

AN INTRODUCTION TO BAYESIAN MCMC ESTIMATION FOR MISSING DATA ANALYSES

Craig Enders

UCLA Department of Psychology

OUTLINE

1

Big Three Missing Data Methods

2

Introduction to MCMC Estimation

3

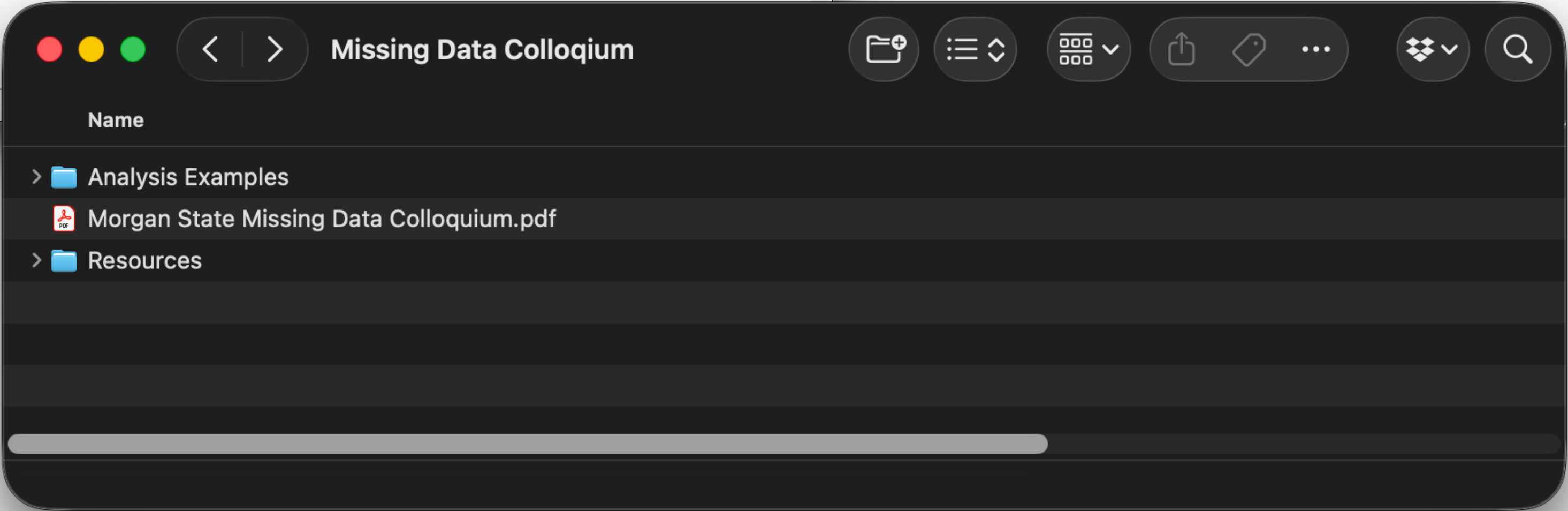
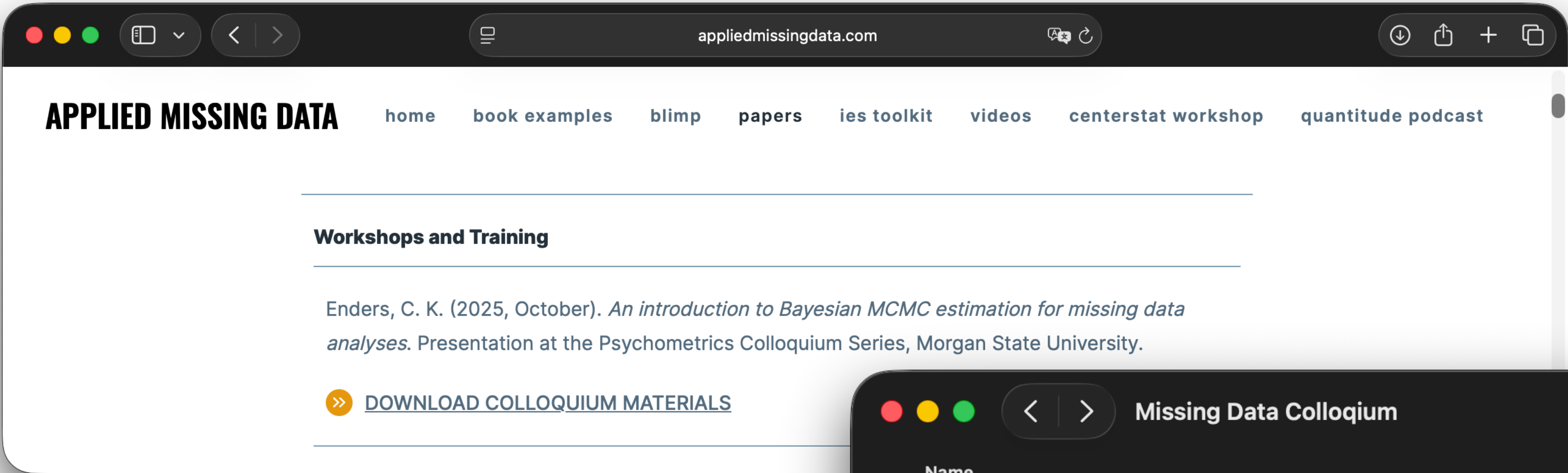
Application: Moderated Regression

4

Application: Structural Equation Model

COLLOQUIUM MATERIALS

WWW.APPLIEDMISSINGDATA.COM/BLIMP-PAPERS



OUTLINE

1

Big Three Missing Data Methods

2

Introduction to MCMC Estimation

3

Application: Moderated Regression

4

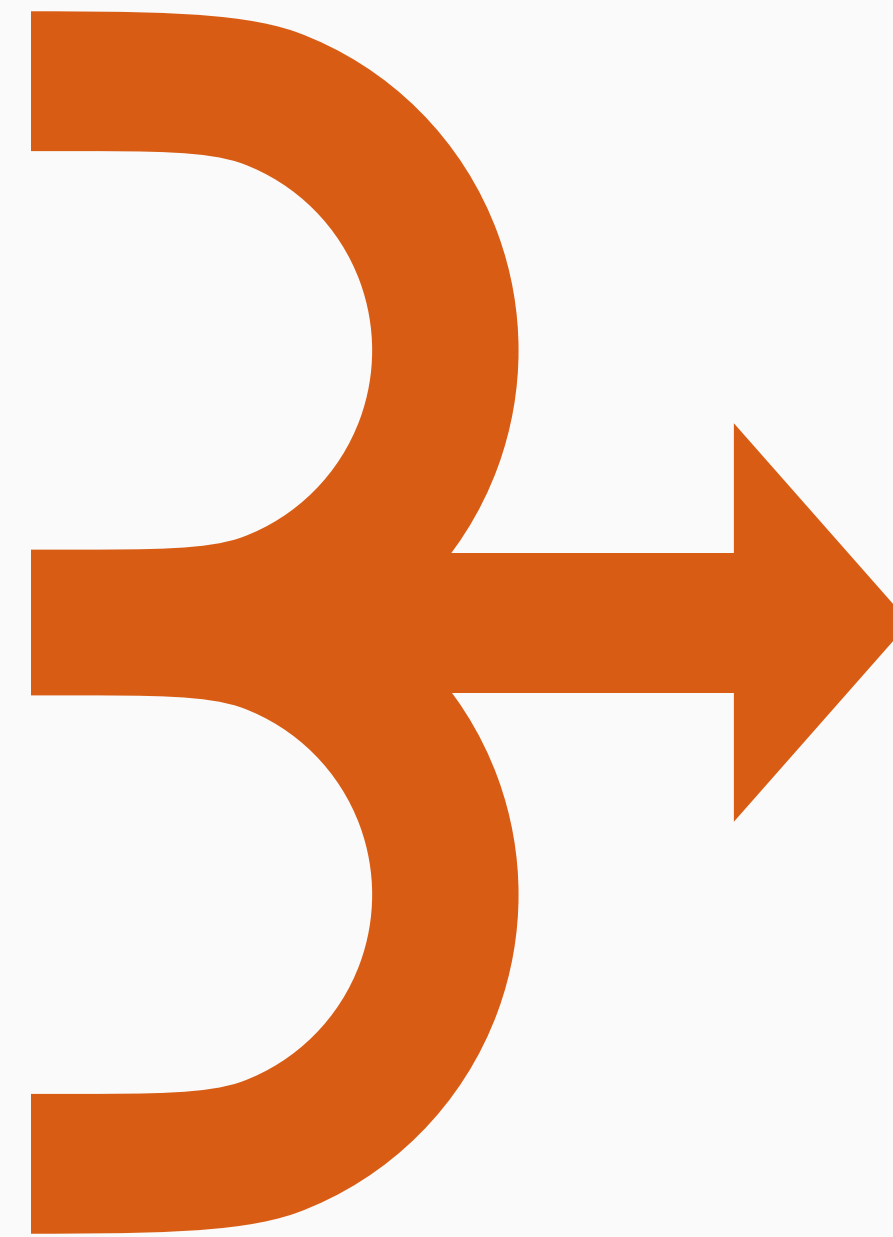
Application: Structural Equation Model

MODERN MISSING DATA METHODS

Maximum likelihood

Multiple imputation

Bayesian MCMC estimation



**the
Big
Three**

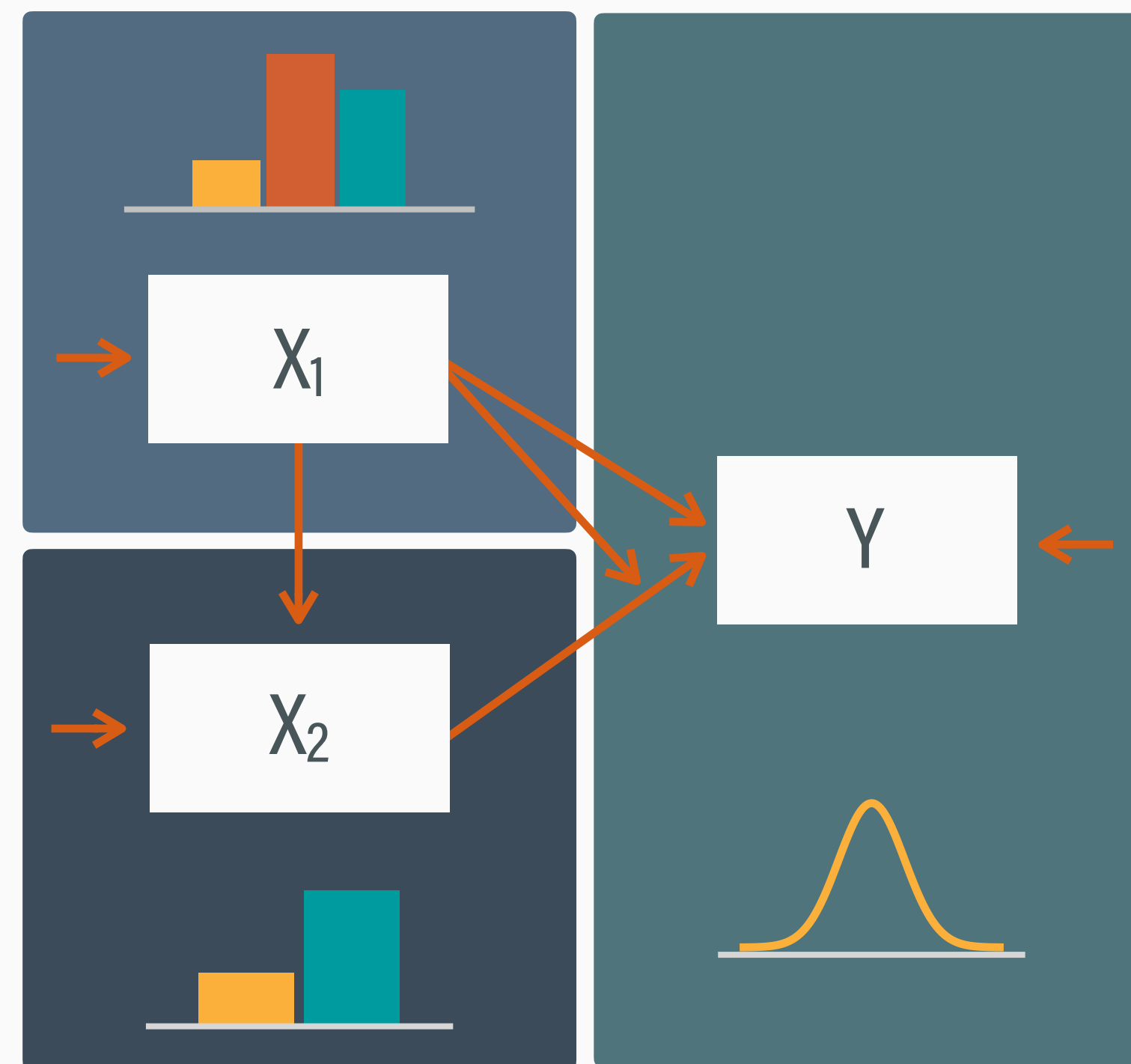
KEY ADVANTAGES OF BIG THREE

- Use all available data, maximize power, no wasted resources
- Unbiased under systematic nonresponse process where missingness related to observed data but not to unseen scores (conditionally missing at random)
- Allow for alternate assumptions about nonresponse process
- Accessible software tools and a large supporting literature

MODELING FRAMEWORKS

Multivariate modeling

Factored regression specification



- Factored specifications invoke a unique submodel and distribution for each variable
- Submodels can include terms that are at odds with multivariate normality (e.g., discrete variables, interactions, random slopes)

CHOOSING A MISSING DATA METHOD

- All things being equal—same data, same variables, same assumptions—the Big Three rarely produce different results
- Missing data analyses require distributional assumptions for all incomplete variables, including those that wouldn't usually require assumptions (e.g., predictors)
- Factored specifications provide the greatest flexibility for modeling complex data structures

MCMC + FACTORED REGRESSION

- ◉ MCMC with factored specifications accommodates a broader range of applications than other Big Three methods:
- ◉ Mixed metrics (normal, ordinal, nominal, skewed, count, and latent)
- ◉ Nonlinear effects (interactions and curvilinear trends)
- ◉ Multilevel structures (random coefficients, interactions, variance heterogeneity, and dynamic effects)
- ◉ Latent variable models incorporating any of these features

OUTLINE

1

Big Three Missing Data Methods

2

Introduction to MCMC Estimation

3

Application: Moderated Regression

4

Application: Structural Equation Model

SIMPLE REGRESSION ILLUSTRATION

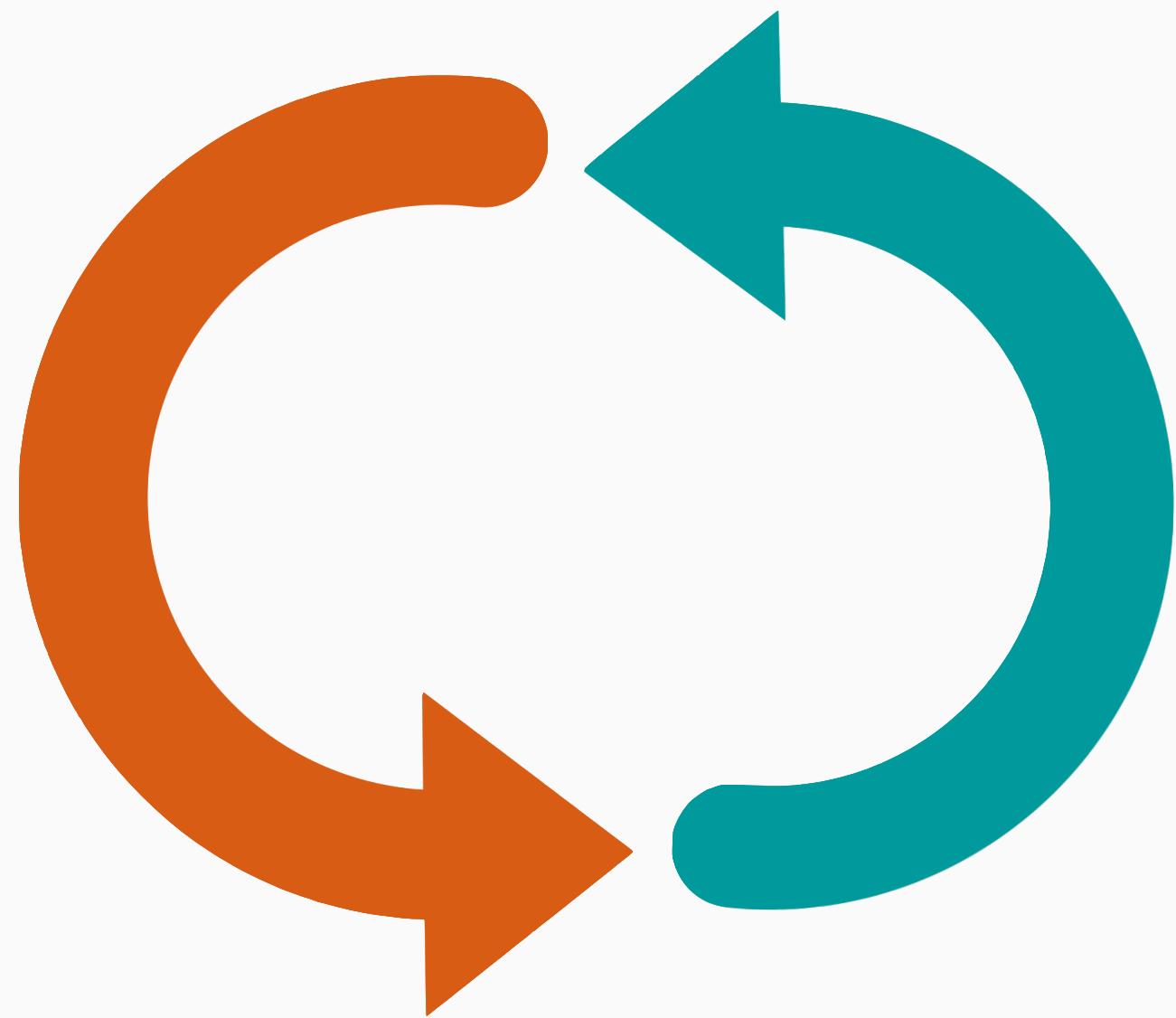
- Study that seeks to determine whether learning problem in 1st grade predict 9th grade reading achievement in middle school

$$\text{read}_9 = \beta_0 + \beta_1(\text{lrnprob}_1) + \varepsilon$$

Variable	Definition	Missing %	Scale
atrisk	Emotional/behavioral risk code	2.2	0 = Low, 1 = Medium/high
read1	1st grade broad reading composite	6.5	Numeric (39 to 153)
lrnprob1	1st grade learning problems	2.2	Numeric (31 to 88)
read9	9th grade broad reading composite	17.4	Numeric (41 to 123)

MCMC ESTIMATION

Estimate model parameters



Impute missing values

Do for $t = 1$ to 10,000 iterations

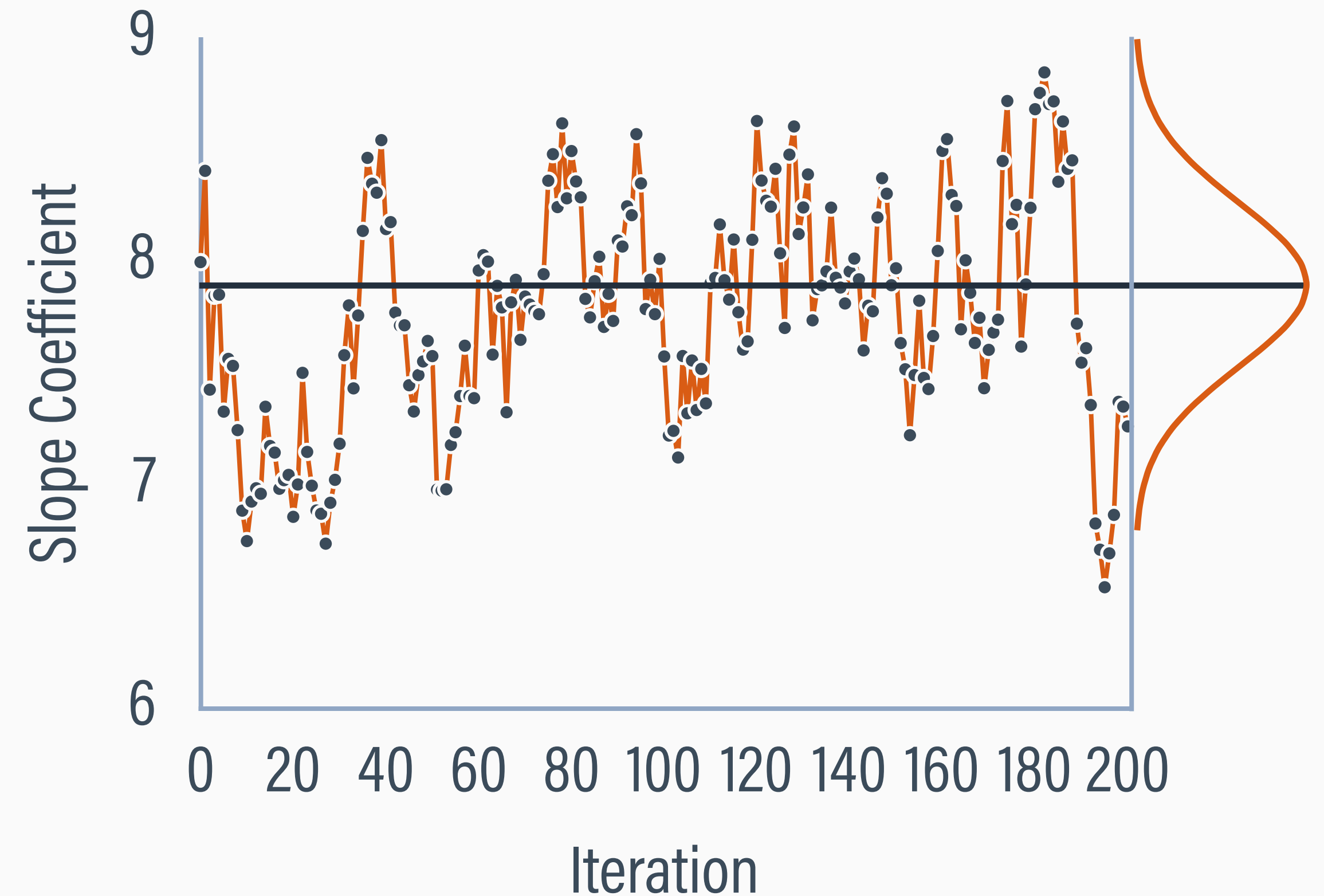
- » Estimate model parameters from the filled-in data
- » Impute missing values based on the model parameters

Repeat

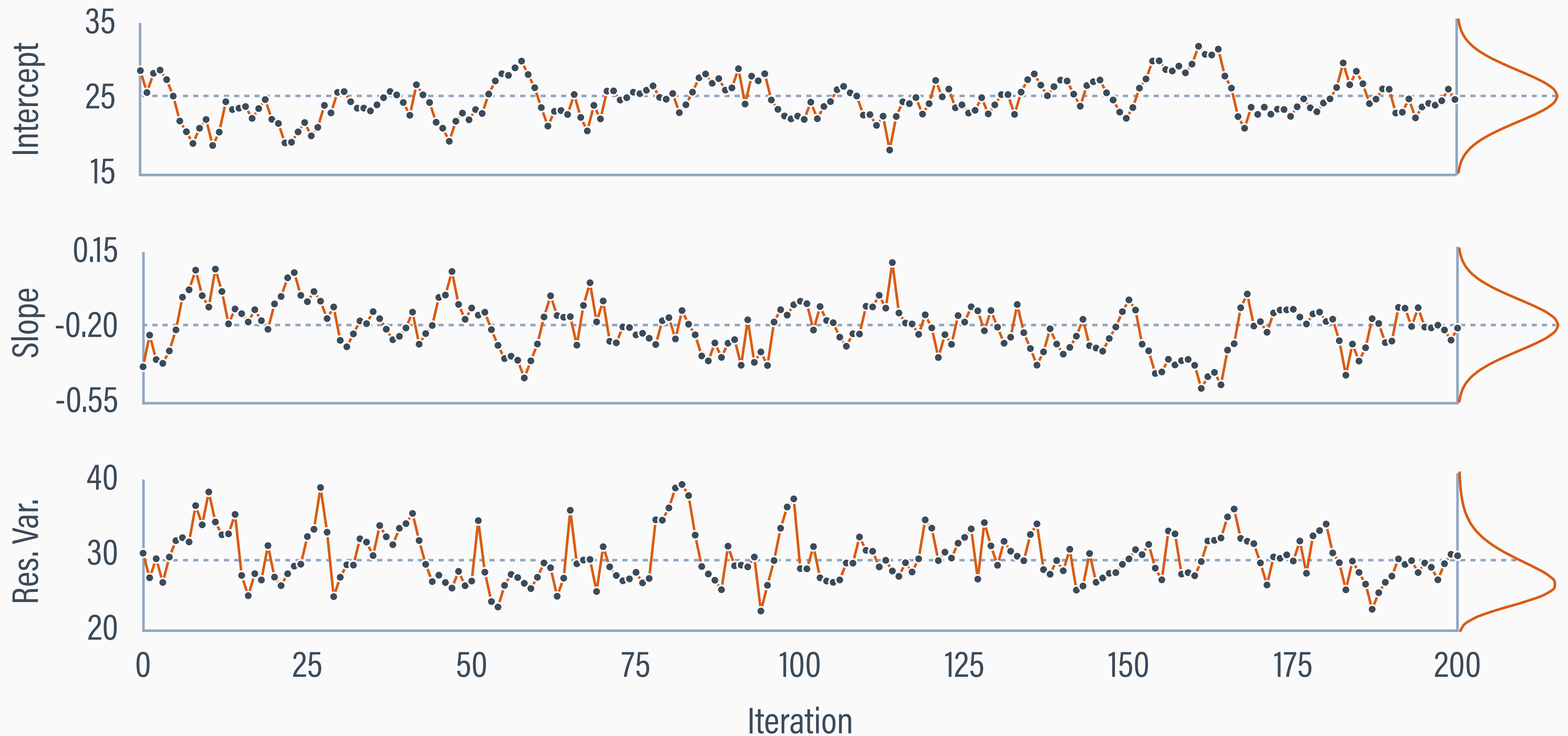
Summarize model parameters

MEANING OF ESTIMATION

- MCMC uses computer simulation to “sample” parameters from a distribution
- Each iteration gives plausible parameter values that could have produced our data
- These accumulate into an empirical distribution of plausible parameter values

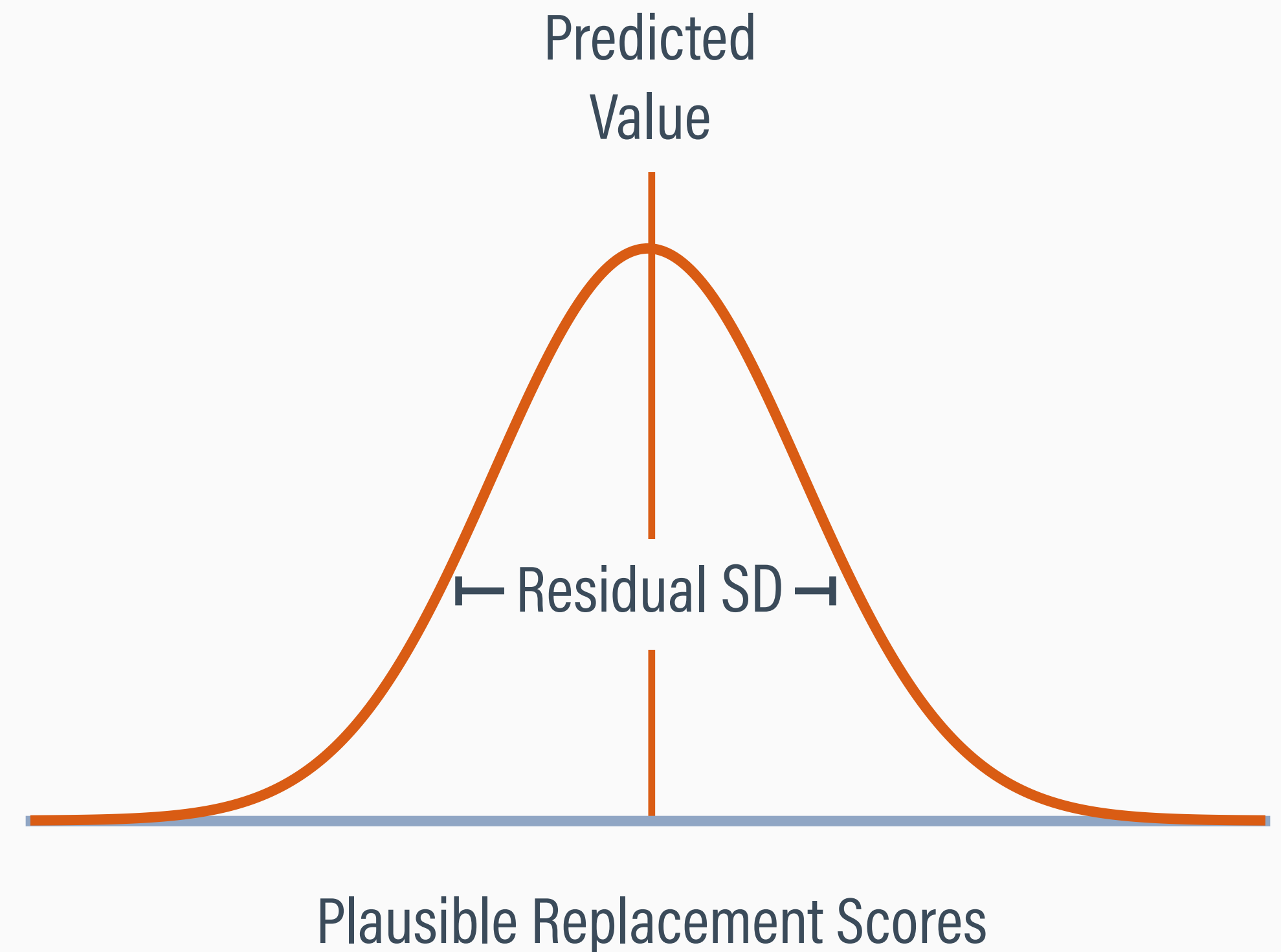


PARAMETERS FROM 200 MCMC CYCLES

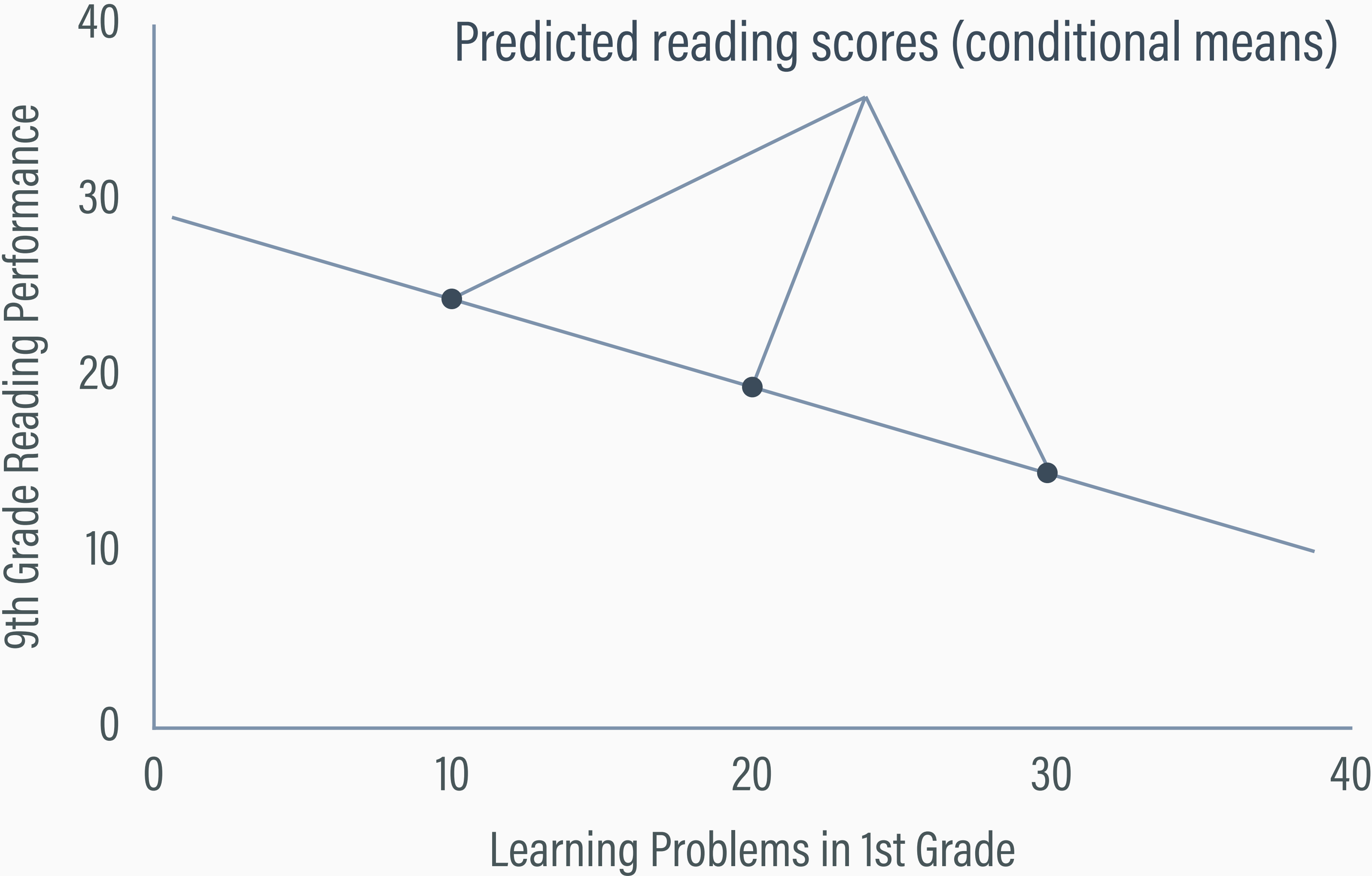


MISSING DATA IMPUTATION

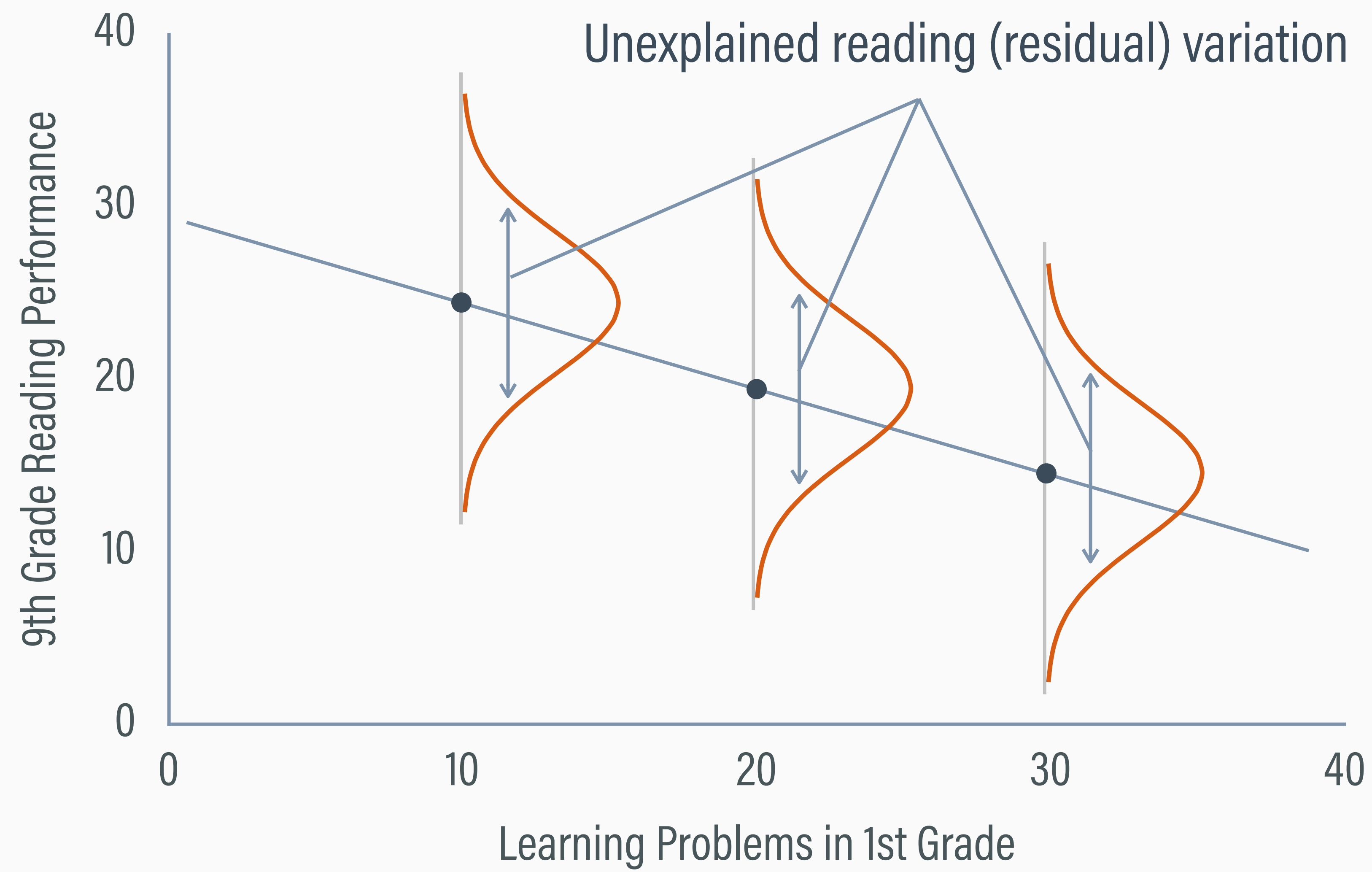
- Missing values are sampled from a distribution of plausible replacement scores
- The model parameters at each iteration define the distribution's shape
- Imputation = predicted value + computer-generated residual term



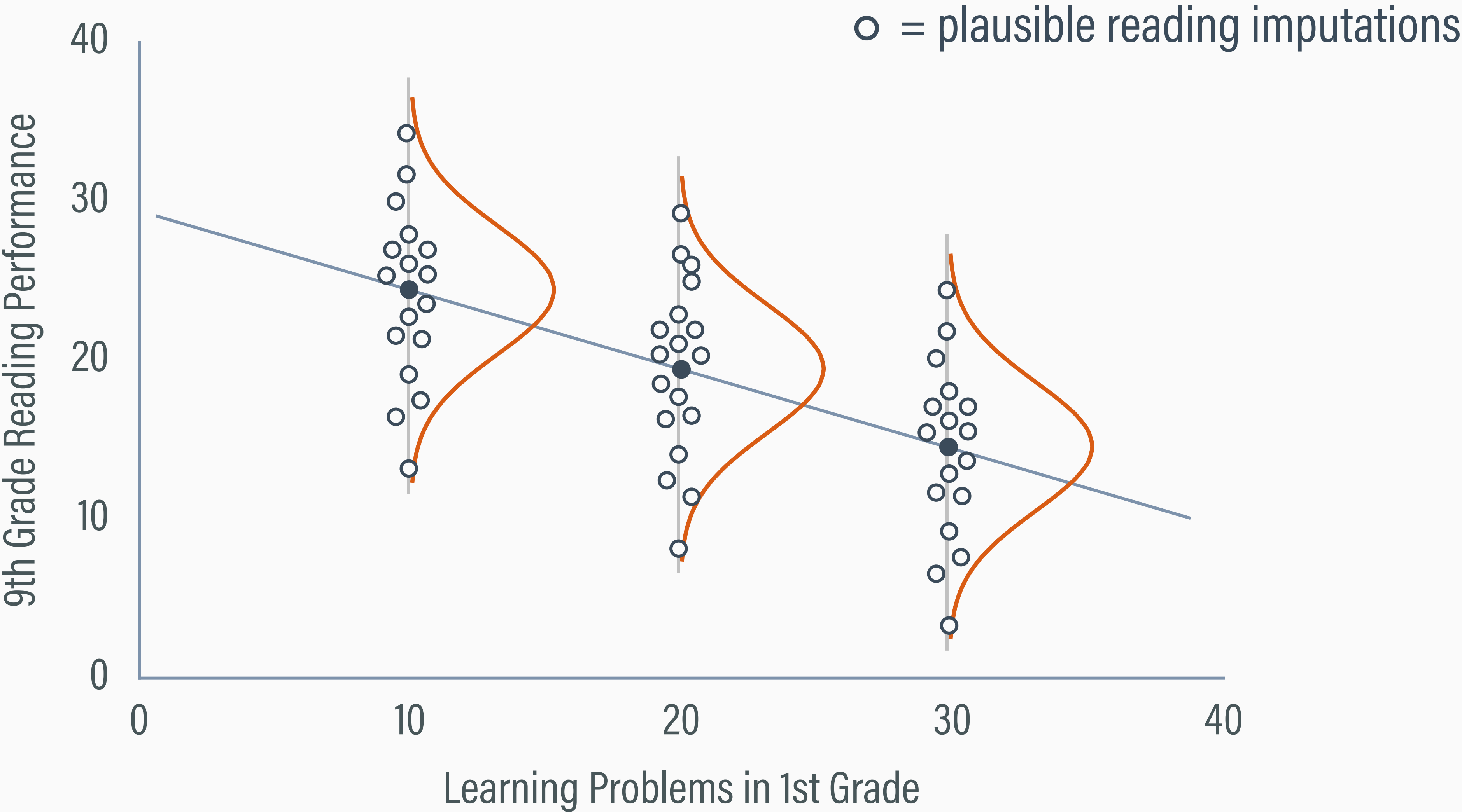
PREDICTED VALUES



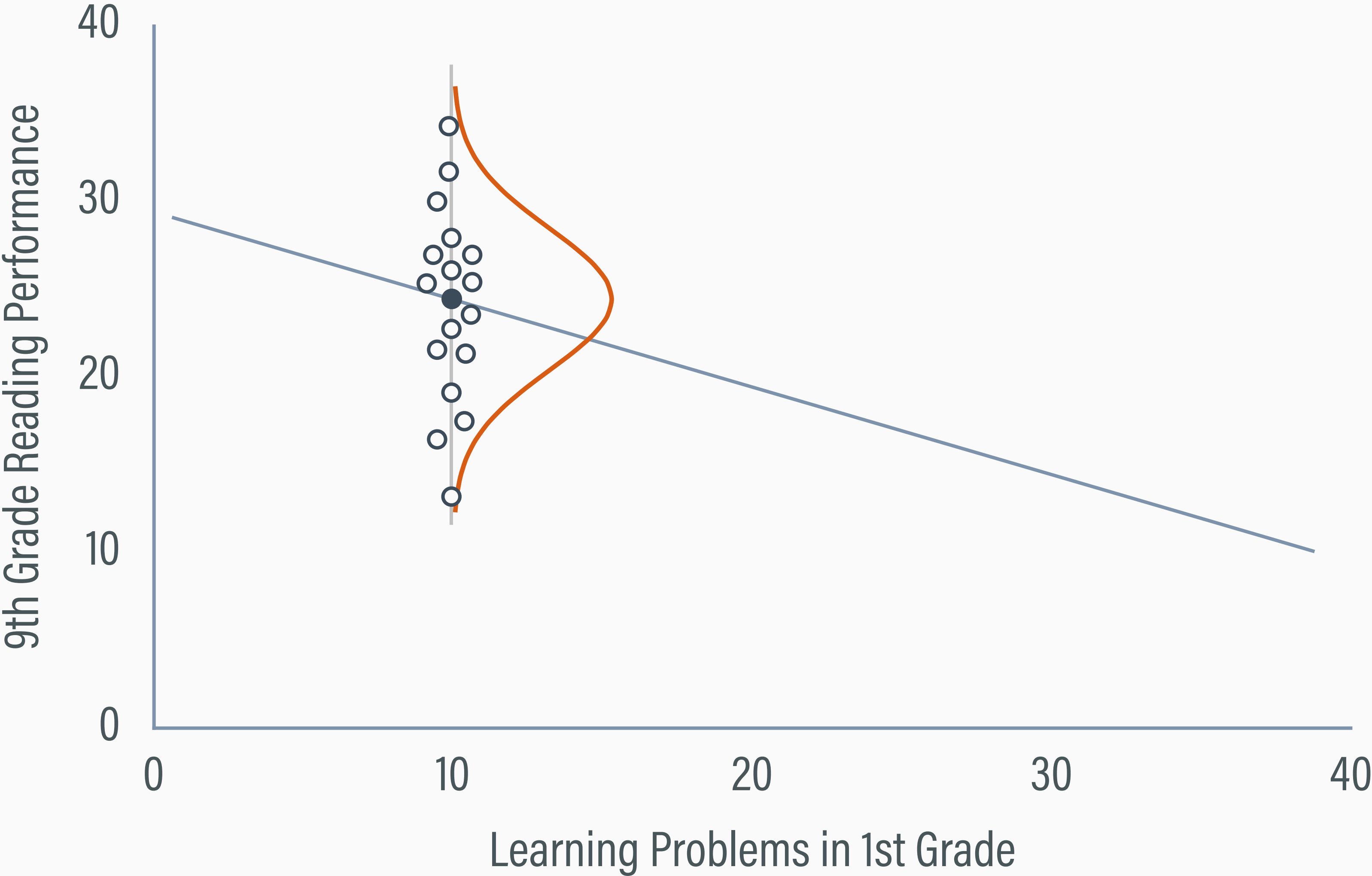
RESIDUAL VARIATION



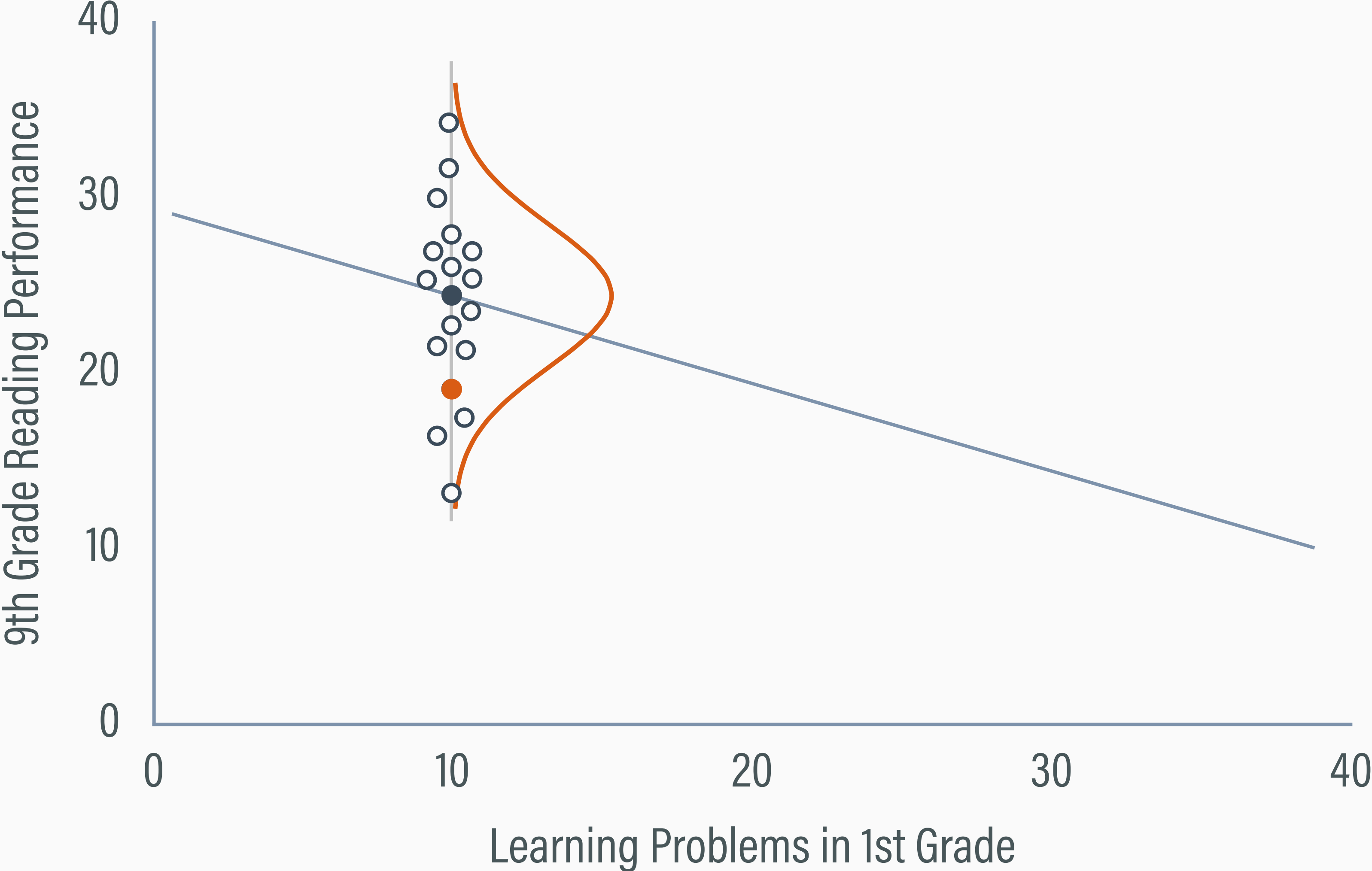
DISTRIBUTIONS OF IMPUTATIONS



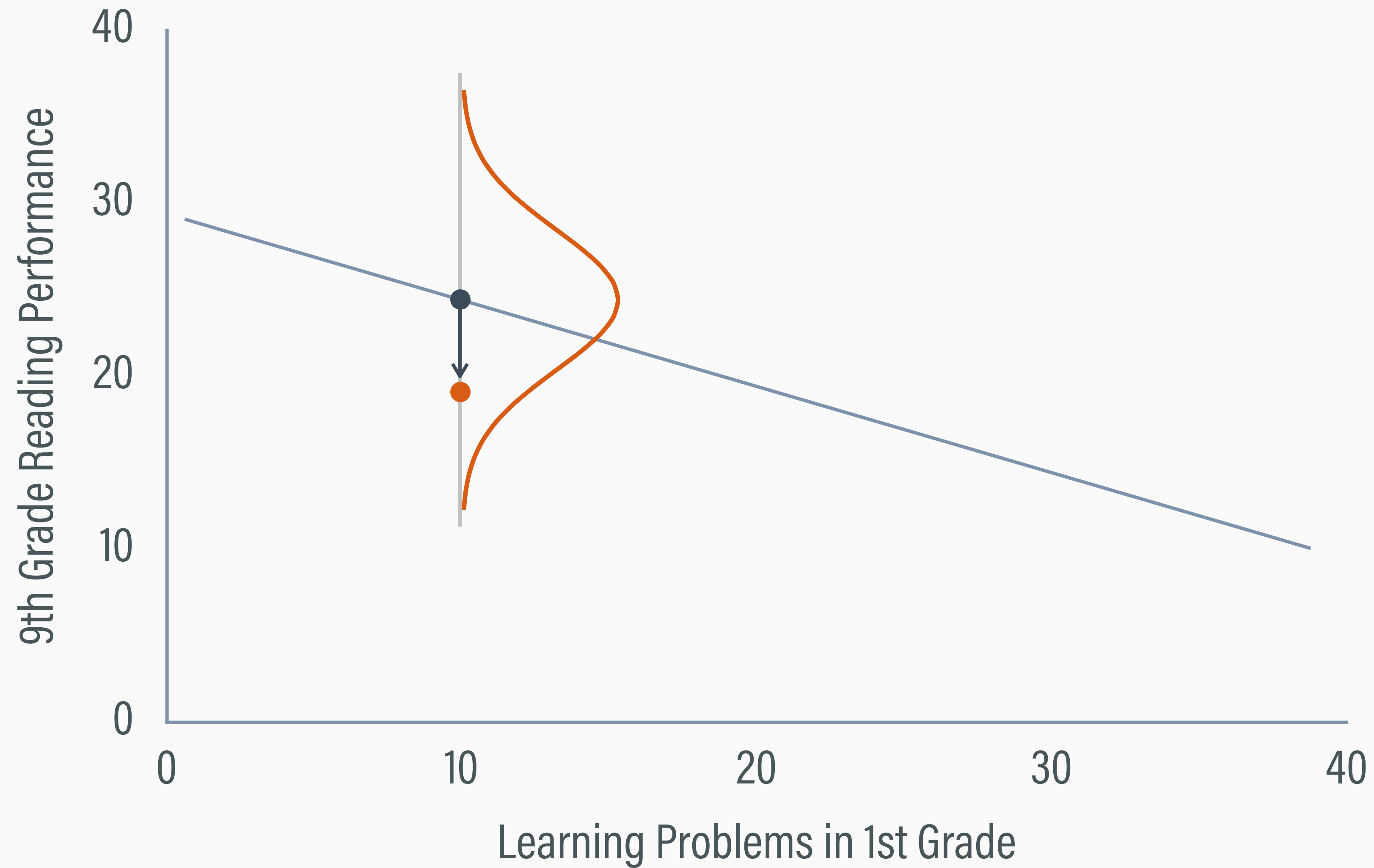
IMPUTATION FOR LOW LEARNING PROBLEMS



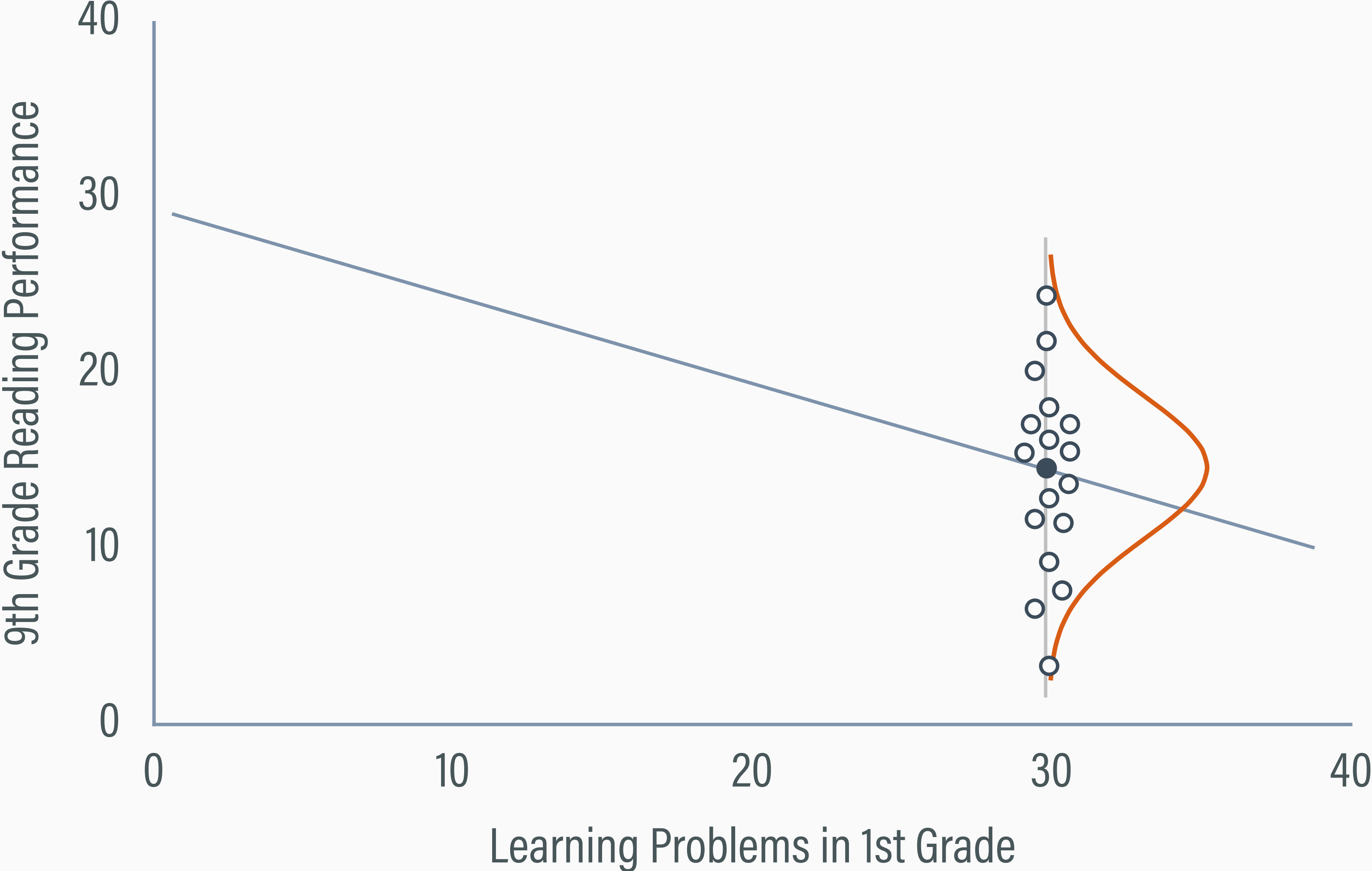
SAMPLE IMPUTATION AT RANDOM



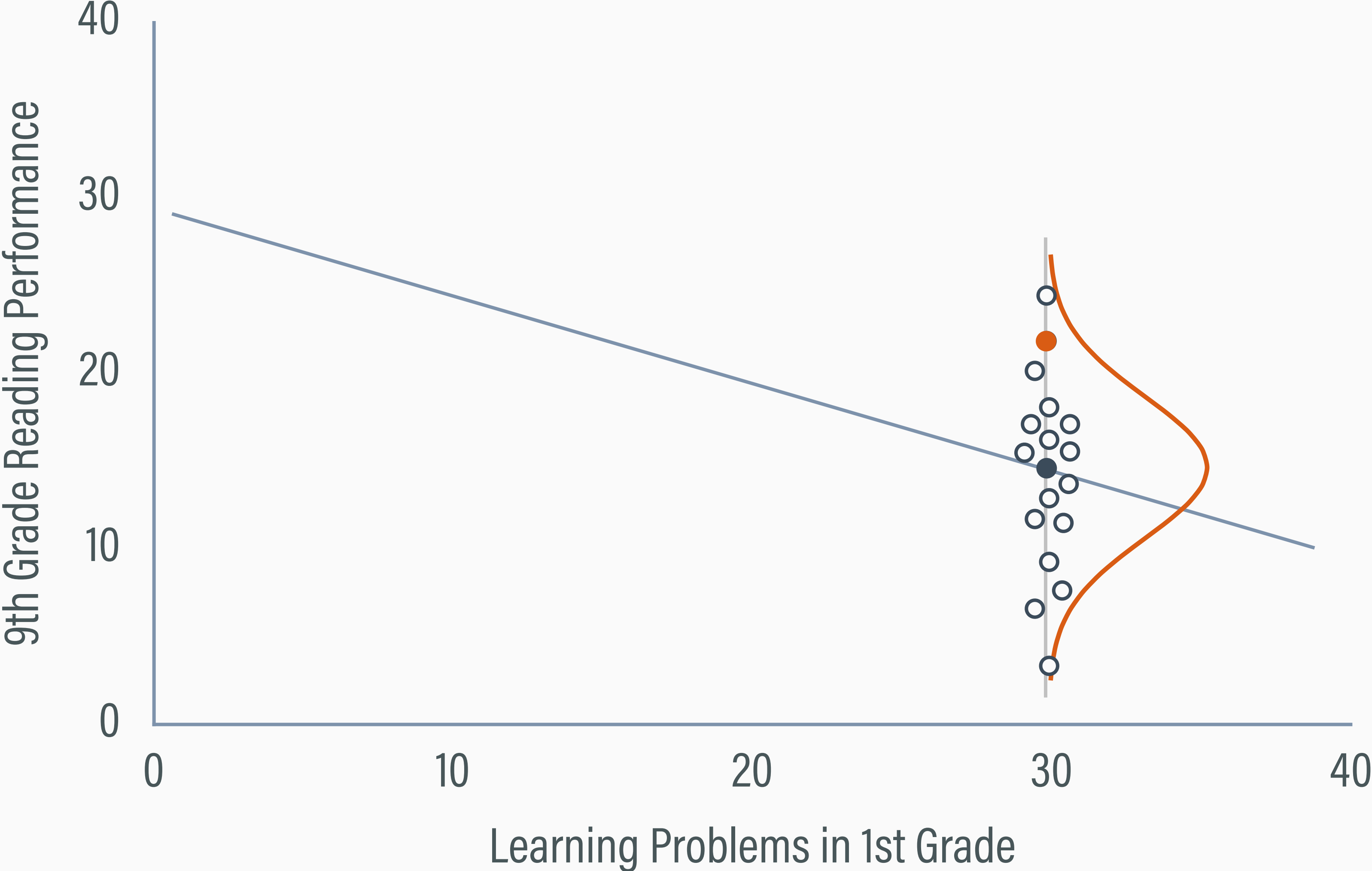
IMPUTATION = PREDICTION + NOISE



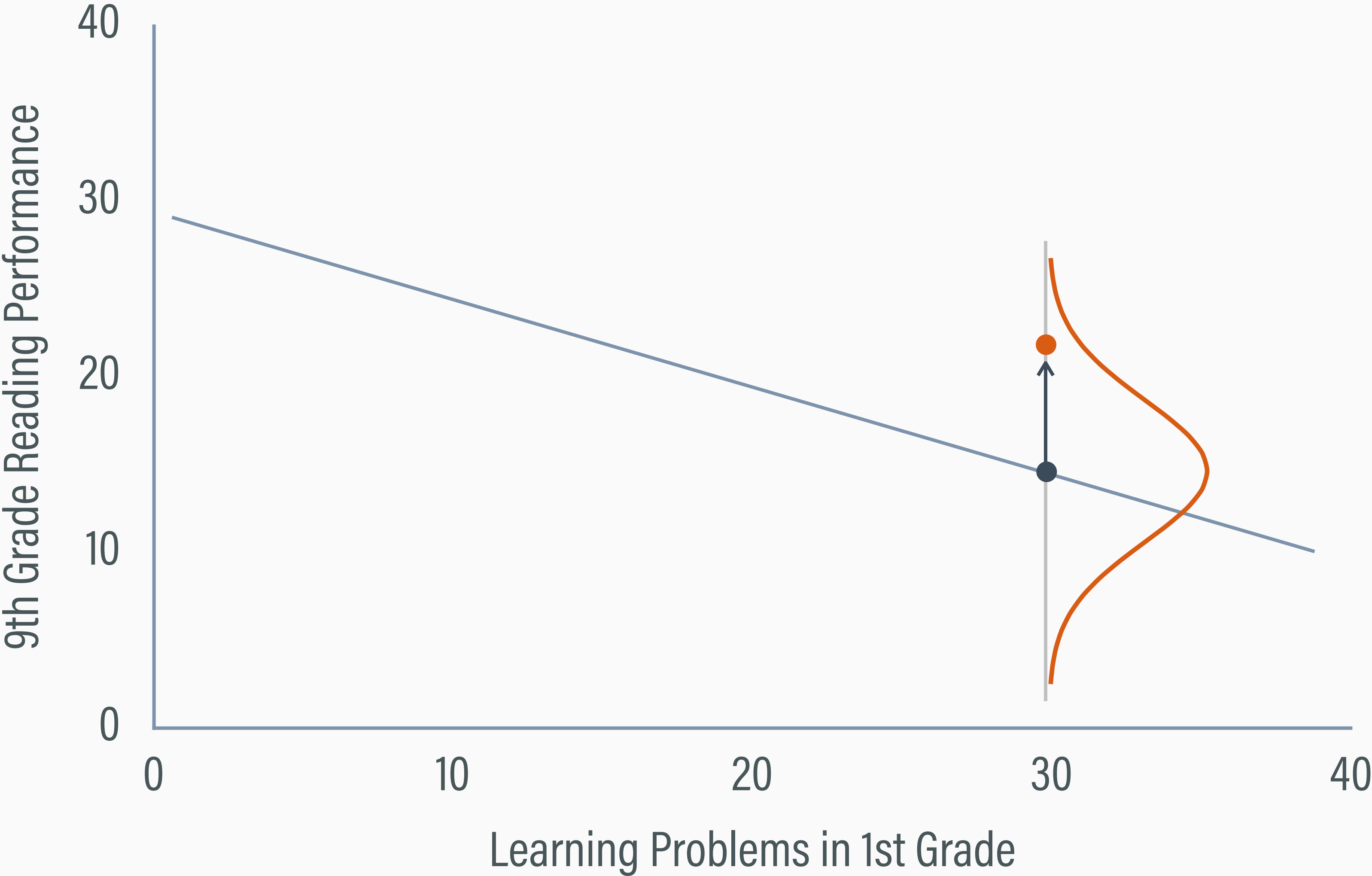
IMPUTATION FOR HIGH LEARNING PROBLEMS



SAMPLE IMPUTATION AT RANDOM



IMPUTATION = PREDICTION + NOISE

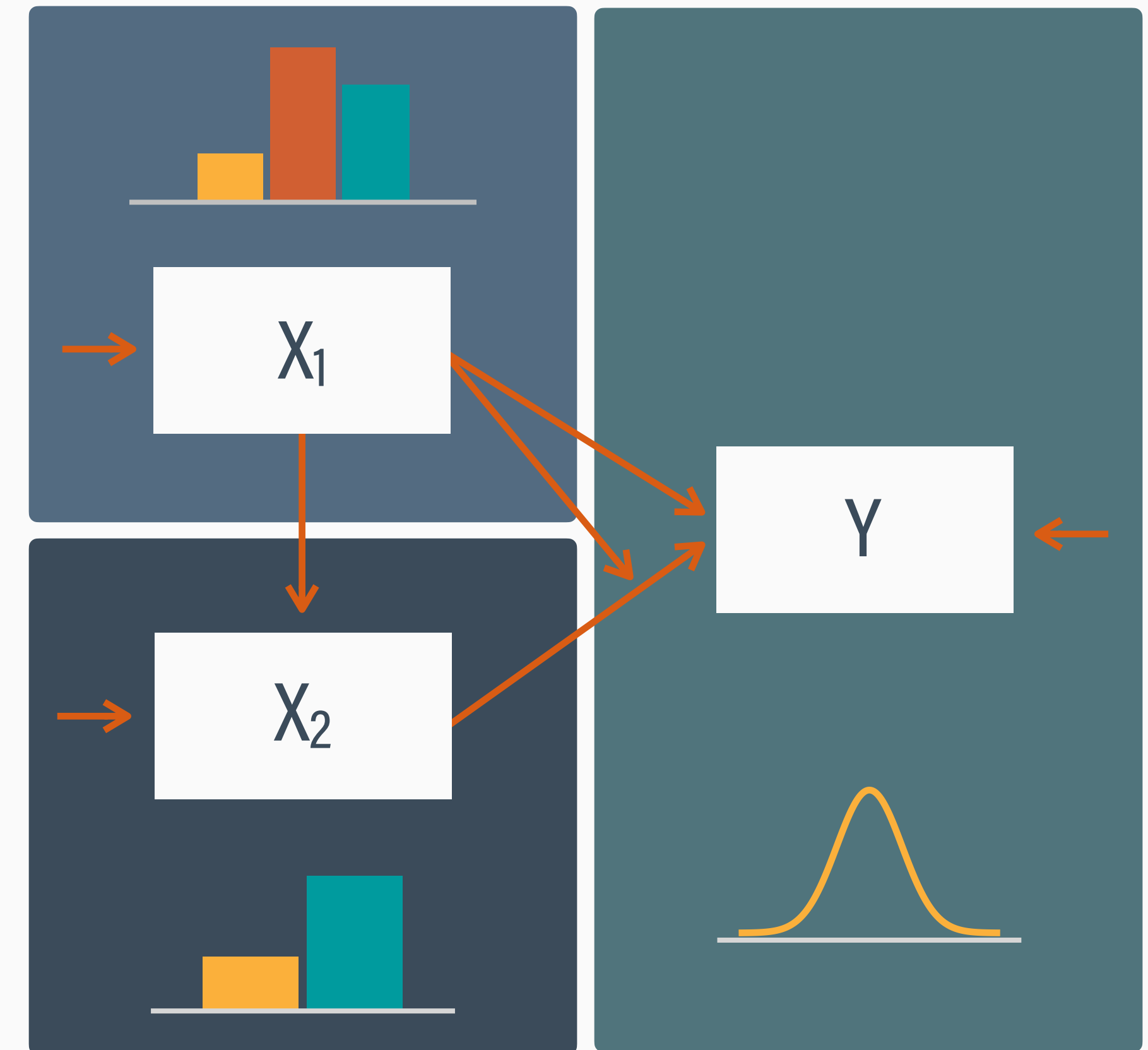


INCOMPLETE PREDICTORS

- Incomplete predictors require their own model and distributional assumptions
- Multivariate methods can misspecify these distributions in a way that introduces bias
- Factored regression combines information from the focal and predictor models to get predicted values and residual variation

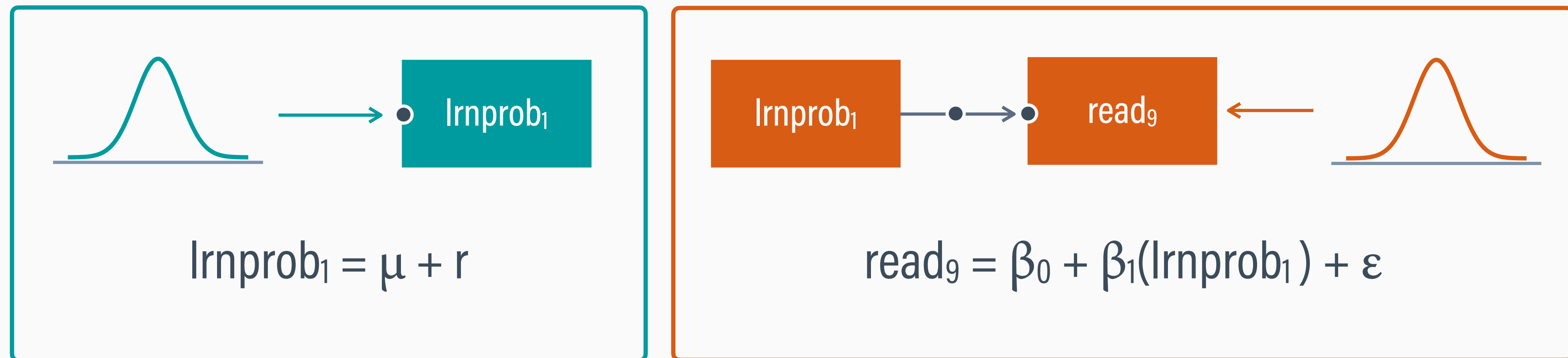
FACTORED REGRESSION SPECIFICATIONS

- Factored specifications invoke a unique submodel and distribution for each variable
- Submodels can include terms that are at odds with multivariate normality (e.g., discrete variables, interactions, random slopes)

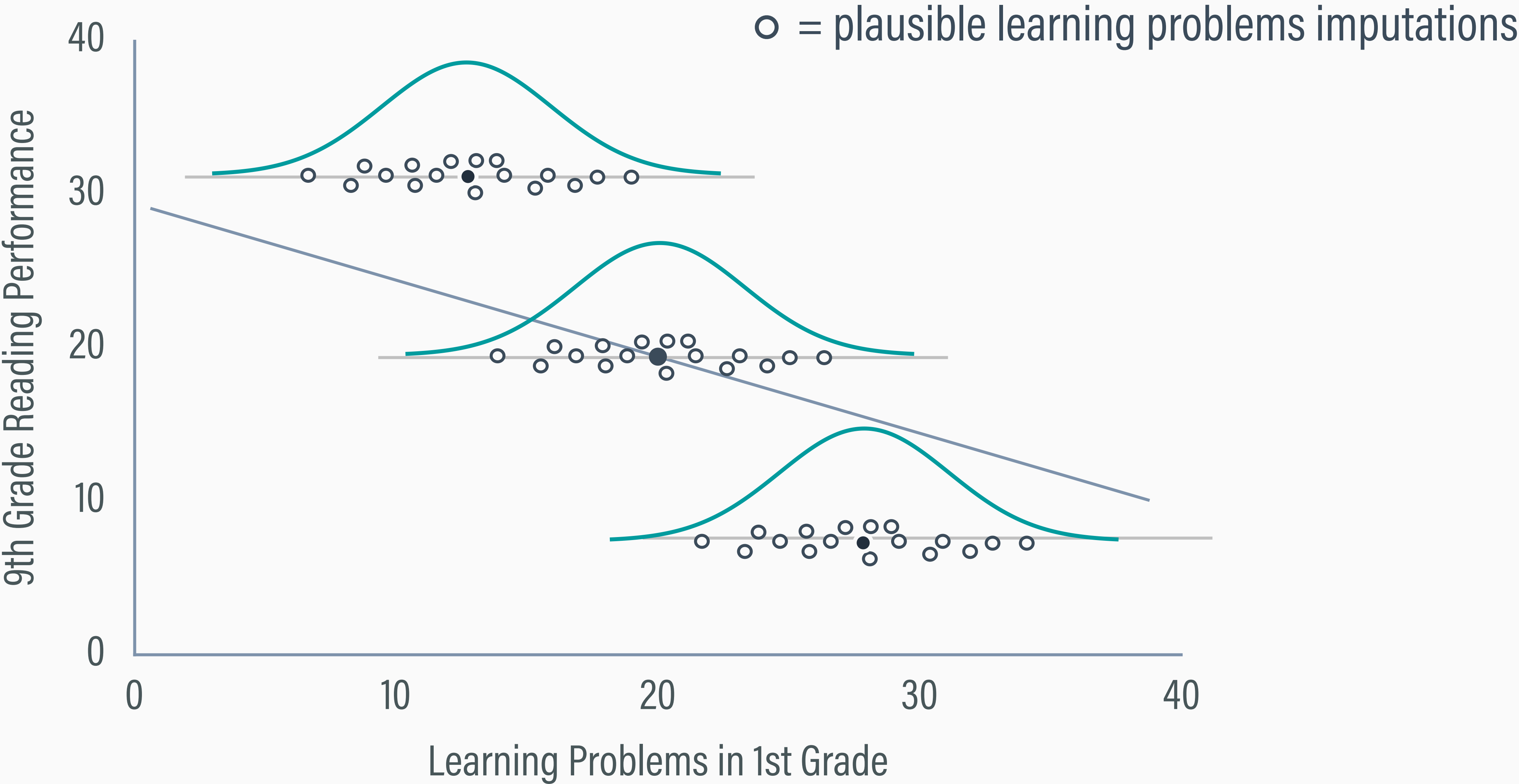


FACTORED SPECIFICATION

- A bivariate distribution equals the product of two univariate distributions: $f(\text{read}_9, \text{lnnprob}_1) = f(\text{read}_9 | \text{lnnprob}_1) \times f(\text{lnnprob}_1)$
- Each submodel distribution is potentially unique

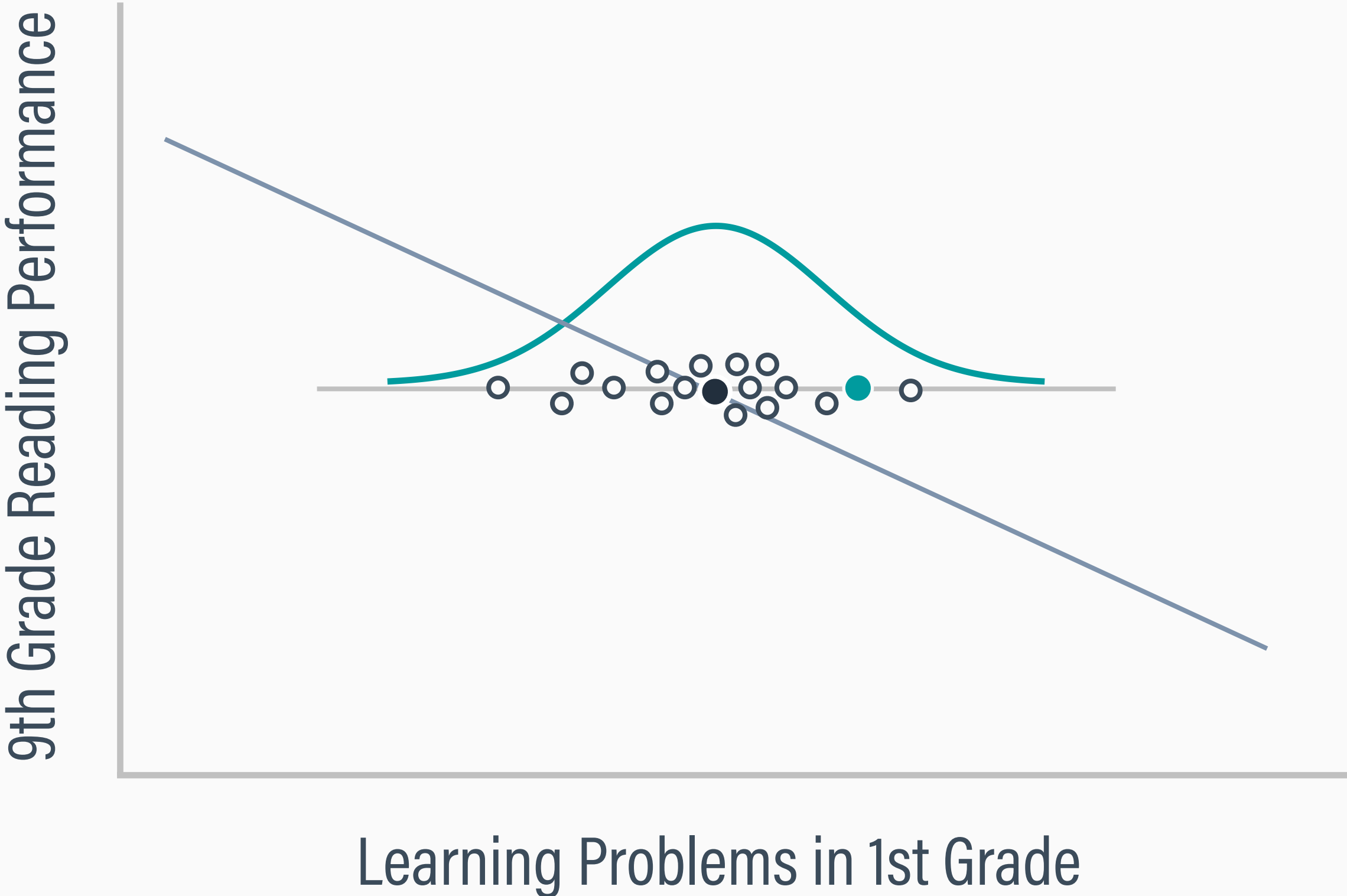


DISTRIBUTIONS OF IMPUTATIONS

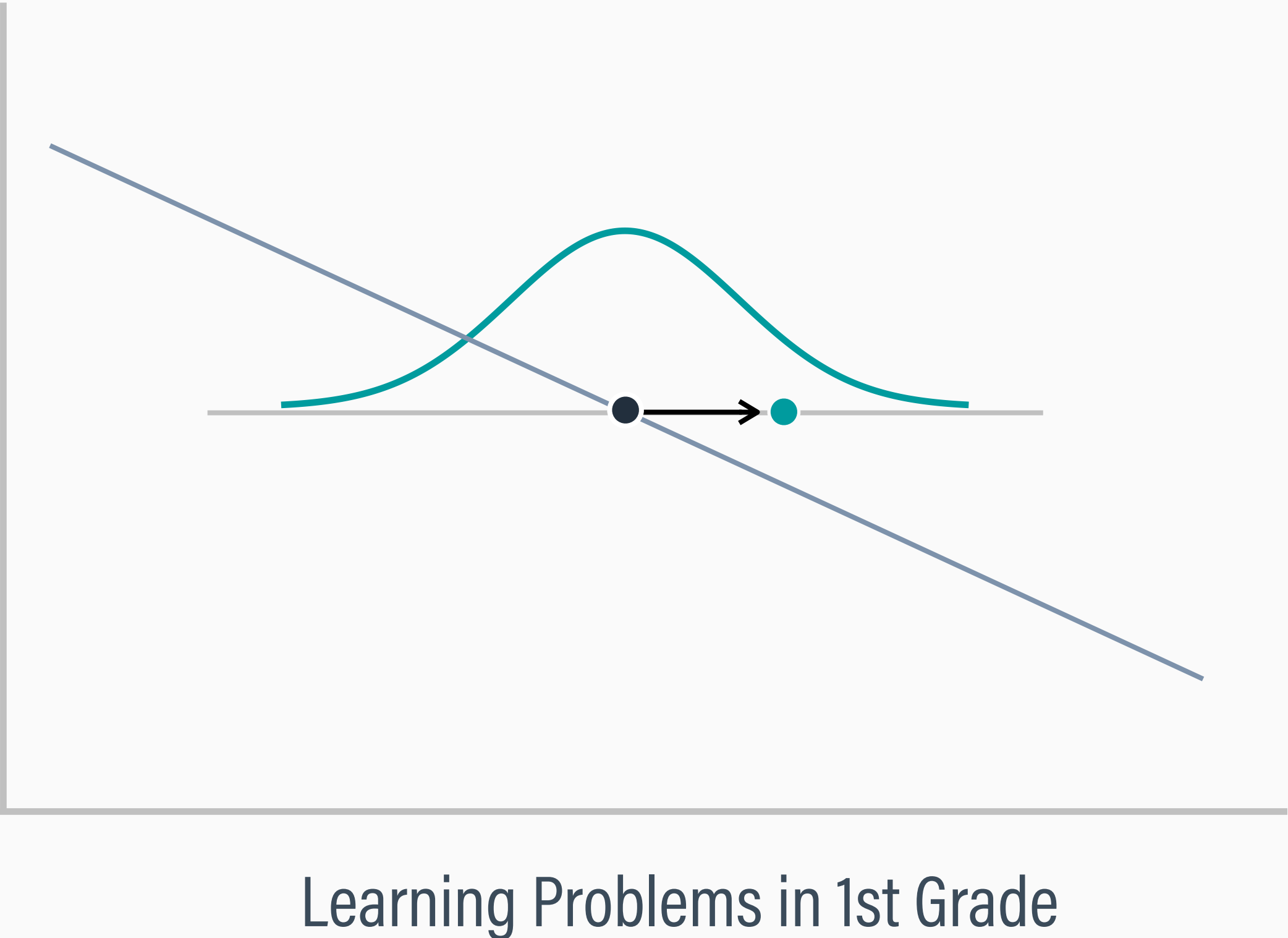


IMPUTATION EXAMPLE

Sample imputation at random

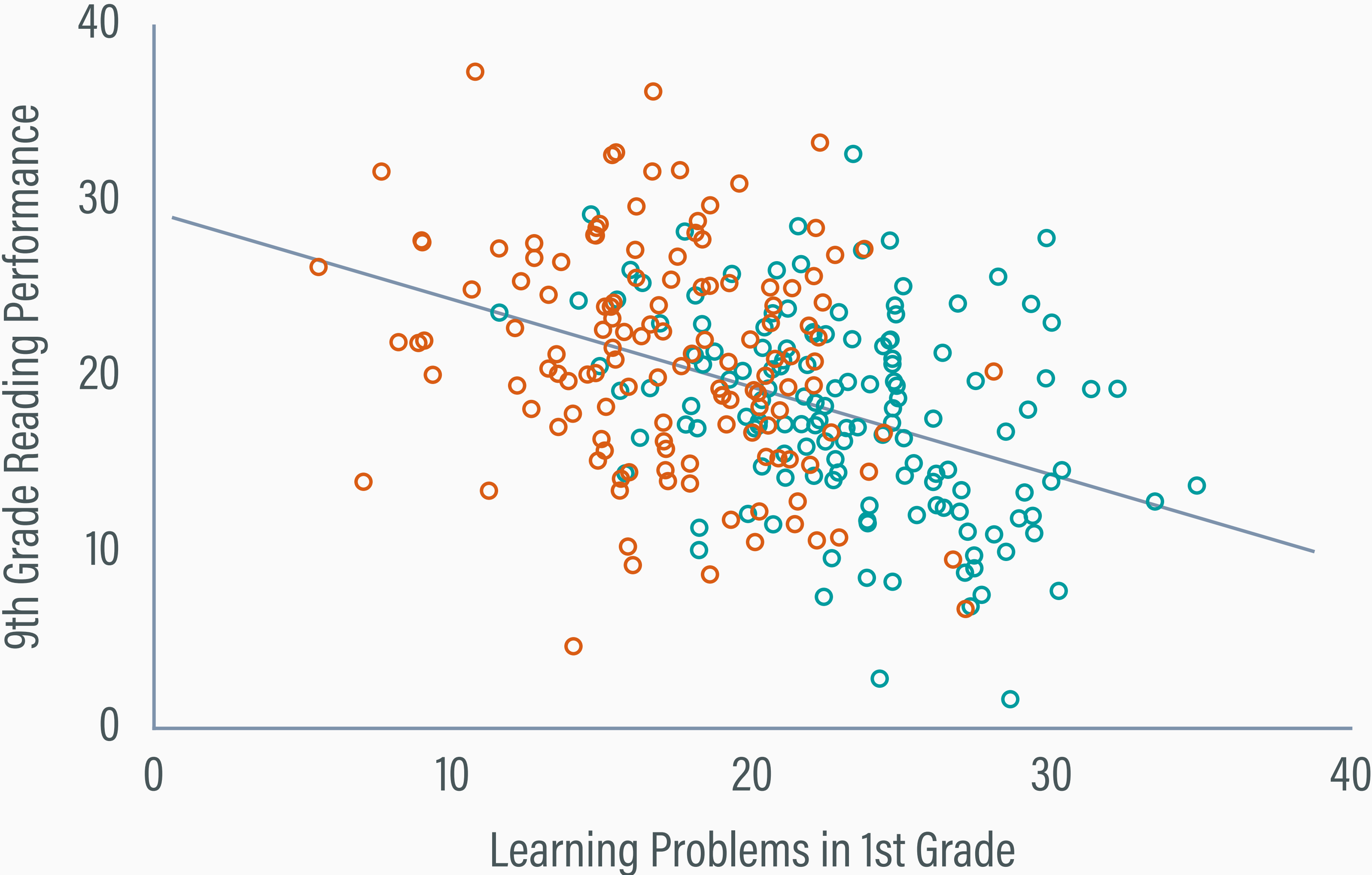


Imputation = predicted value + noise



FILLED-IN DATA FROM ONE ITERATION

- Cases with imputed scores
- Cases with complete data



BLIMP STUDIO SCRIPT

DATA: reading.dat;

VARIABLES: id atrisk lnrprob1 read1 read9;

MISSING: 999;

MODEL: read9 ~ lnrprob1;

BURN: 10000;

ITER: 10000;

SEED: 90291;

MODEL DETAILS

DATA: reading.dat;

VARIABLES: id atrisk lnrprob1 read1 read9;

MISSING: 999;

MODEL: read9 ~ lnrprob1;

regression model

BURN: 10000;

ITER: 10000;

SEED: 90291;

OUTLINE

1

Big Three Missing Data Methods

2

Introduction to MCMC Estimation

3

Application: Moderated Regression

4

Application: Structural Equation Model

FACTORED SPECIFICATION

Incomplete Predictor Model

Outcome Model

