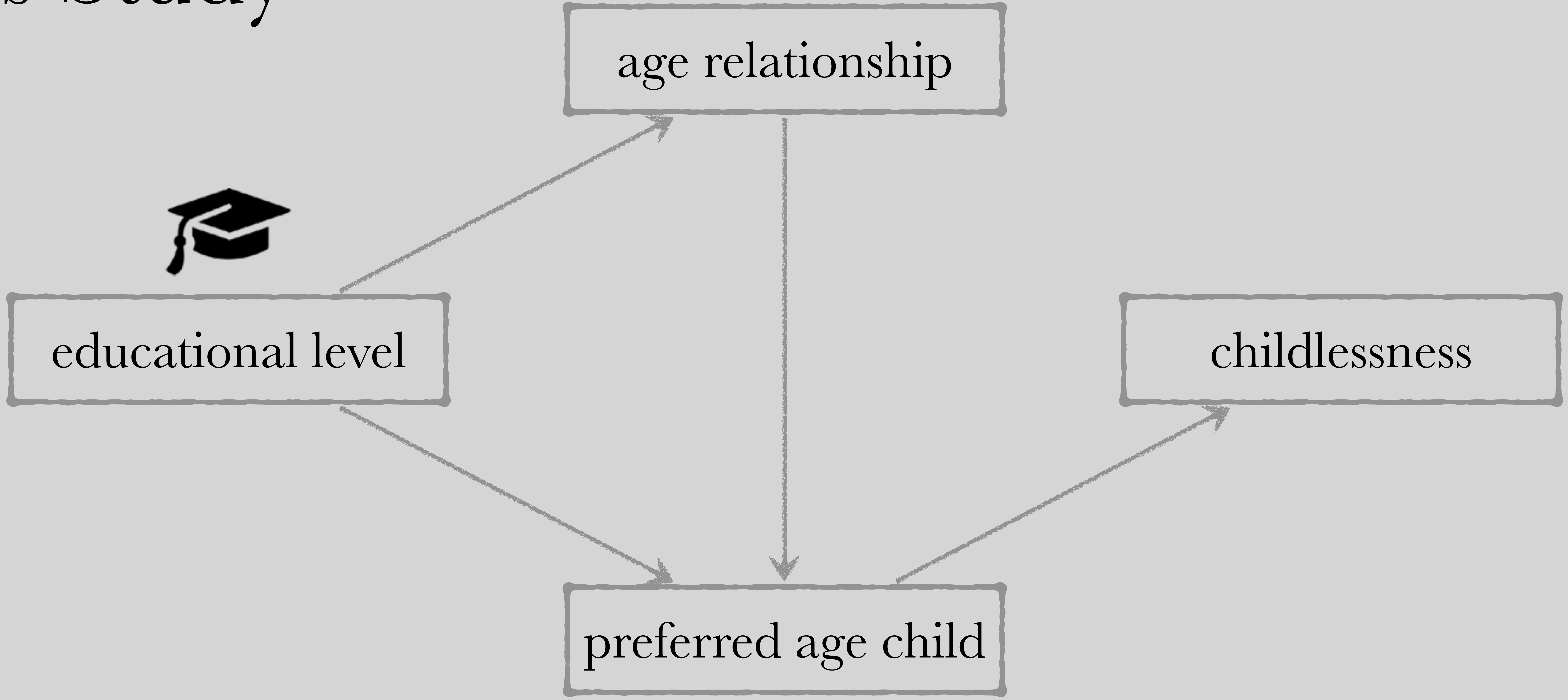
What Kind of Data Would We need to Address This Model?

This Study



years education
innate factors
postponement
fertility
preferences

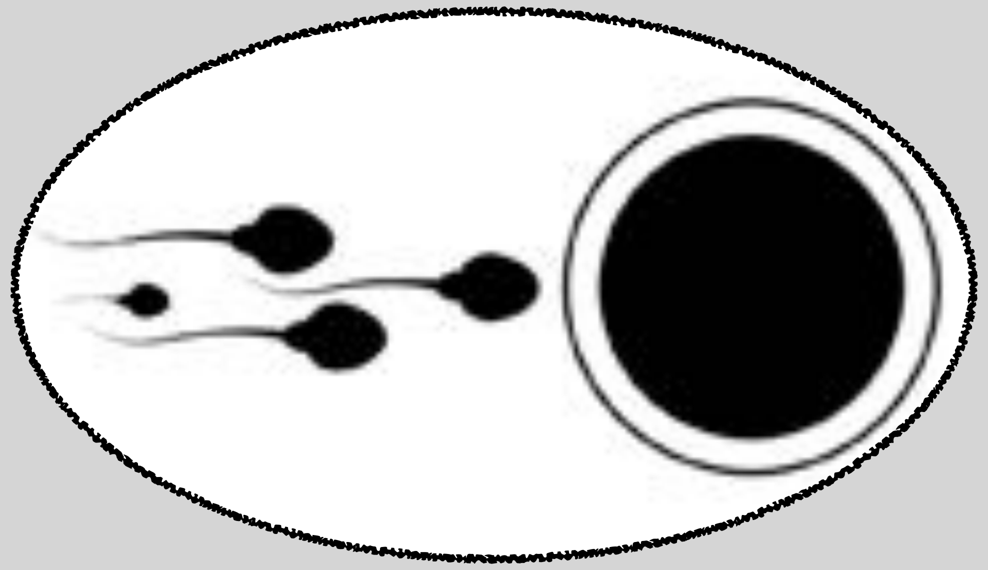
This Study


age relationship
educational level
childlessness
preferred age child

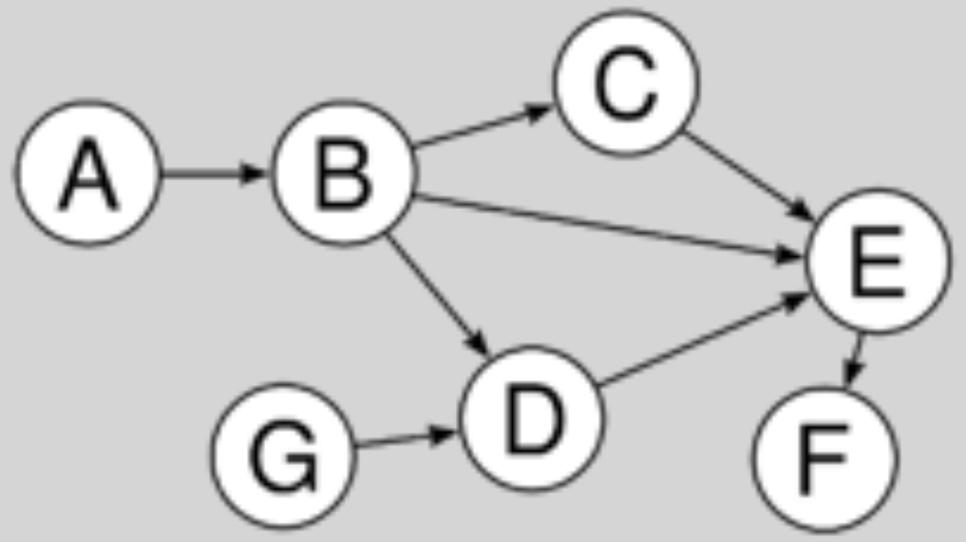
Take-Home Messages

microsimulation can advance traditional statistical modelling

microsimulation can:

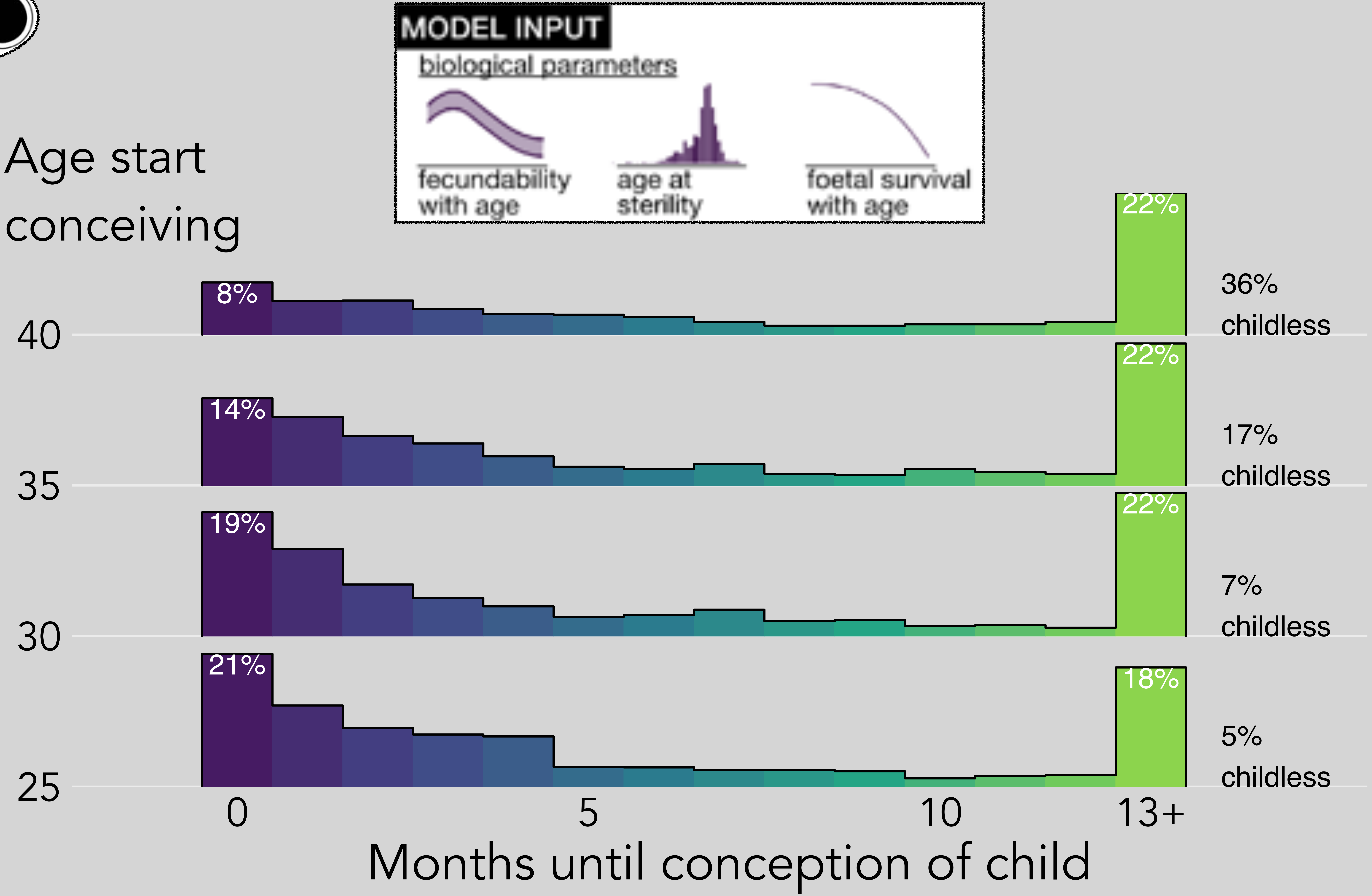


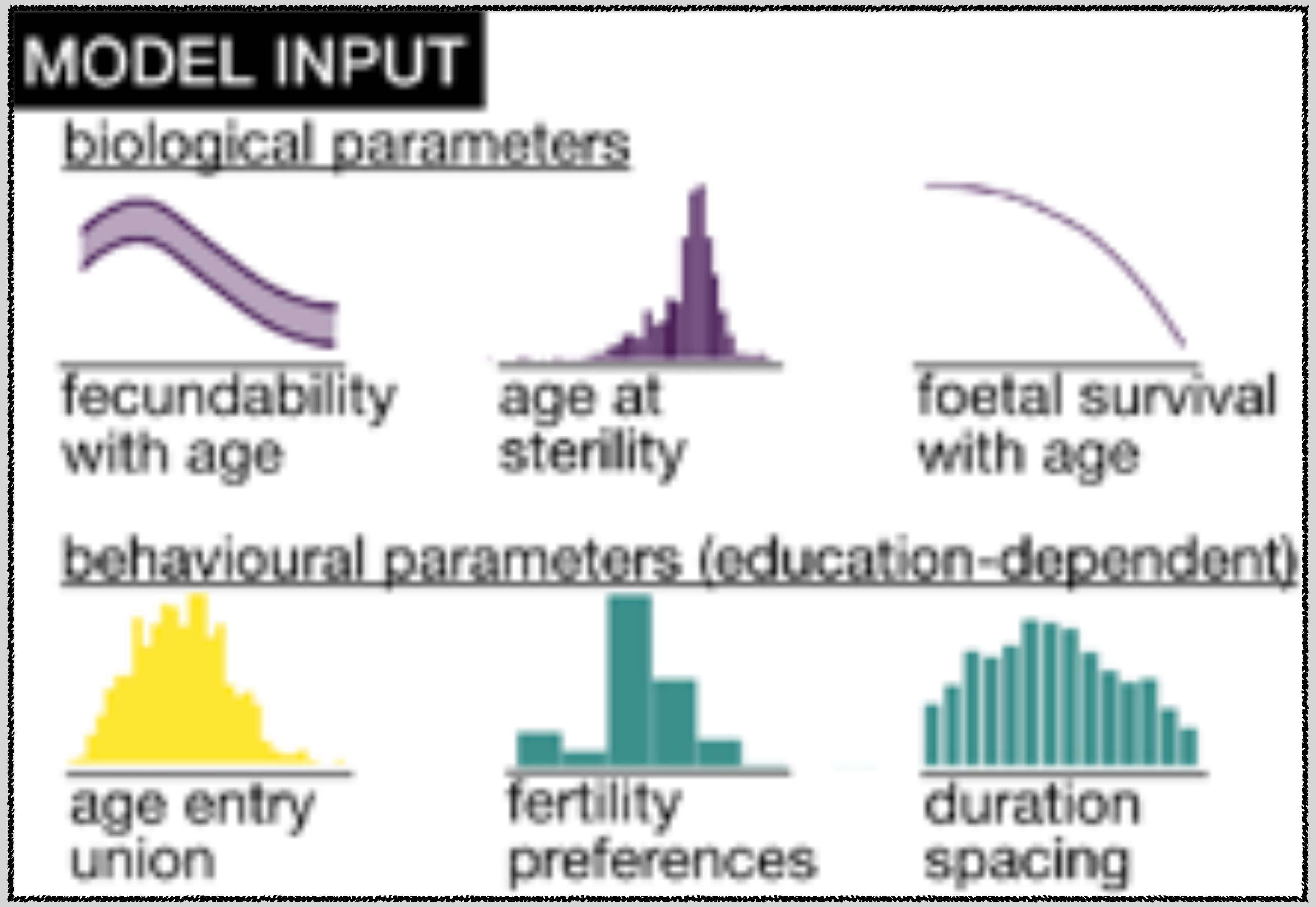
include biological information



test (causal) mechanisms

quantify unpredictability

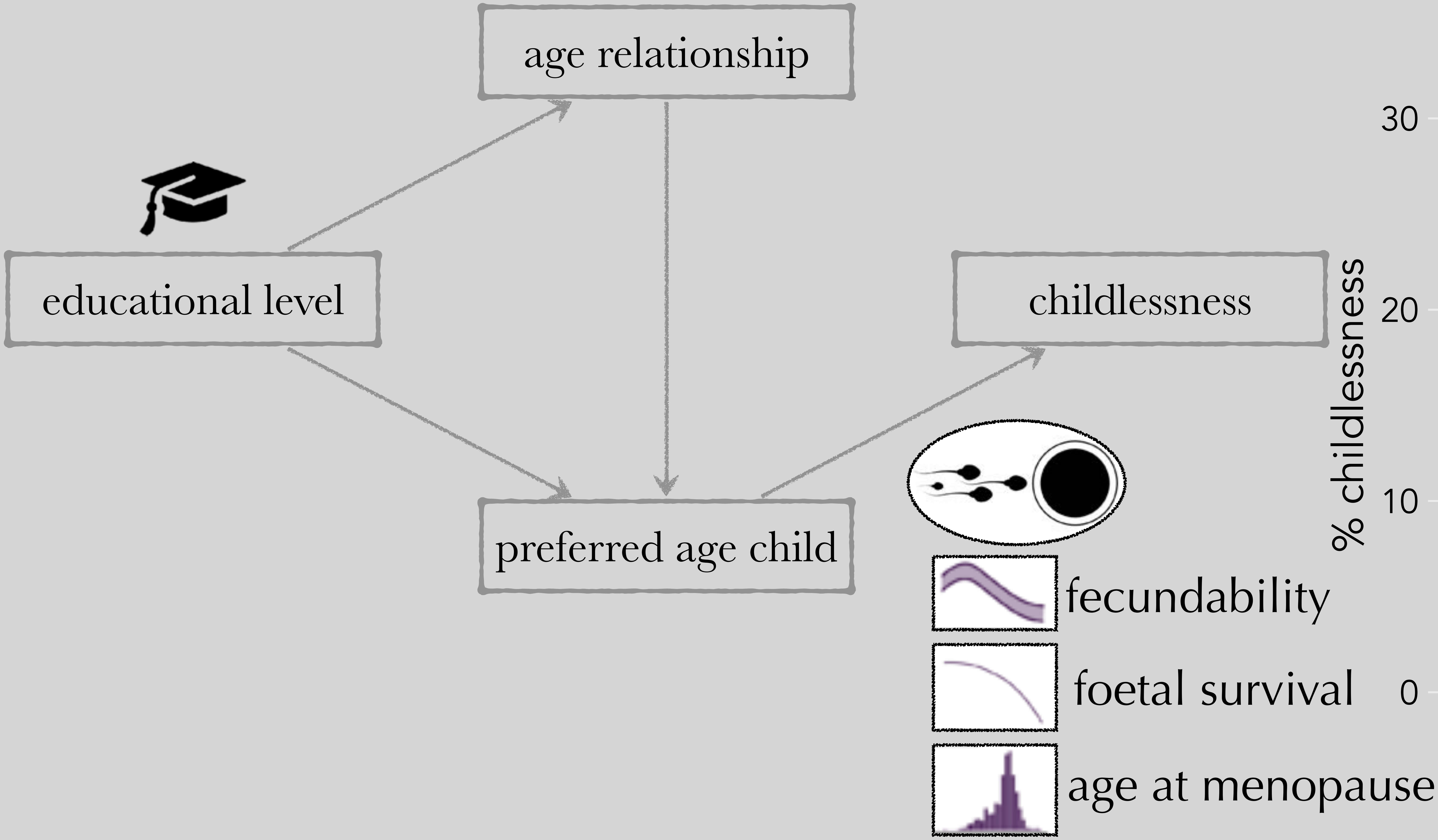

MODEL INPUT
biological parameters
fecundability with age
age at sterility
foetal survival with age
Age start conceiving
40
8%
22%
36% childless
35
14%
22%
17% childless
30
19%
22%
7% childless
25
21%
18%
5% childless
0
5
10
13+
Months until conception of child

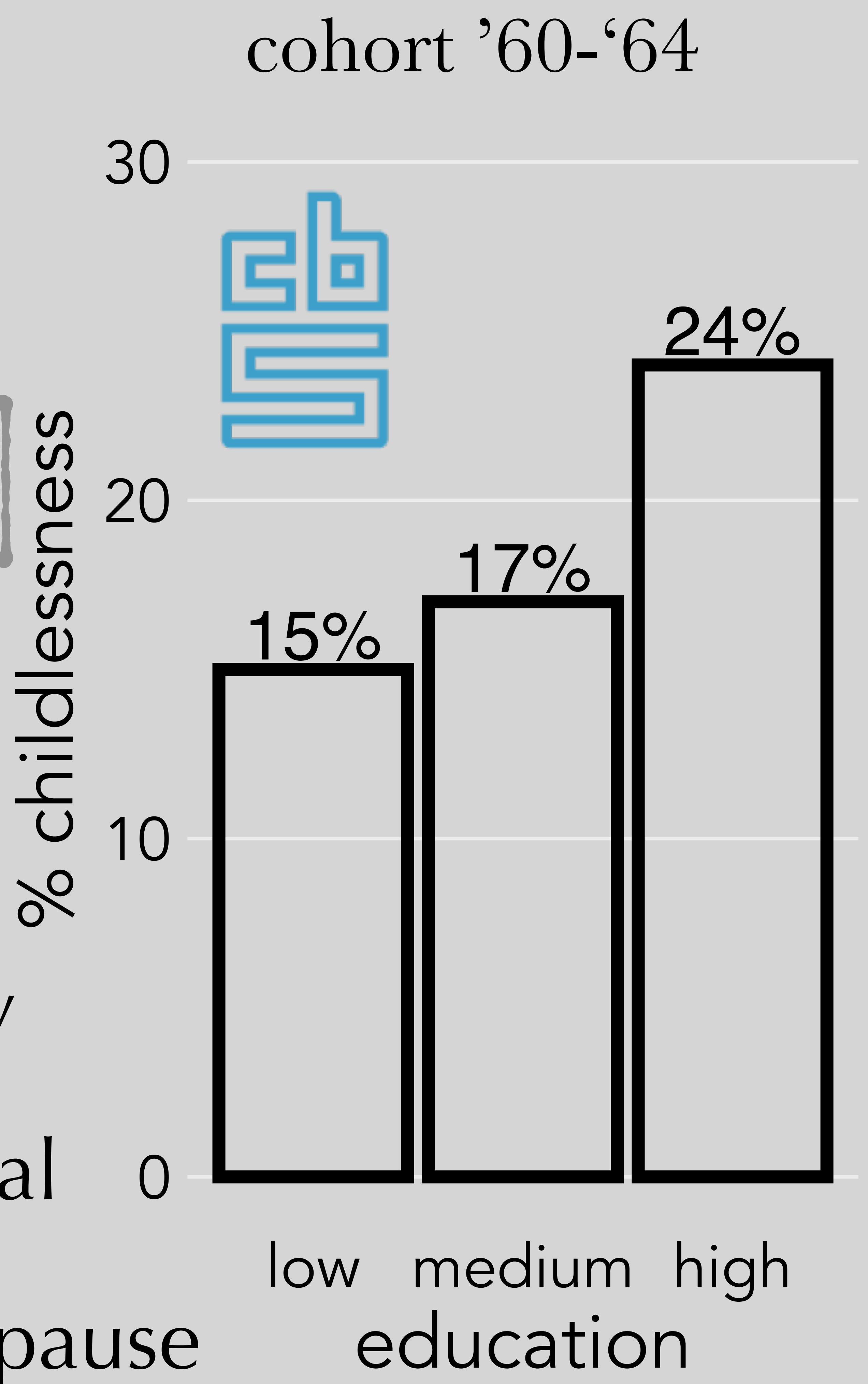


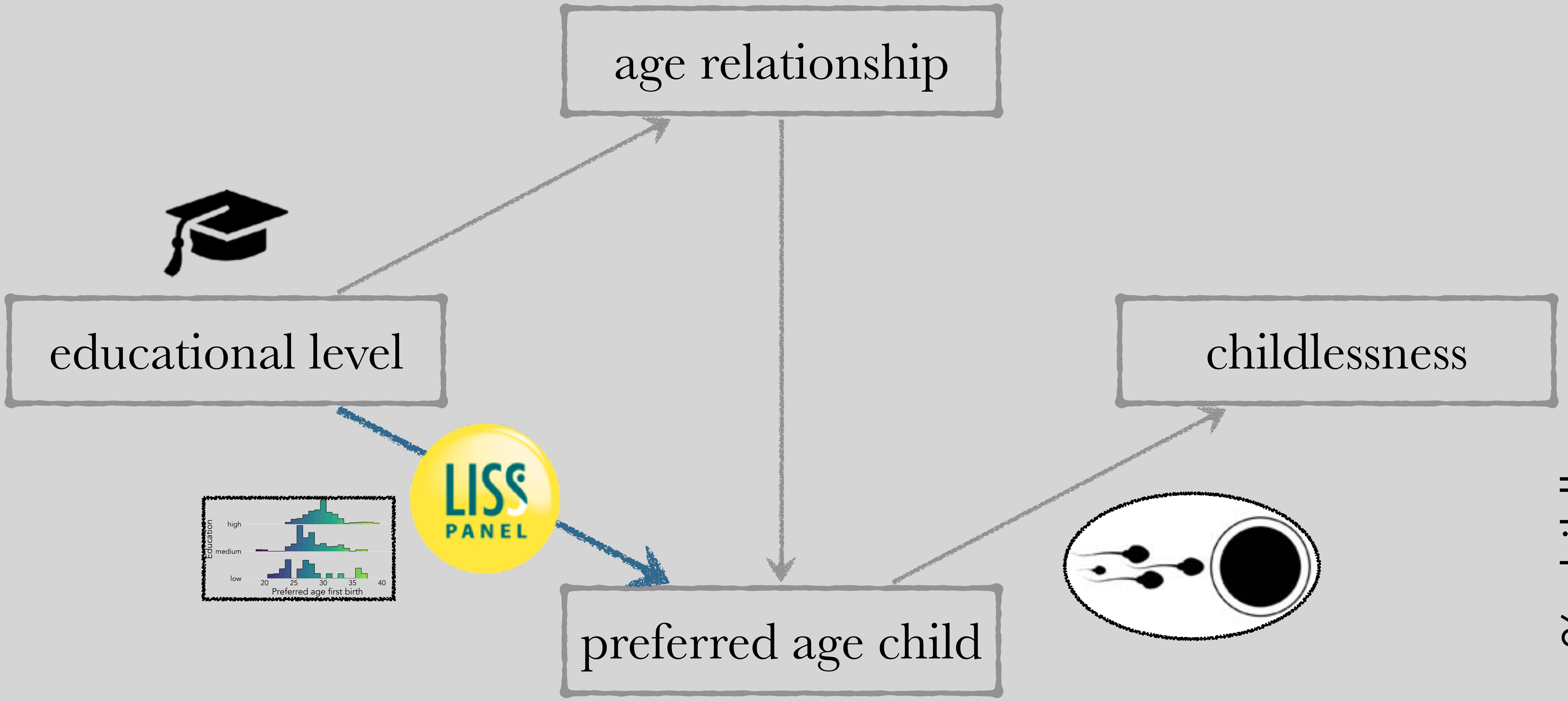
determines whether and when people would like to conceive

determines whether and when people conceive

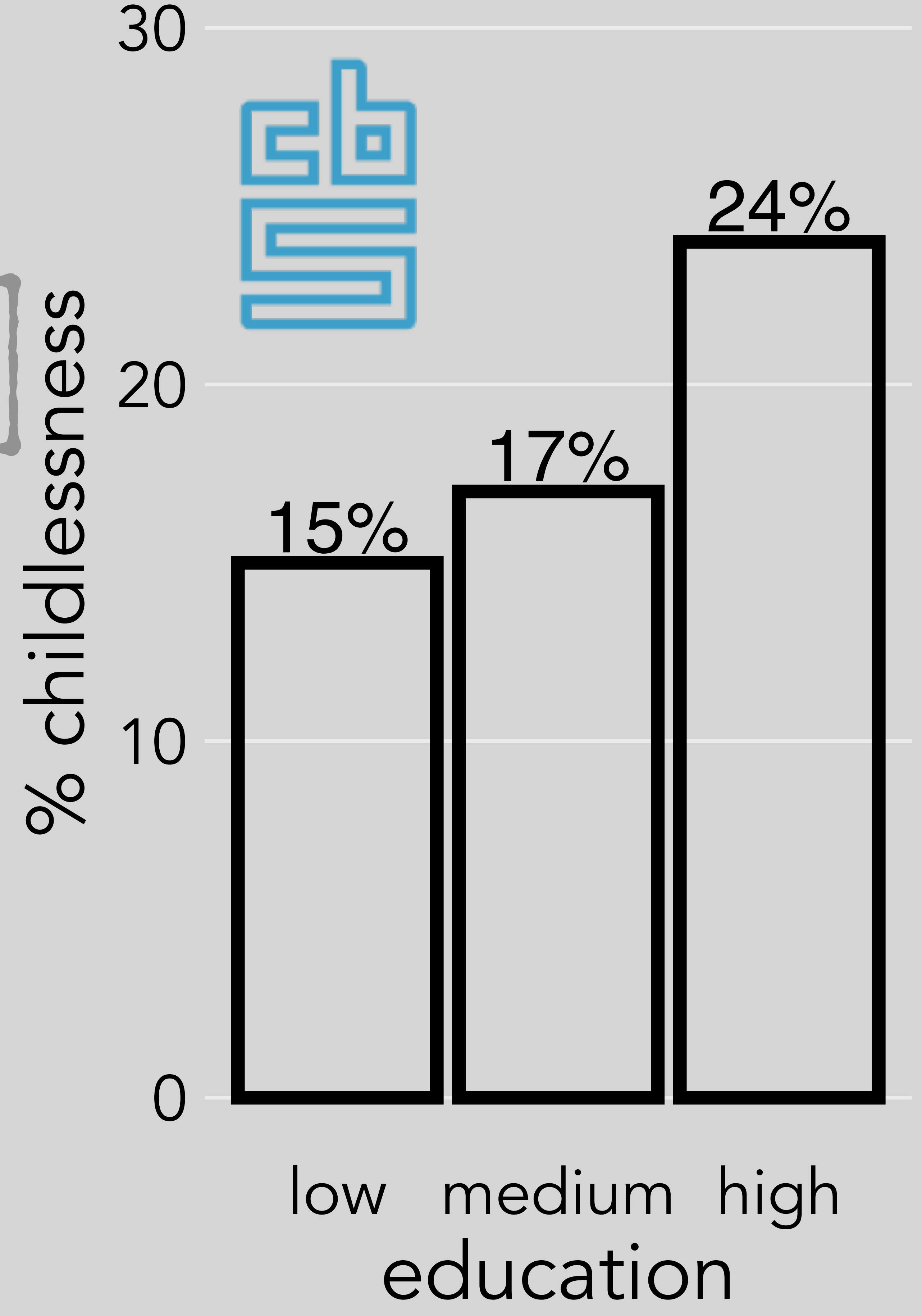

MODEL INPUT
biological parameters
fecundability with age
age at sterility
foetal survival with age
behavioural parameters (education-dependent)
age entry union
fertility preferences
duration spacing
MODEL RUN
Randomly determined traits individual 1
in union =22 | spac. =5 | pref. =2 | fecund. =0.3 | steril. =43 | edu. =high
stochasticity
stochasticity
22 union
27 start conceiving
29 birth
34 start conceiving
38 miscarriage
43 sterility
Randomly determined traits individual 10000
in union =26 | spac. =1 | pref. =1 | fecund. =0.1 | steril. =45 | edu. =low
stochasticity
26 union
27 start conceiving
31 birth
no further children preferred
MODEL OUTPUT
due to:
partner status
preferences
stochasticity
% childless
low high
education
frequency
age first birth
frequency
children
'TRUTH'
% childless
low high
education
frequency
age first birth
frequency
children


age relationship
educational level
childlessness
preferred age child
fecundability
foetal survival
age at menopause

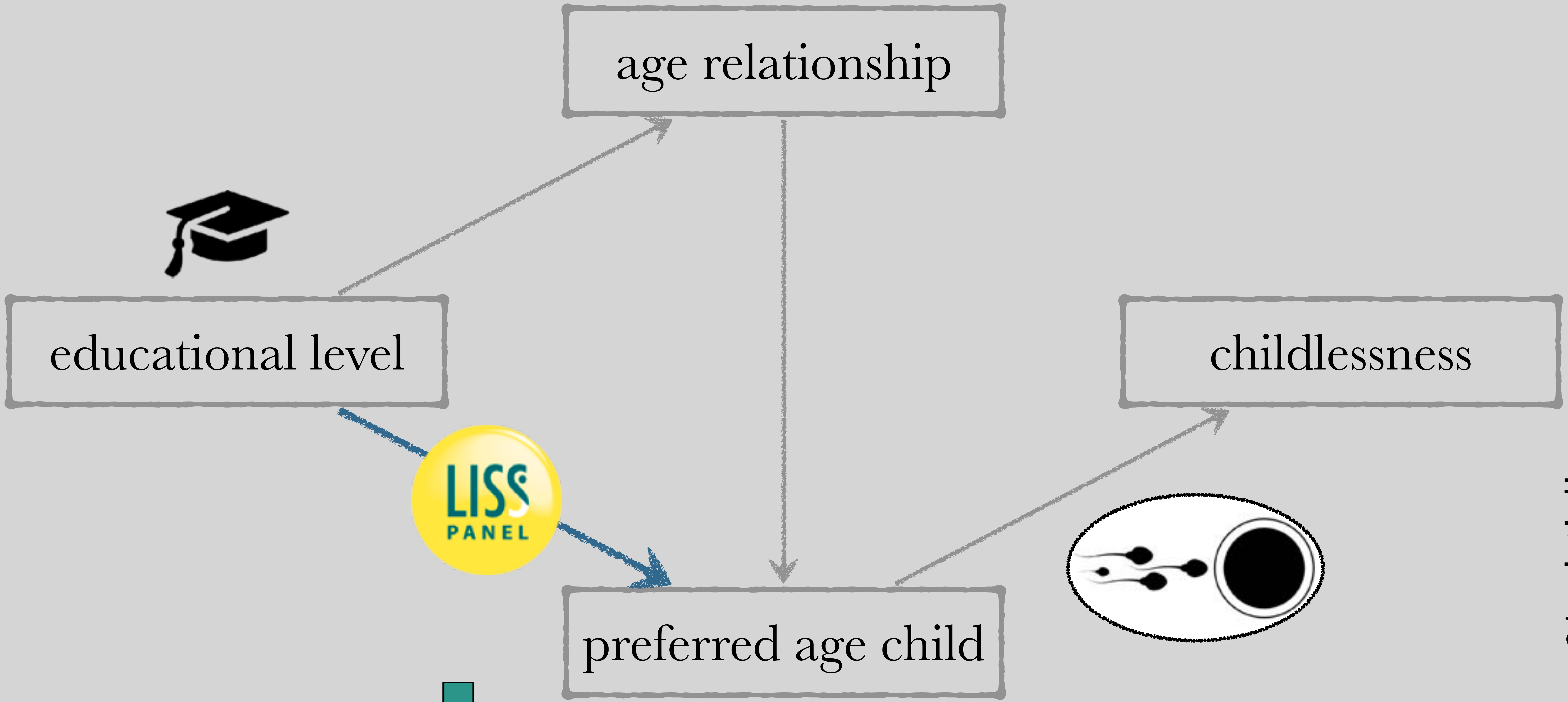

cohort ’60-‘64
30
20
10
0
% childlessness
15%
17%
24%
low
medium
high
education

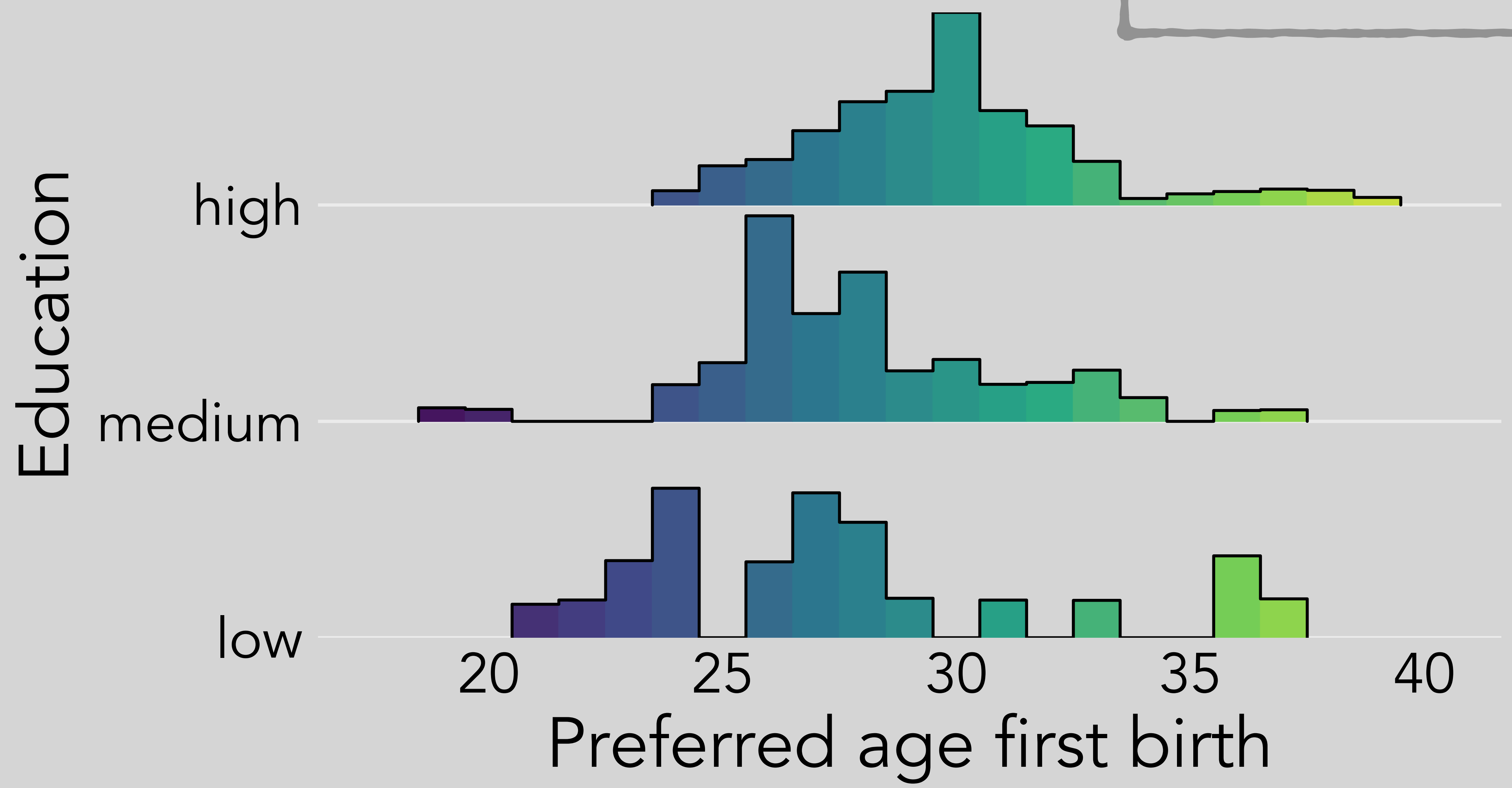


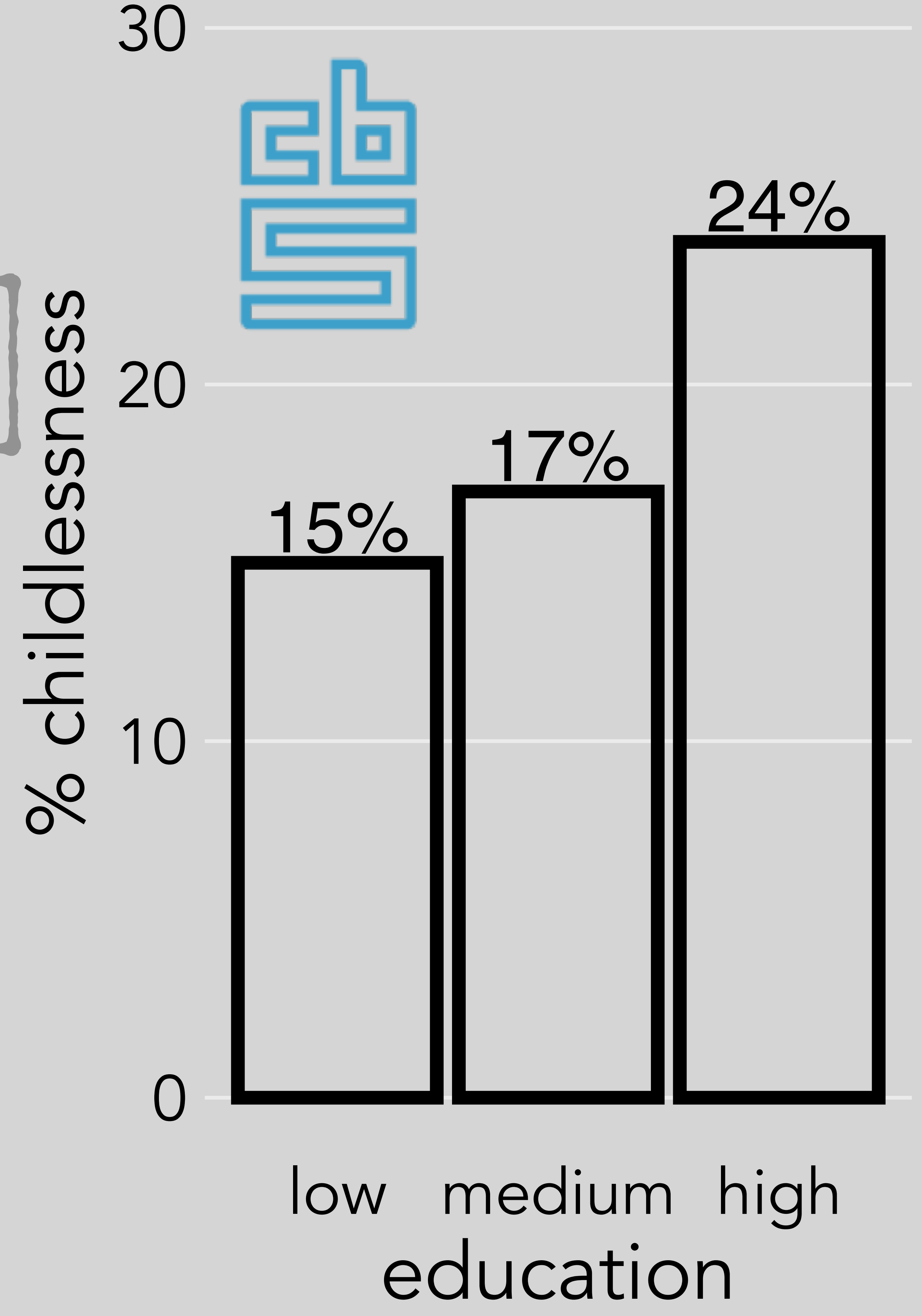
age relationship
educational level
childlessness
preferred age child
LISS PANEL
high
medium
low
Education
20 25 30 35 40
Preferred age first birth

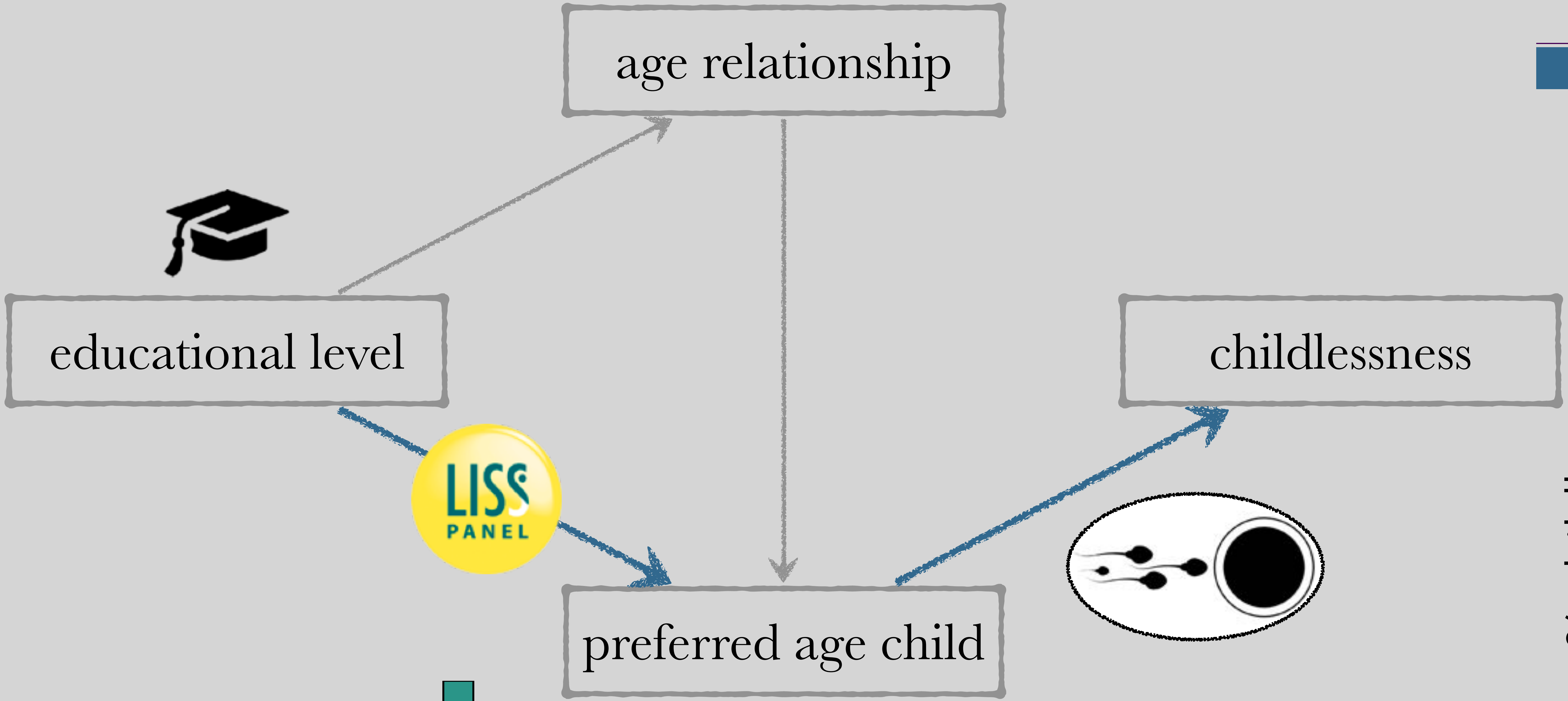


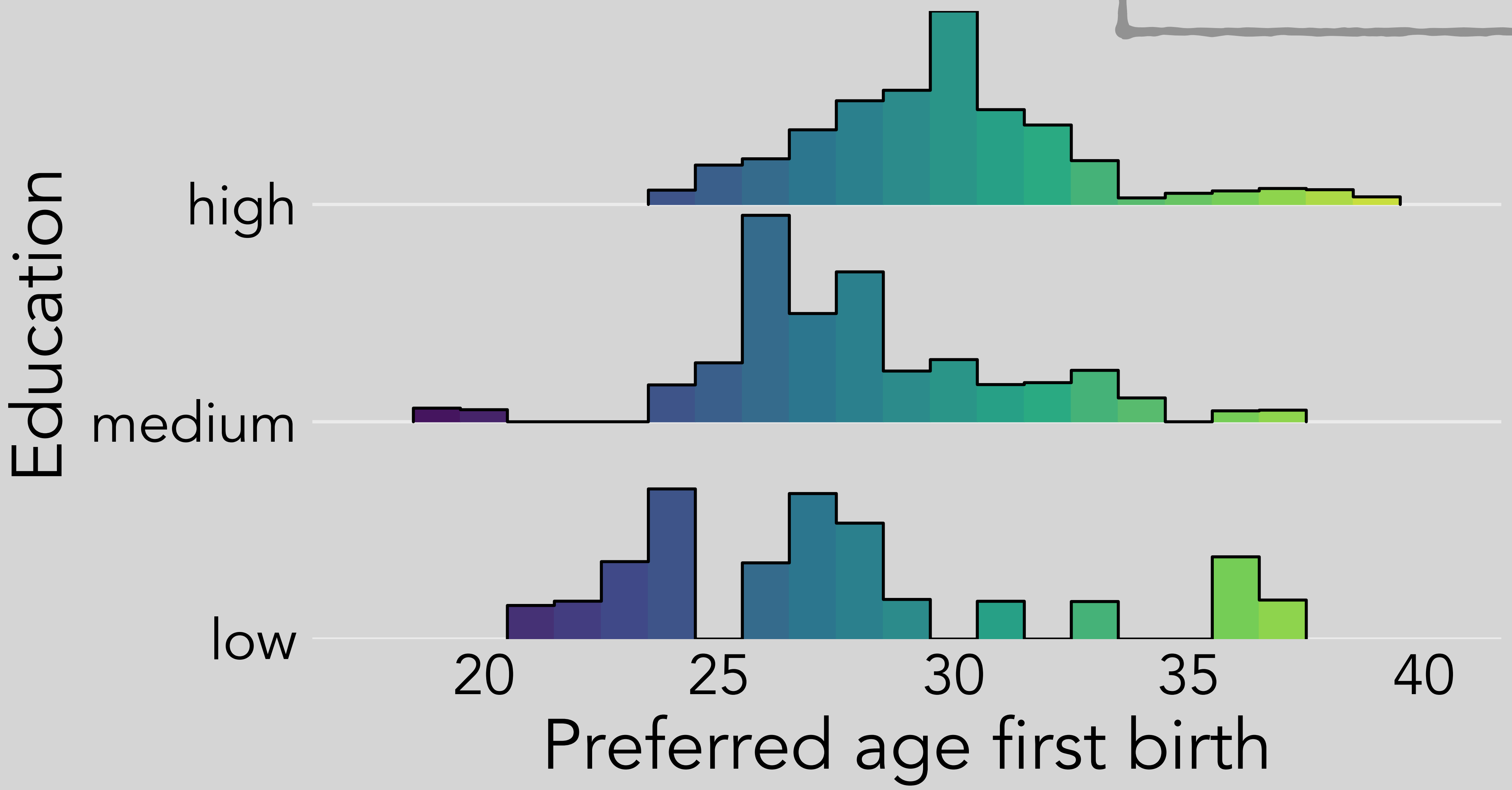
30
20
10
0
% childlessness
15%
17%
24%
low
medium
high
education

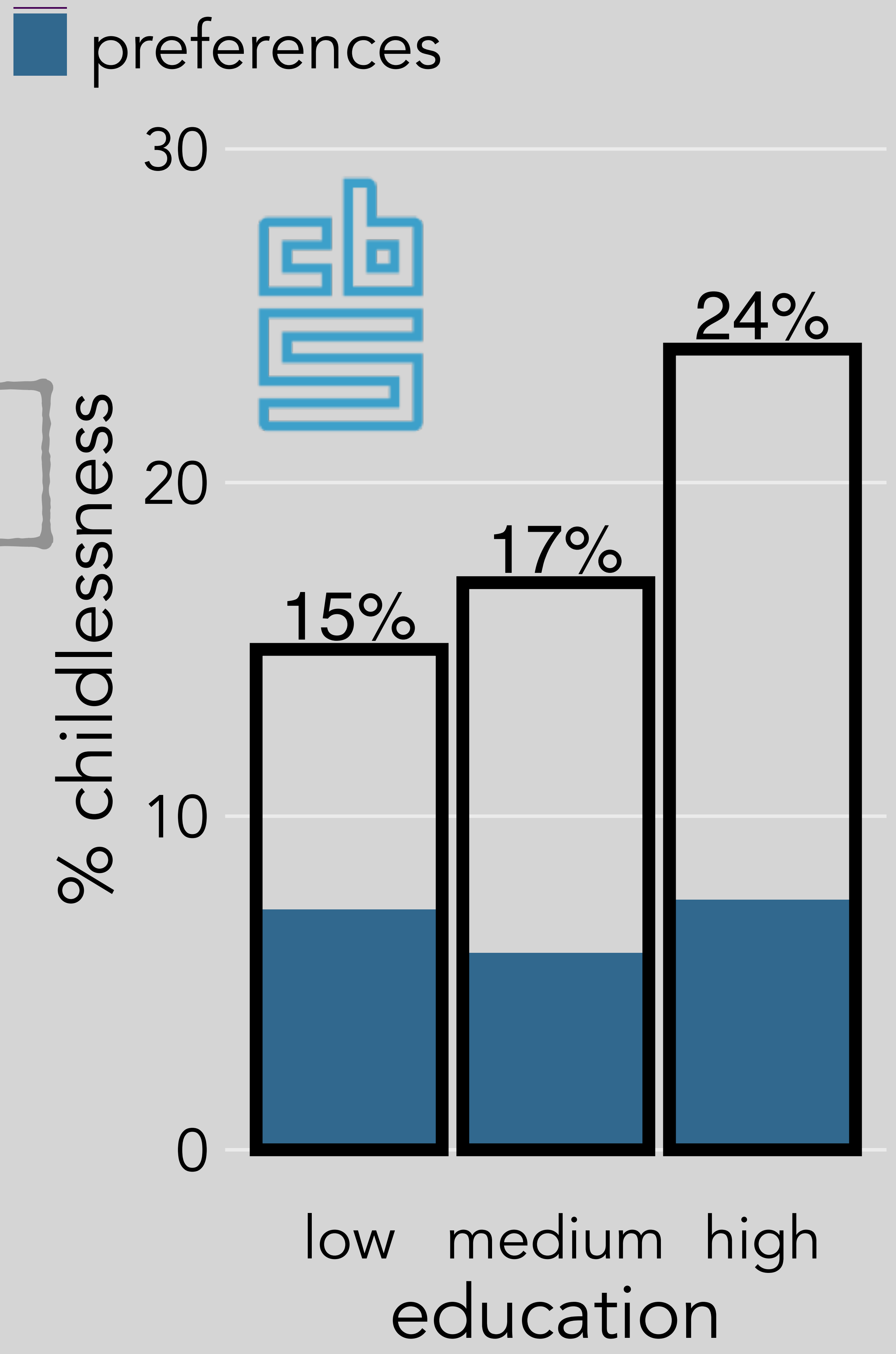

age relationship
educational level
childlessness
preferred age child
LISS
PANEL

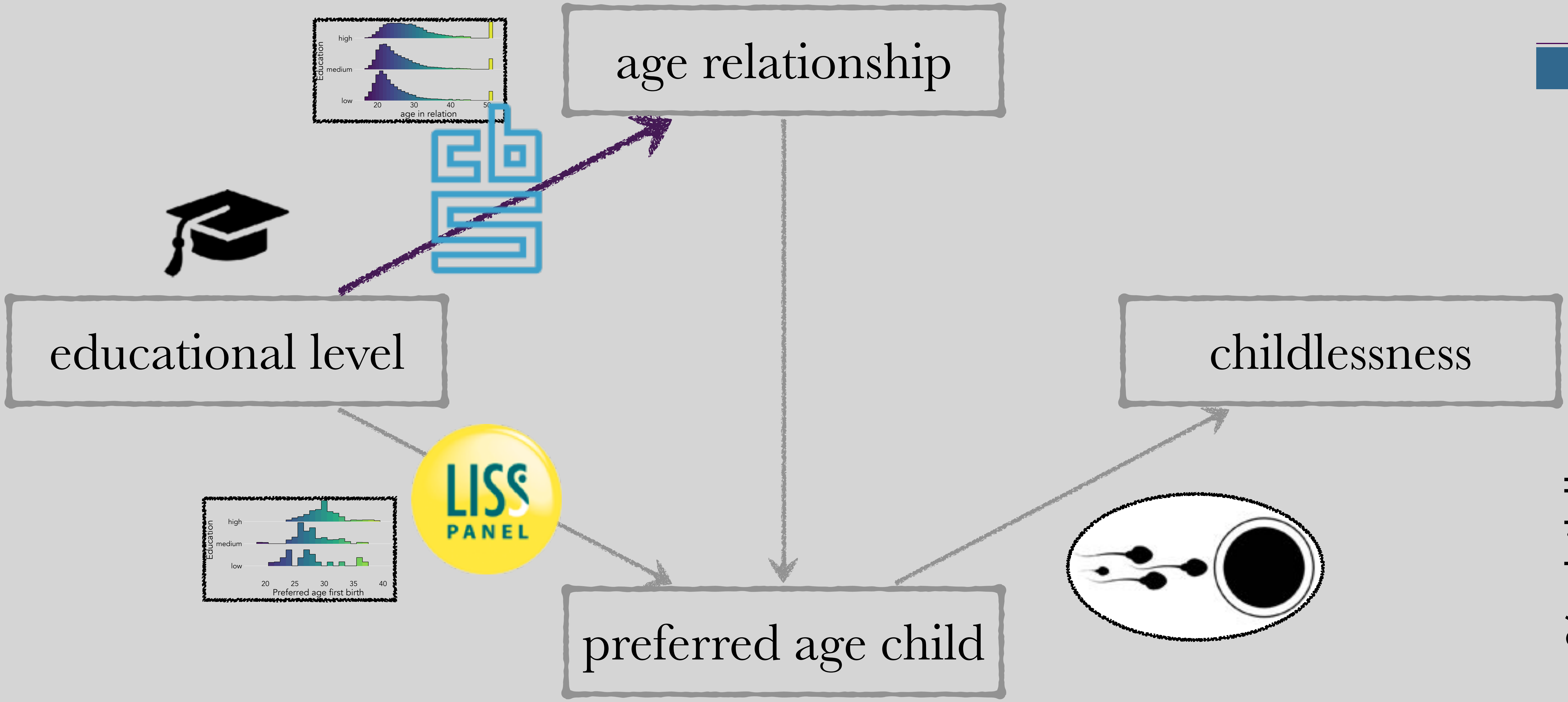

Education
high
medium
low
20
25
30
35
40
Preferred age first birth

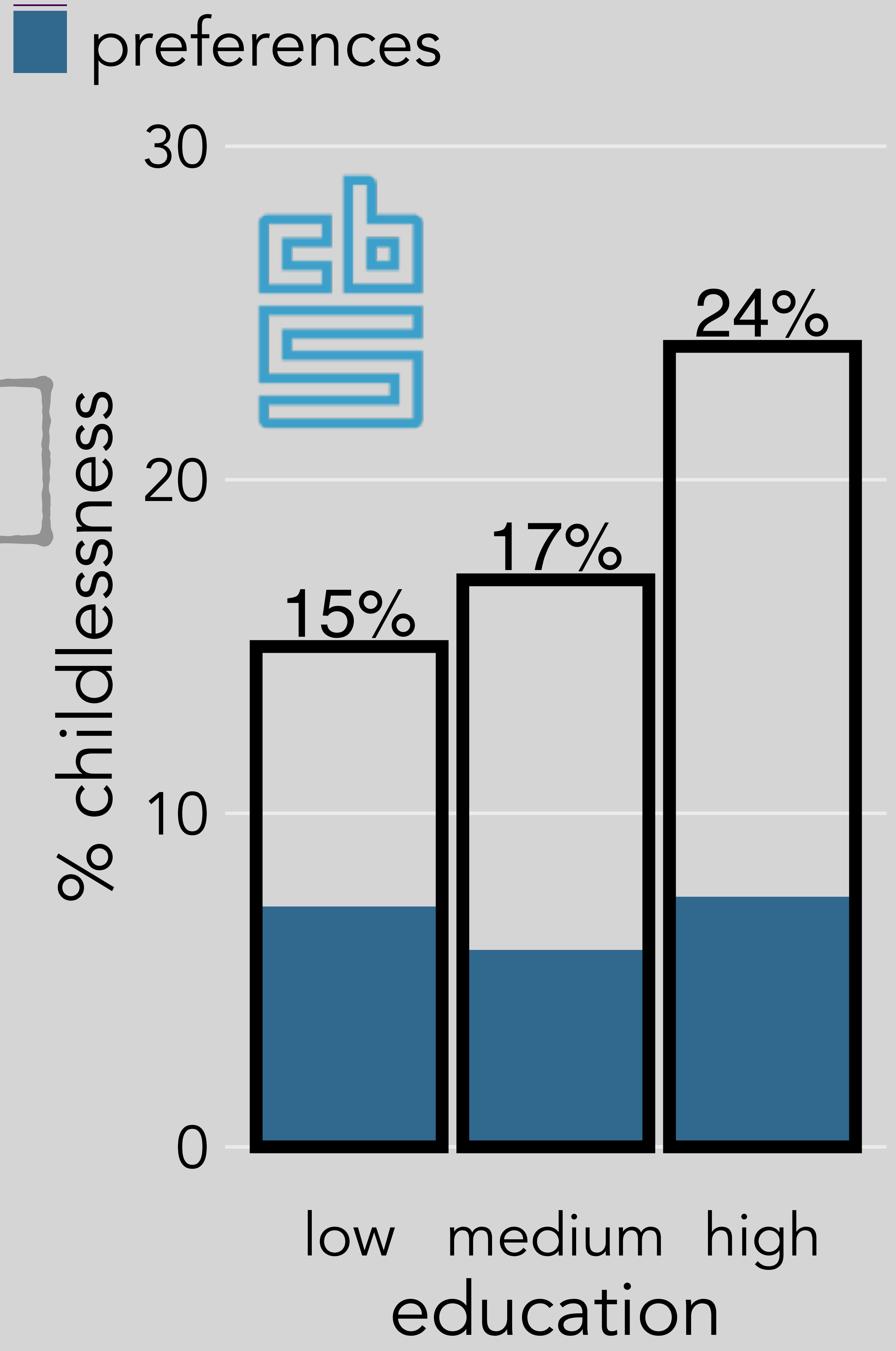

% childlessness
30
20
10
0
15%
17%
24%
low
medium
high
education


age relationship
educational level
childlessness
preferred age child
LISS
PANEL


Education
high
medium
low
20
25
30
35
40
Preferred age first birth

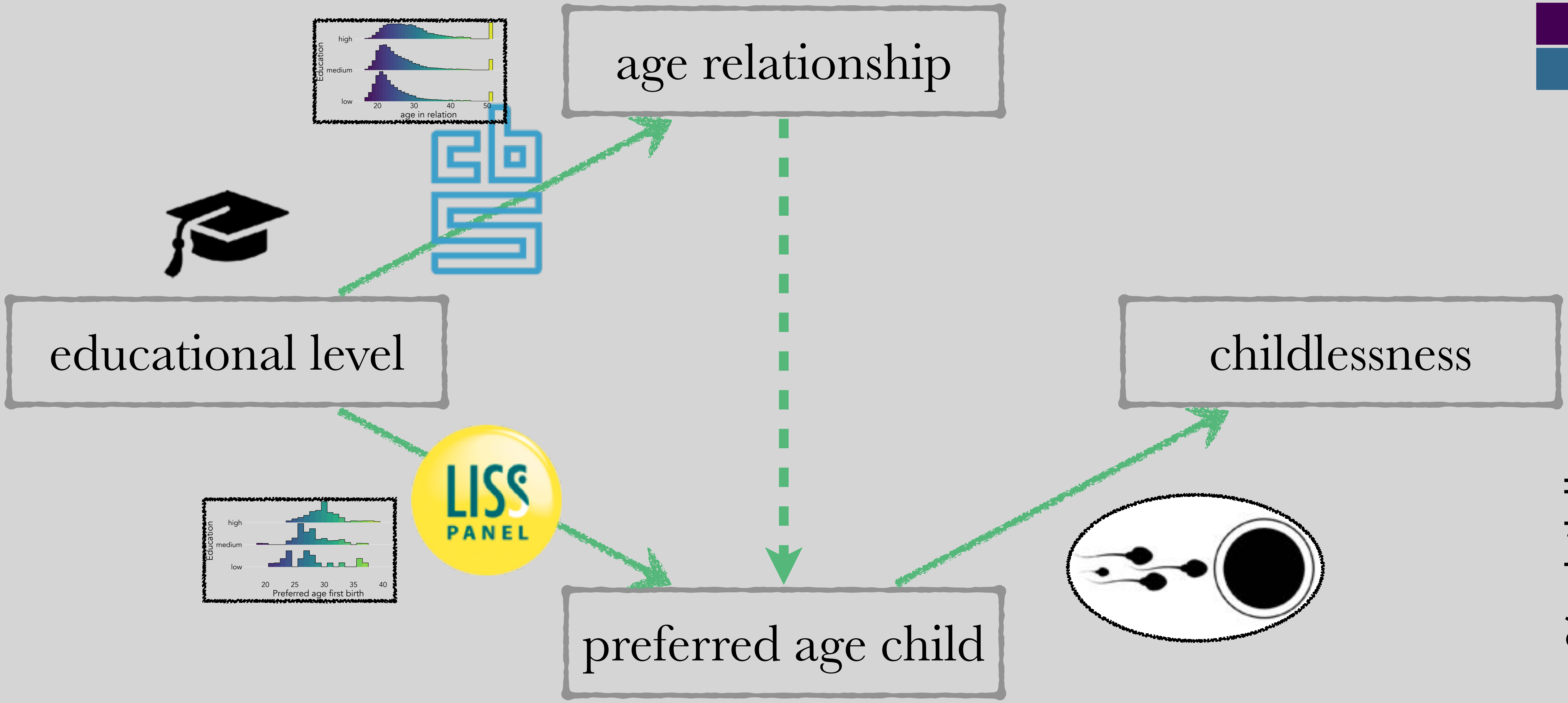

preferences
% childlessness
30
20
10
0
15%
17%
24%
low
medium
high
education

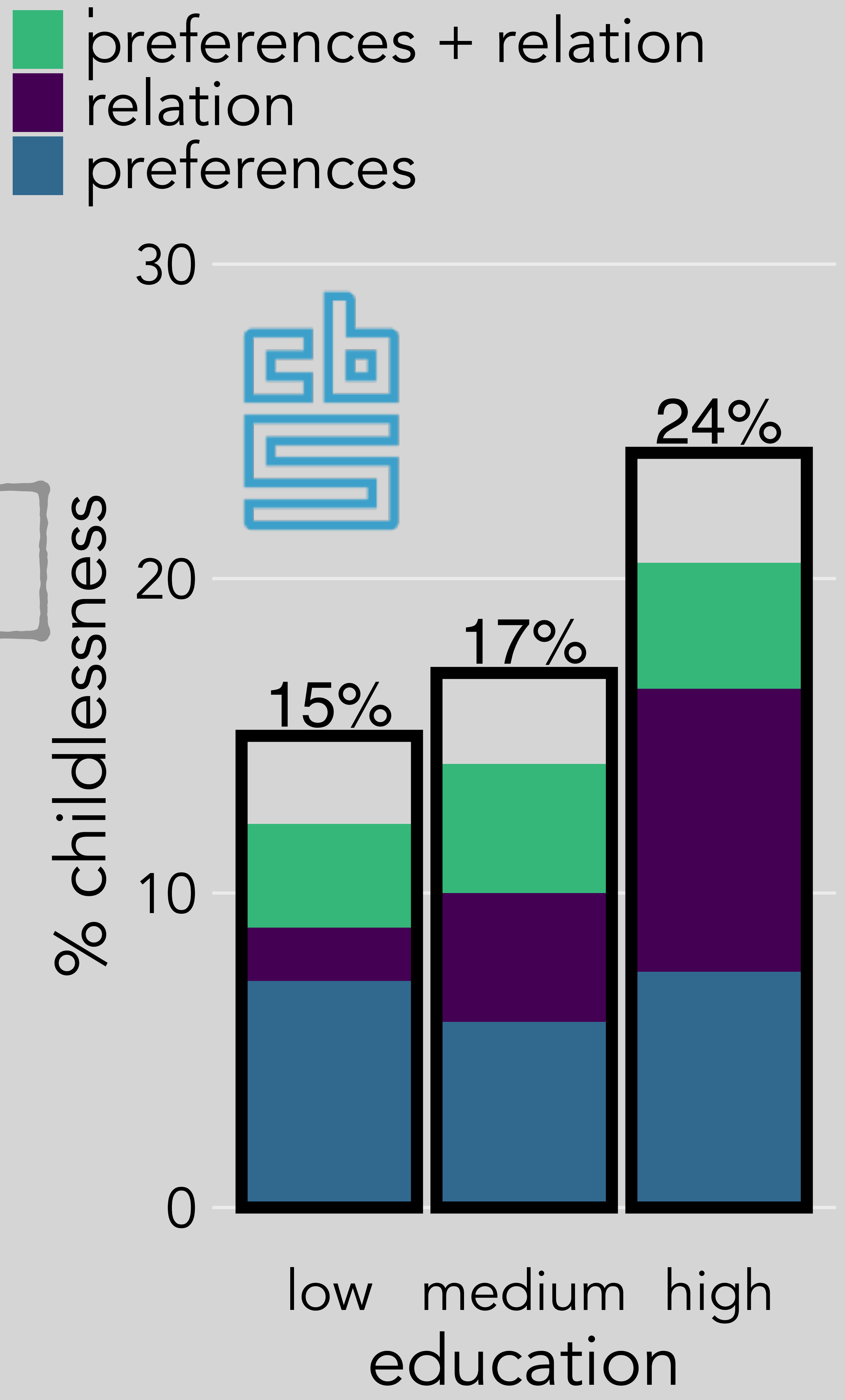

age relationship
educational level
childlessness
preferred age child

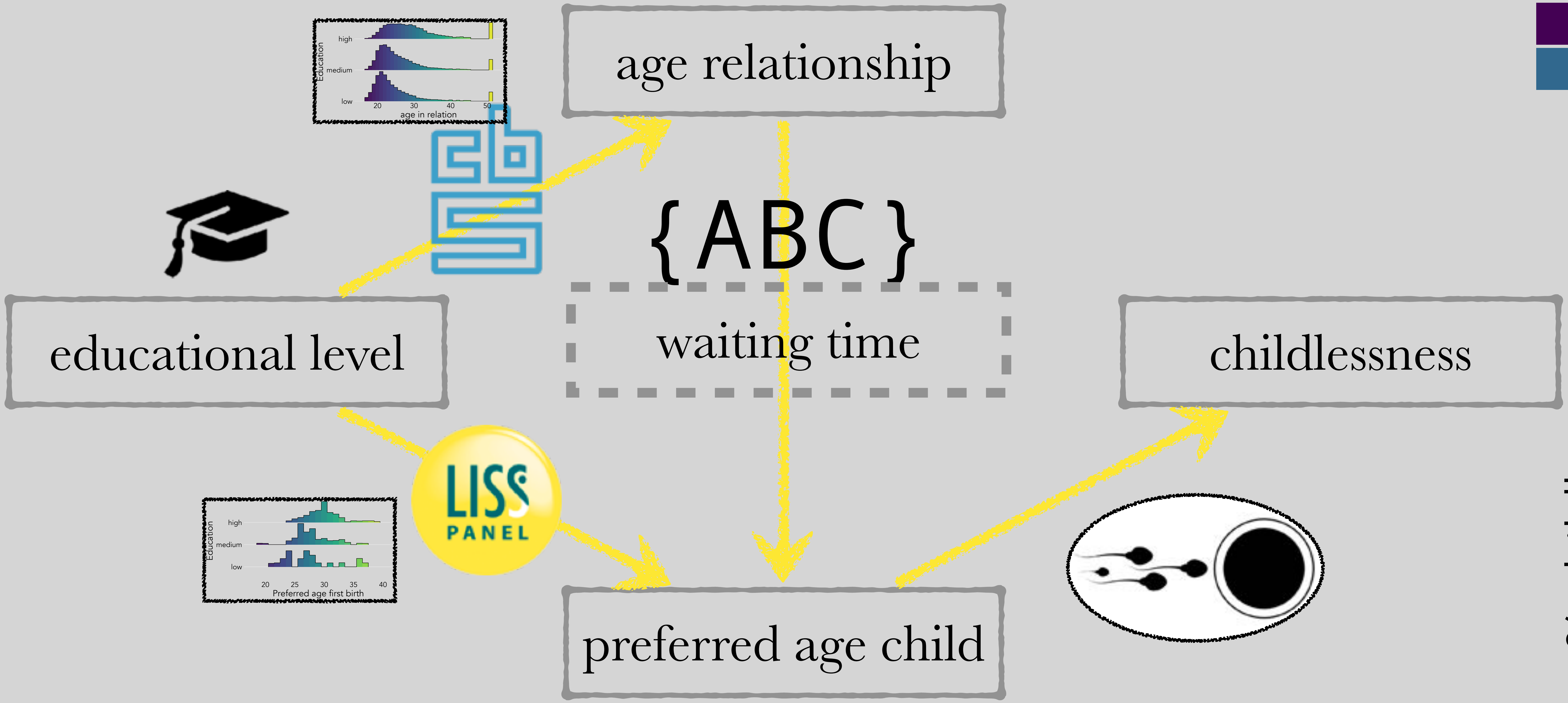

preferences
30
20
10
0
% childlessness
15%
17%
24%
low
medium
high
education

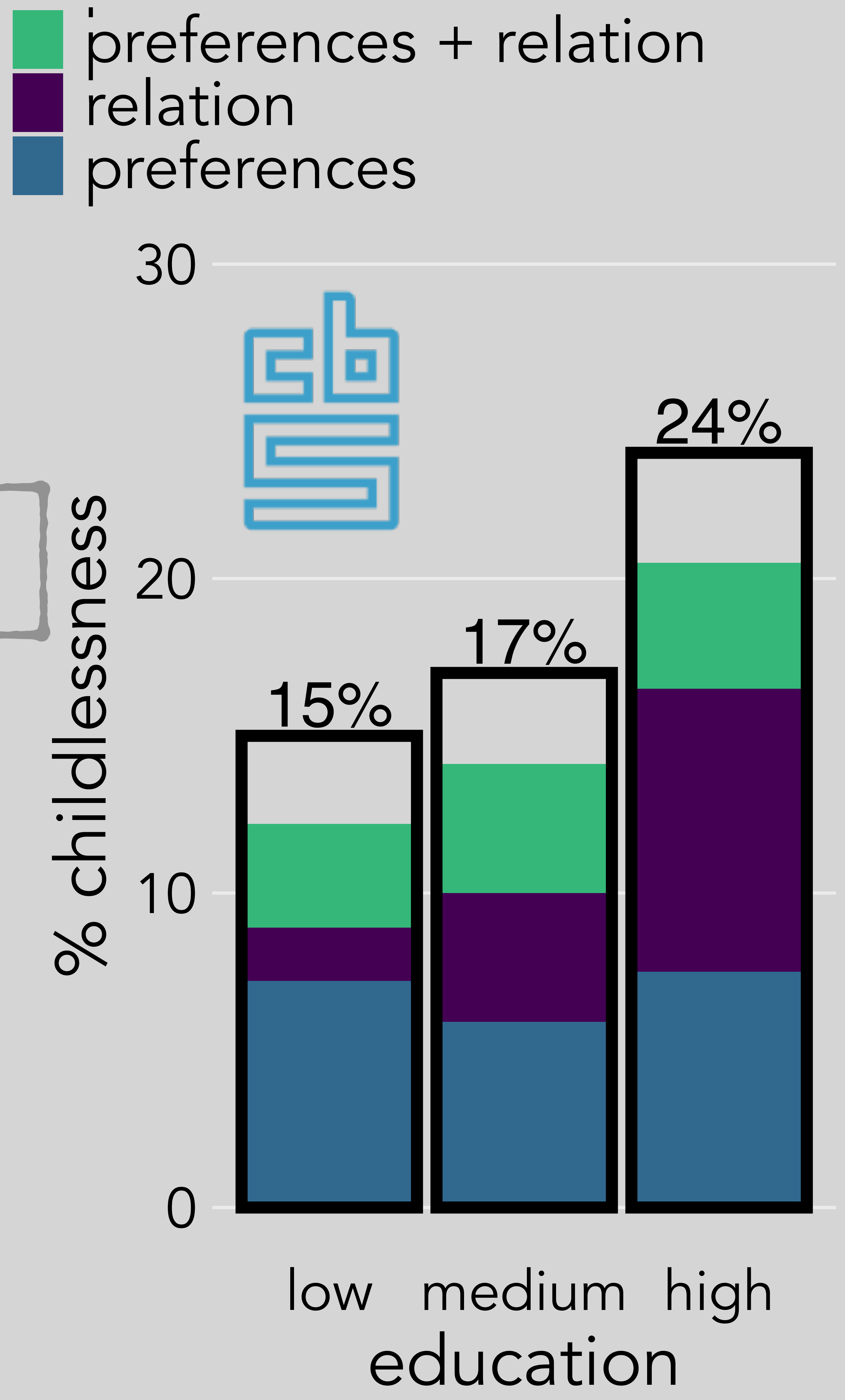

age relationship
educational level
childlessness
preferred age child
LISS
PANEL
high
medium
low
Education
20
30
40
50
age in relation
preferences
30
20
10
0
% childlessness
15%
17%
24%
low
medium
high
education

age relationship
educational level
childlessness
preferred age child
LISS PANEL
high
medium
low
Education
20
30
40
50
age in relation
relation
preferences
30
20
10
0
% childlessness
15%
17%
24%
low
medium
high
education


age relationship
high
medium
low
Education
age in relation
educational level
high
medium
low
Education
Preferred age first birth
LISS
PANEL
preferred age child
childlessness

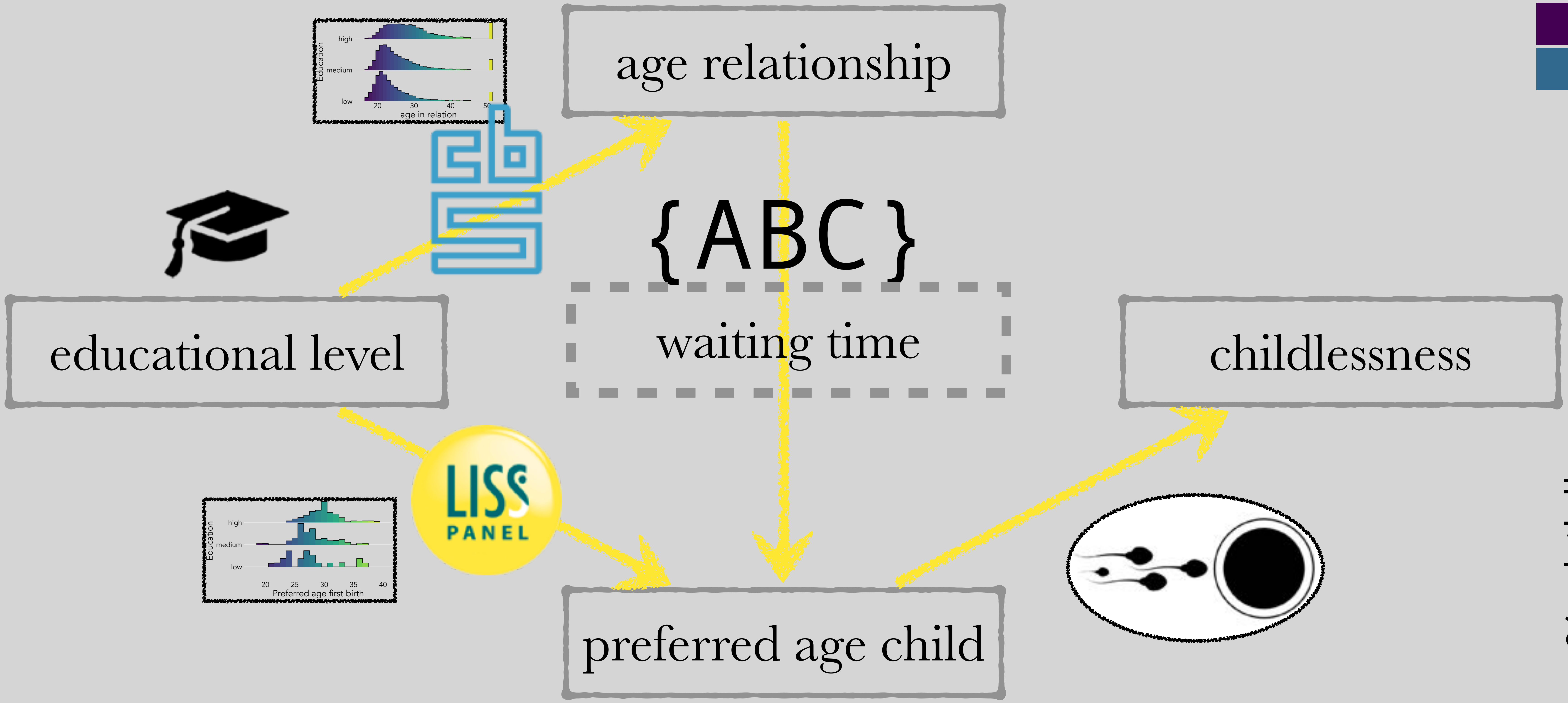

preferences + relation
relation
preferences
% childlessness
30
20
10
0
15%
17%
24%
low
medium
high
education

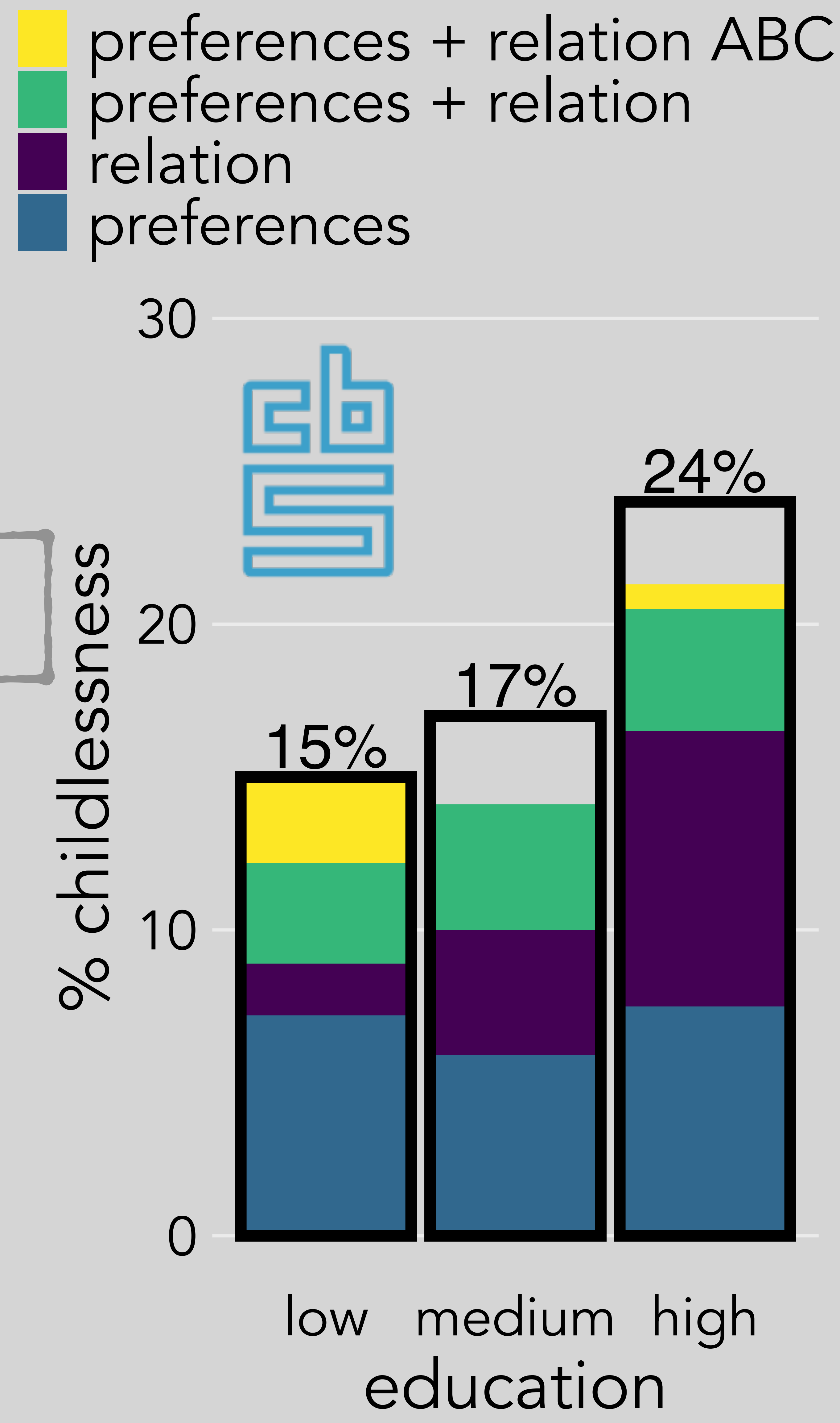




{ABC}
Approximate Bayesian Computation

{ABC}
Variation due to:
preferred waiting time child
differences in fecundability
differences due to stochasticity
age in relation ∝
actual
outcomes
age in relation
fecundability
stochasticity
∝
ABC
model
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
years until birth since relation





{ABC}
Approximate
Bayesian
Computation

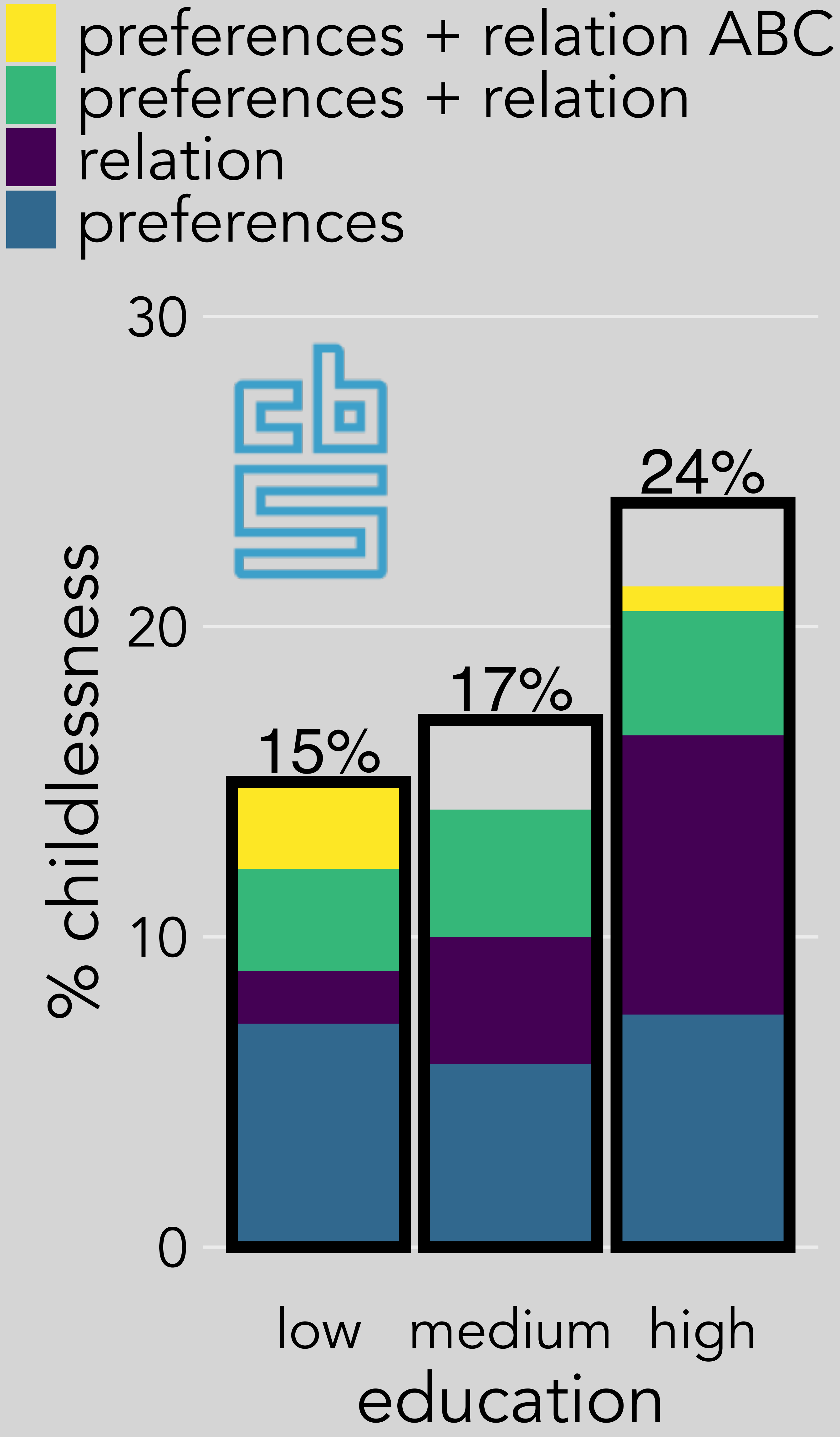
Where Did We Go Wrong?

Assumptions

1. No break-ups
2. All births are preferred
3. Preferences do not determine relationship
4. Preferences do not determine education
5. Preferences do not change
6. Education is not related to ‘biology’
7. Preferences are measured well

Improvements

1. Make waiting time dependent on age and education
2. Better measures of age in relationship



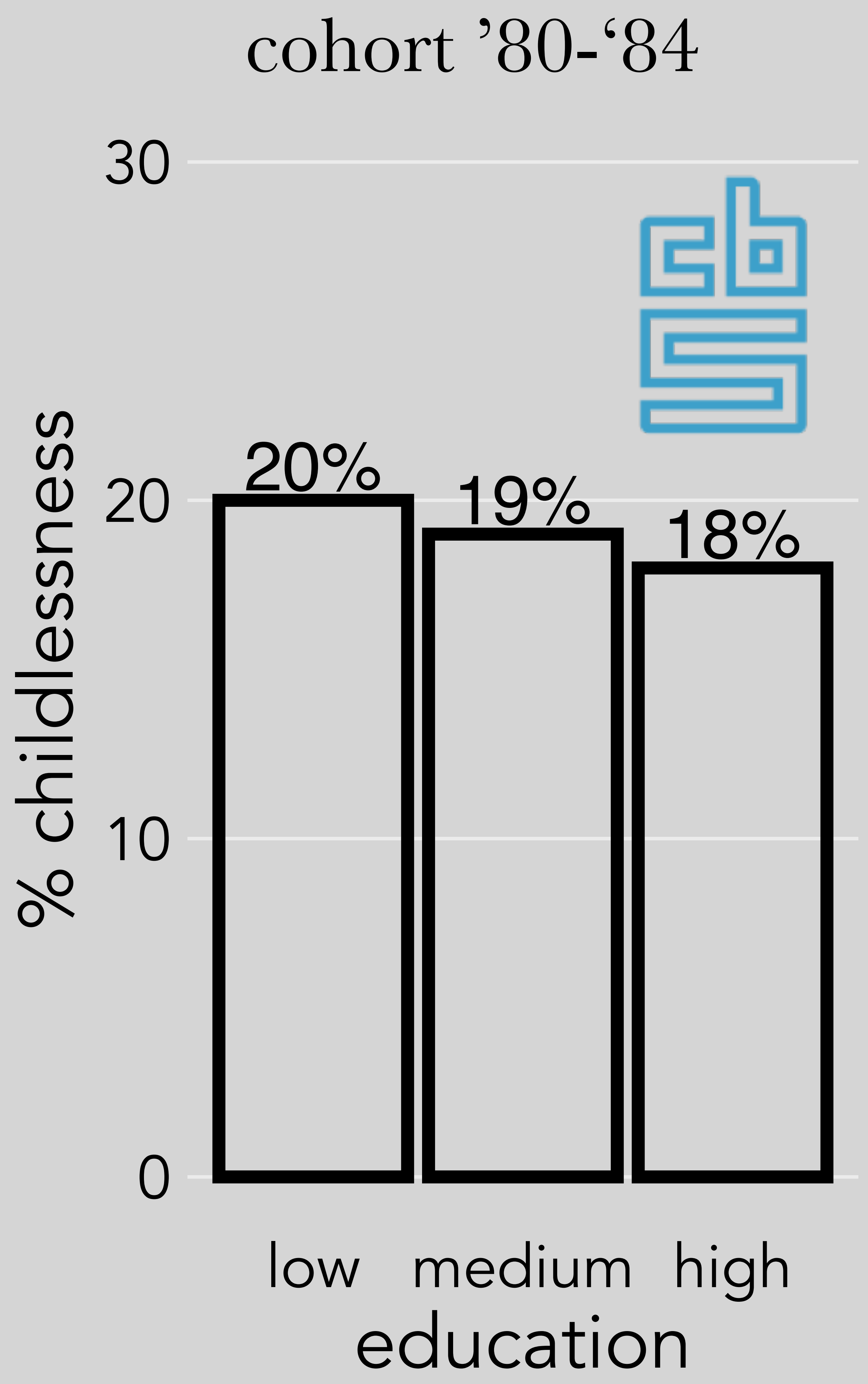
European Journal of Population (2019) 35:563–586
https://doi.org/10.1007/s10680-018-9492-2

CrossMark

Education, Gender, and Cohort Fertility in the Nordic Countries

Marika Jalovaara[1] · Gerda Neyer[2] · Gunnar Andersson[2] · Johan Dahlberg[2] · Lars Dommermuth[3] · Peter Fallesen[2,4] · Trude Lappegård[5]

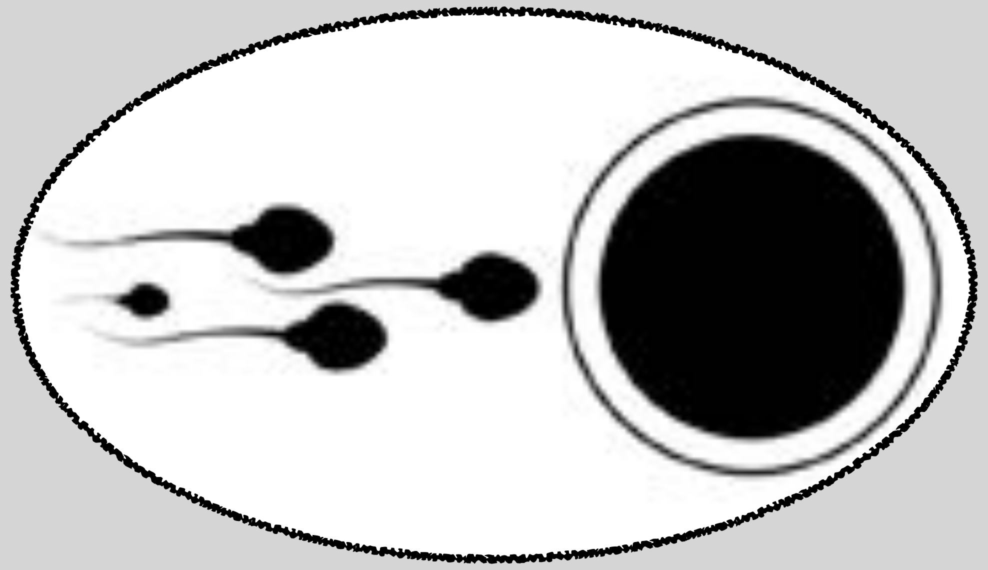
“In Denmark, Norway and Sweden, childlessness is now highest among the least educated women



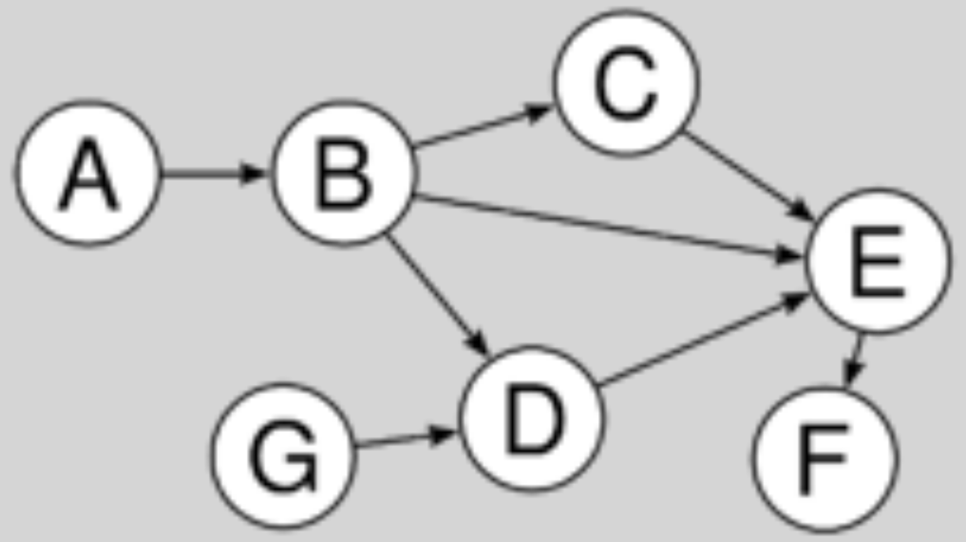
Take-Home Messages

microsimulation can advance traditional statistical modelling

microsimulation can:



include biological information

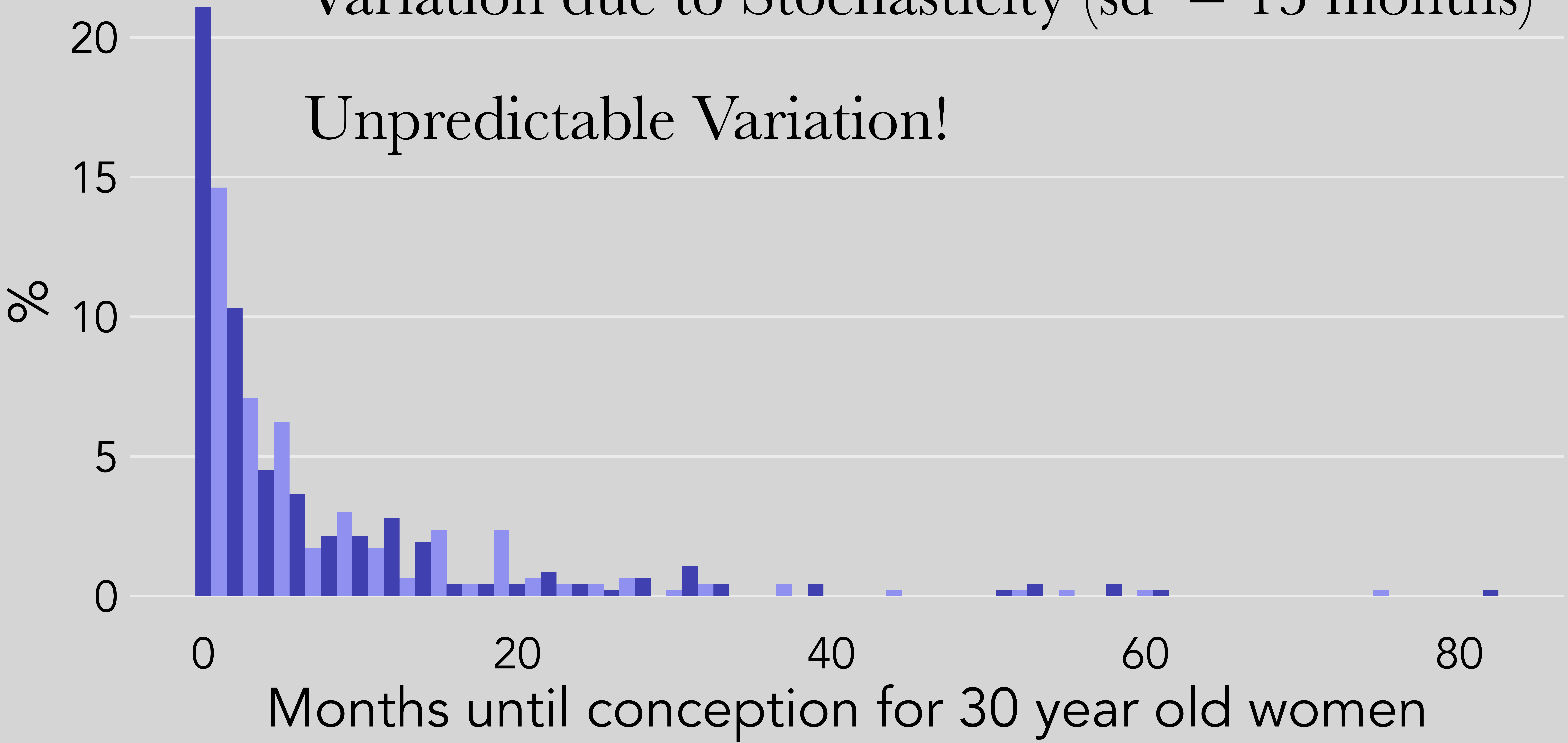


test (causal) mechanisms

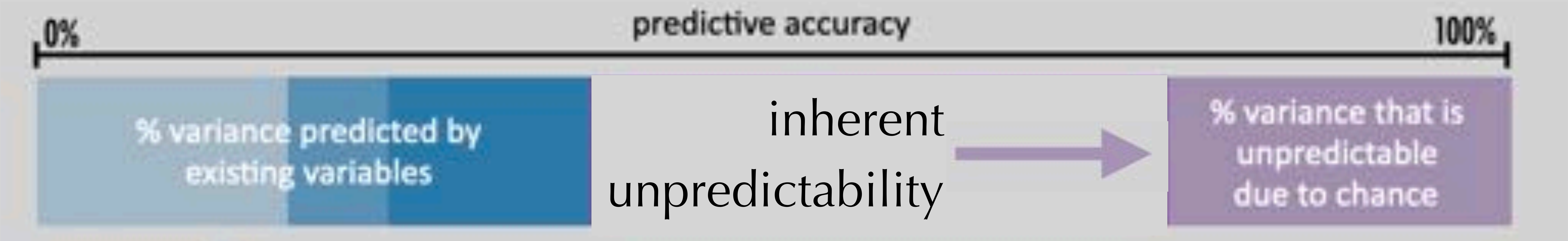
quantify unpredictability

Variation due to Stochasticity (sd = 13 months)

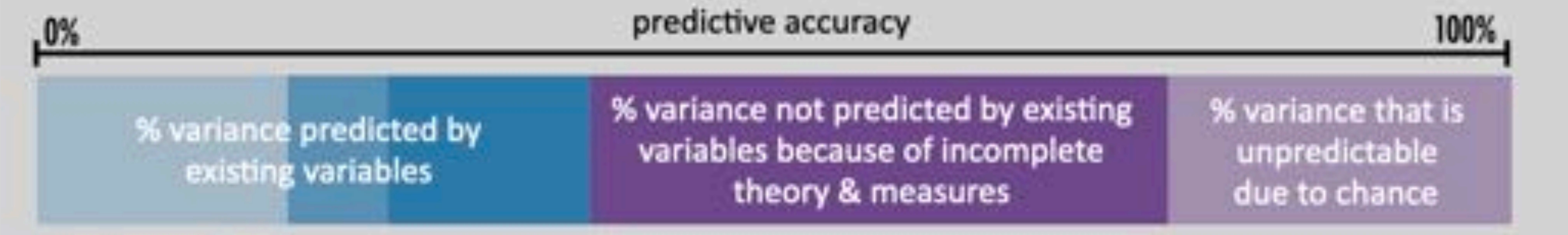
Unpredictable Variation!



Unpredictable Variation



Unique Insight into State of Field



?

The Proposal

a shift towards **prediction** leads to a more reliable and useful social science

microsimulation can advance traditional statistical modelling

FERTILITY PREDICTION CHALLENGE

March-August 2024

University of Groningen, Netherlands

0.54*

Is the current best [known to us] F1-score of a classifier that predicts who is going to have a child in the next three years

CAN YOU BEAT THIS SCORE?

Do you want to contribute to research on fertility behavior and the methodology of using prediction in social sciences?

Are you interested in working with unique registry-based datasets, including a social network for the entire Dutch population?

Are you looking for an engaging practical task for your machine learning course or workshop?

Or are you simply curious about the challenge and want to learn more about its design and prizes?

Sign up here to receive an update when the registration for the challenge opens and details are available

Contacts:
Gert Stulp g.stulp@rug.nl
Elizaveta Sivak e.sivak@rug.nl

university of groningen ODISSEI Eyra

* This result was obtained by the STL Trio Titans team at the data challenge at the SICSS-ODISSEI summer school in June 2023.