

AN INTRODUCTION TO MISSING DATA ANALYSES WITH MCMC ESTIMATION

Craig Enders

UCLA Department of Psychology

OUTLINE

1

Big Three Missing Data Methods

2

Introduction to MCMC Estimation

3

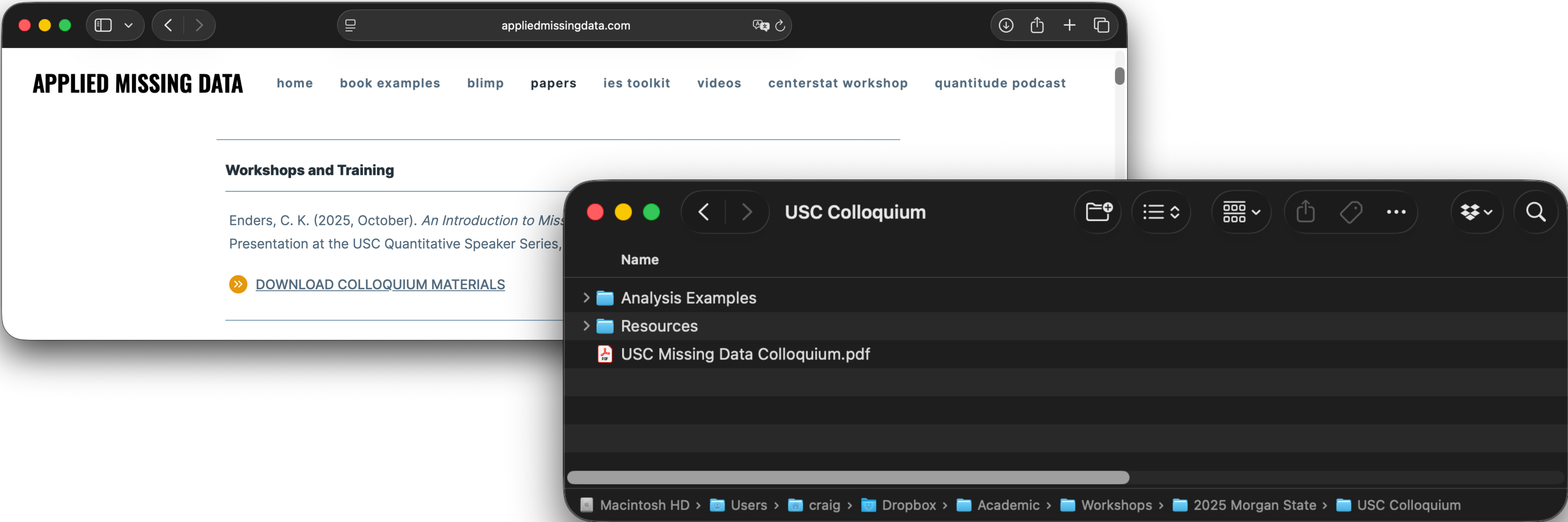
Application: Moderated Regression

4

Application: Structural Equation Model

COLLOQUIUM MATERIALS

WWW.APPLIEDMISSINGDATA.COM/BLIMP-PAPERS



WWW.APPLIEDMISSINGDATA.COM/IES-TOOLKIT

DEALING WITH MISSING DATA IN EDUCATIONAL RESEARCH

METHODOLOGICAL INNOVATIONS AND
CONTEMPORARY RECOMMENDATIONS

CRAIG K. ENDERS, PHD

DEALING WITH MISSING DATA IN EDUCATIONAL RESEARCH

SOFTWARE TUTORIALS

CRAIG K. ENDERS
REMUS MITCHELL
TONATIUH XOCHIHUA-TLECUITL
MICHAEL P. WOLLER

WWW.APPLIEDMISSINGDATA.COM/VIDEOS

BLIMP VIDEO SERIES

The Blimp video series and corresponding YouTube channel provide researchers with training for using the Blimp software. Each video provides a short, step-by-step tutorial that walks viewers through a particular aspect of a missing data analysis. Check back, as new videos are continually added.

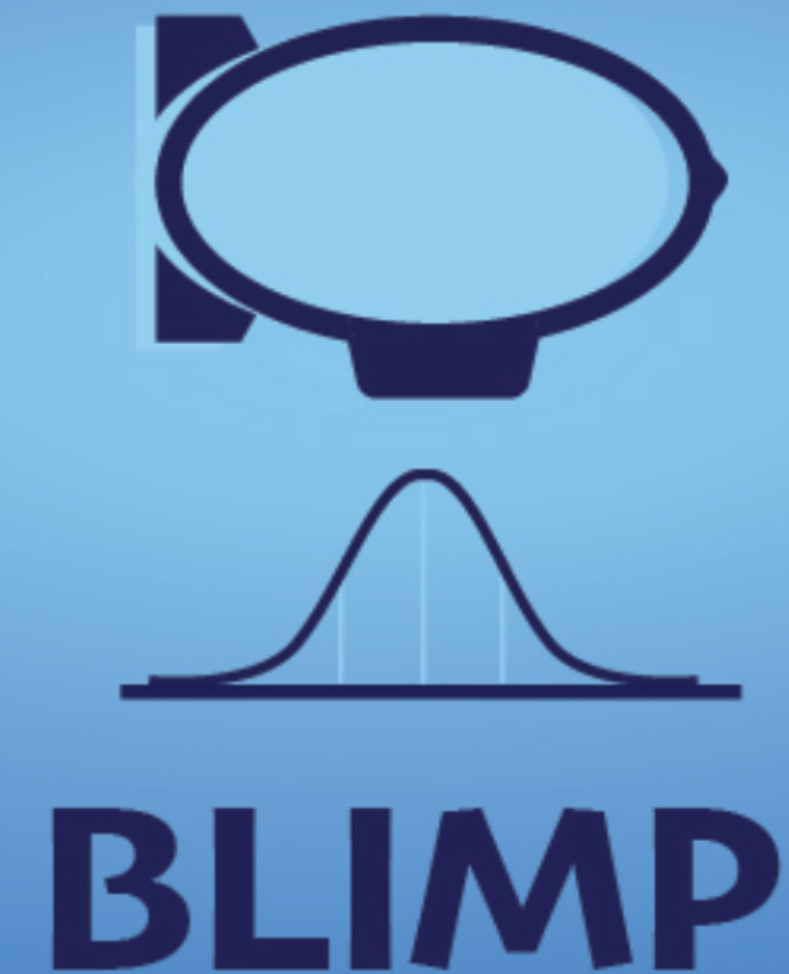
WWW.APPLIEDMISSINGDATA.COM/BLIMP

BLIMP 3.0

Blimp 3 offers powerful latent variable modeling and imputation for incomplete data sets with up to three levels. Blimp's unique Bayesian computational architecture allows easy specification of complex analyses that are difficult or impossible to fit in other software packages.

[Download Now](#)

[User's Guide](#)



OUTLINE

1

Big Three Missing Data Methods

2

Introduction to MCMC Estimation

3

Application: Moderated Regression

4

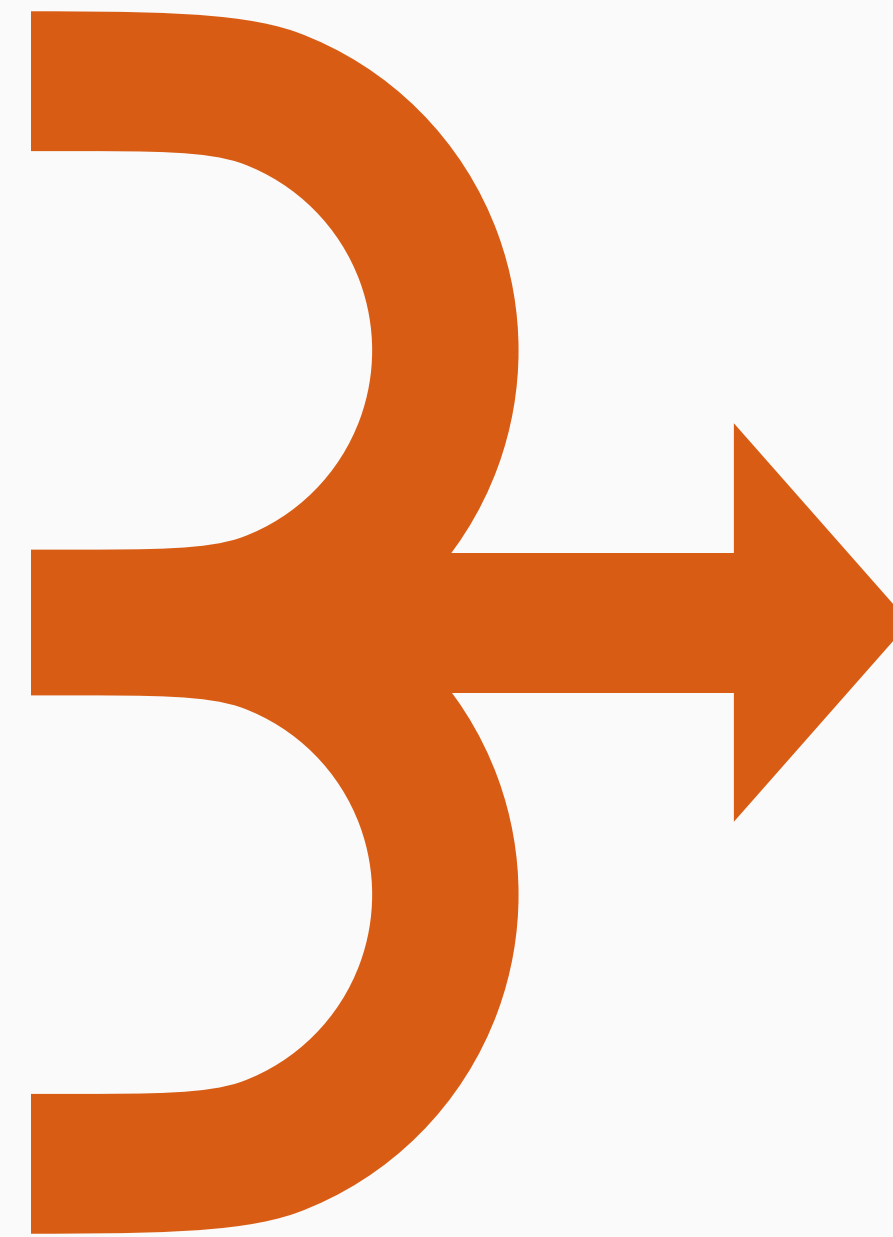
Application: Structural Equation Model

MODERN MISSING DATA METHODS

Maximum likelihood

Multiple imputation

Bayesian MCMC estimation



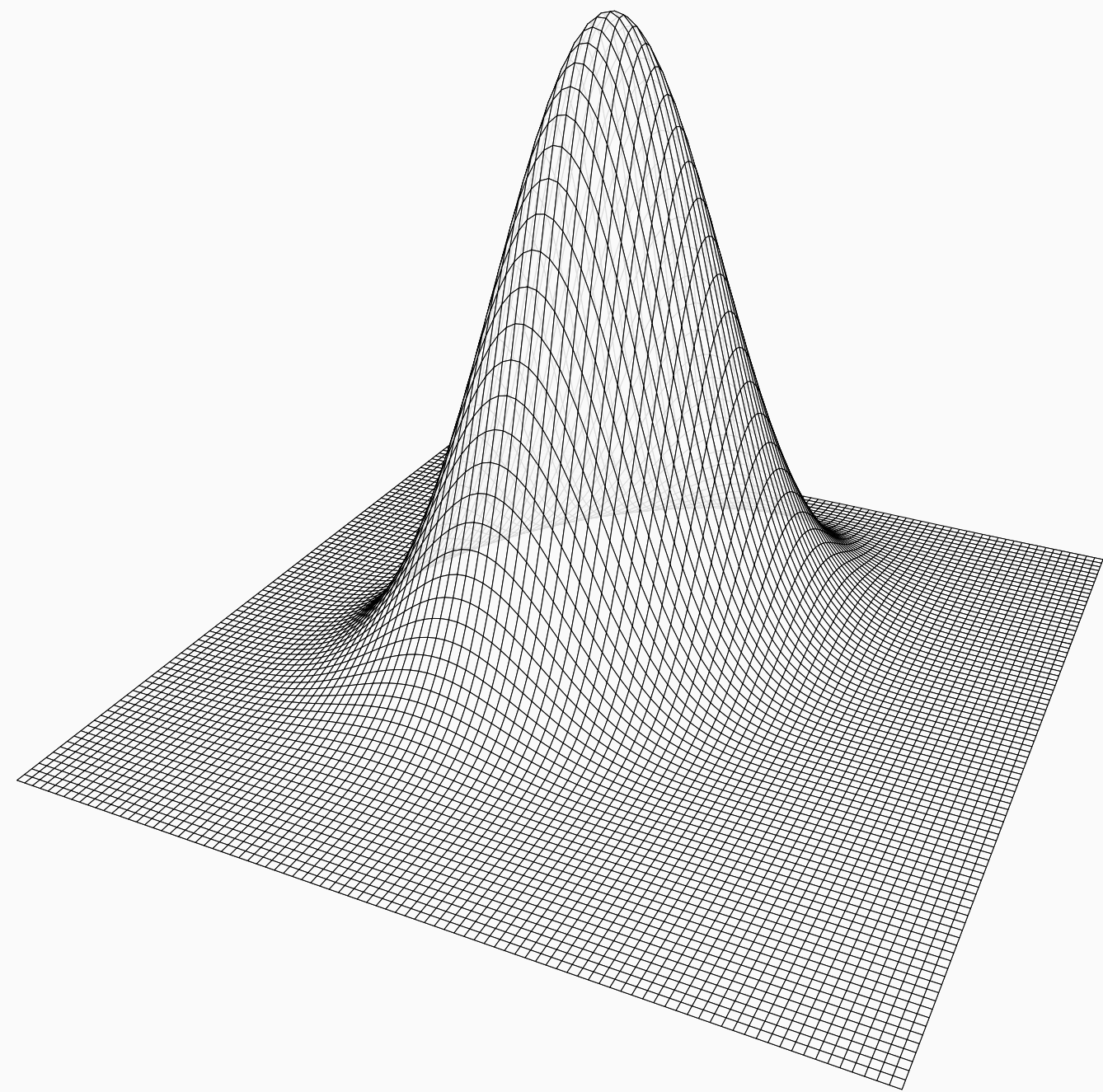
**the
Big
Three**

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

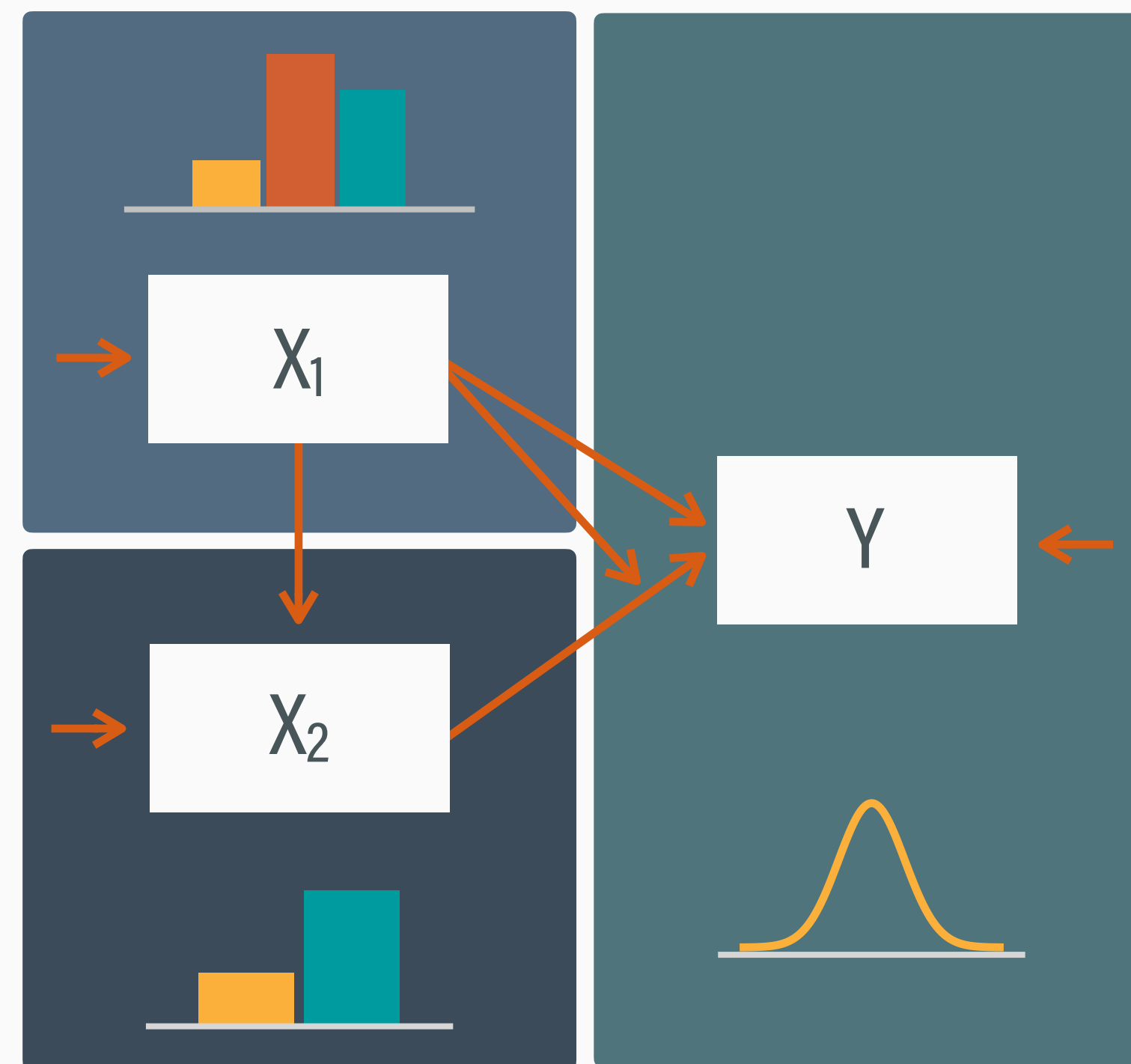


- Classic versions of the Big Three often assume multivariate normality
- Includes FIML and multiple imputation
- Multivariate assumptions restrict the range of problems that can be addressed

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

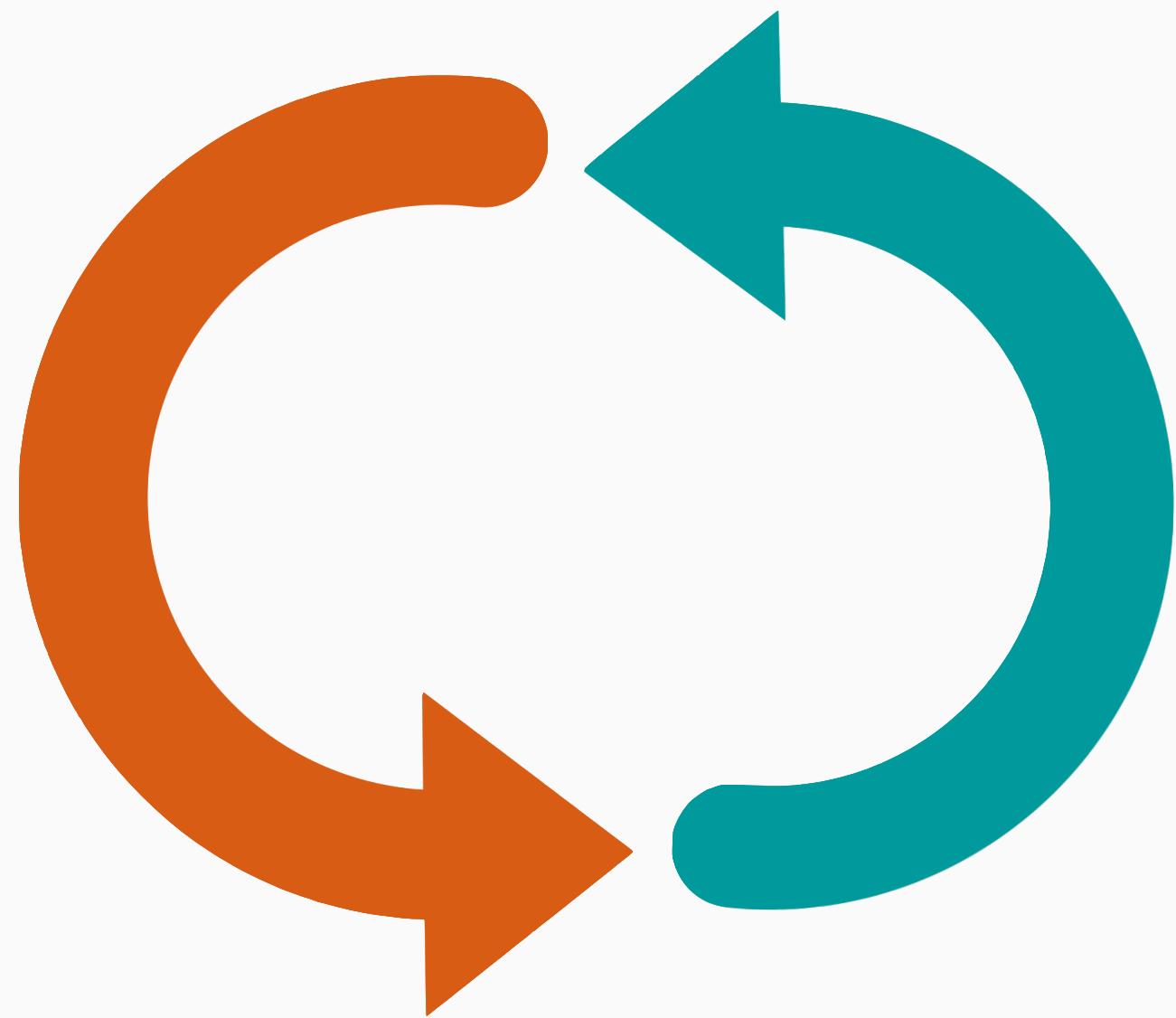
- Study that seeks to determine whether learning problem in 1st grade predict 9th grade reading achievement in middle school

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

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

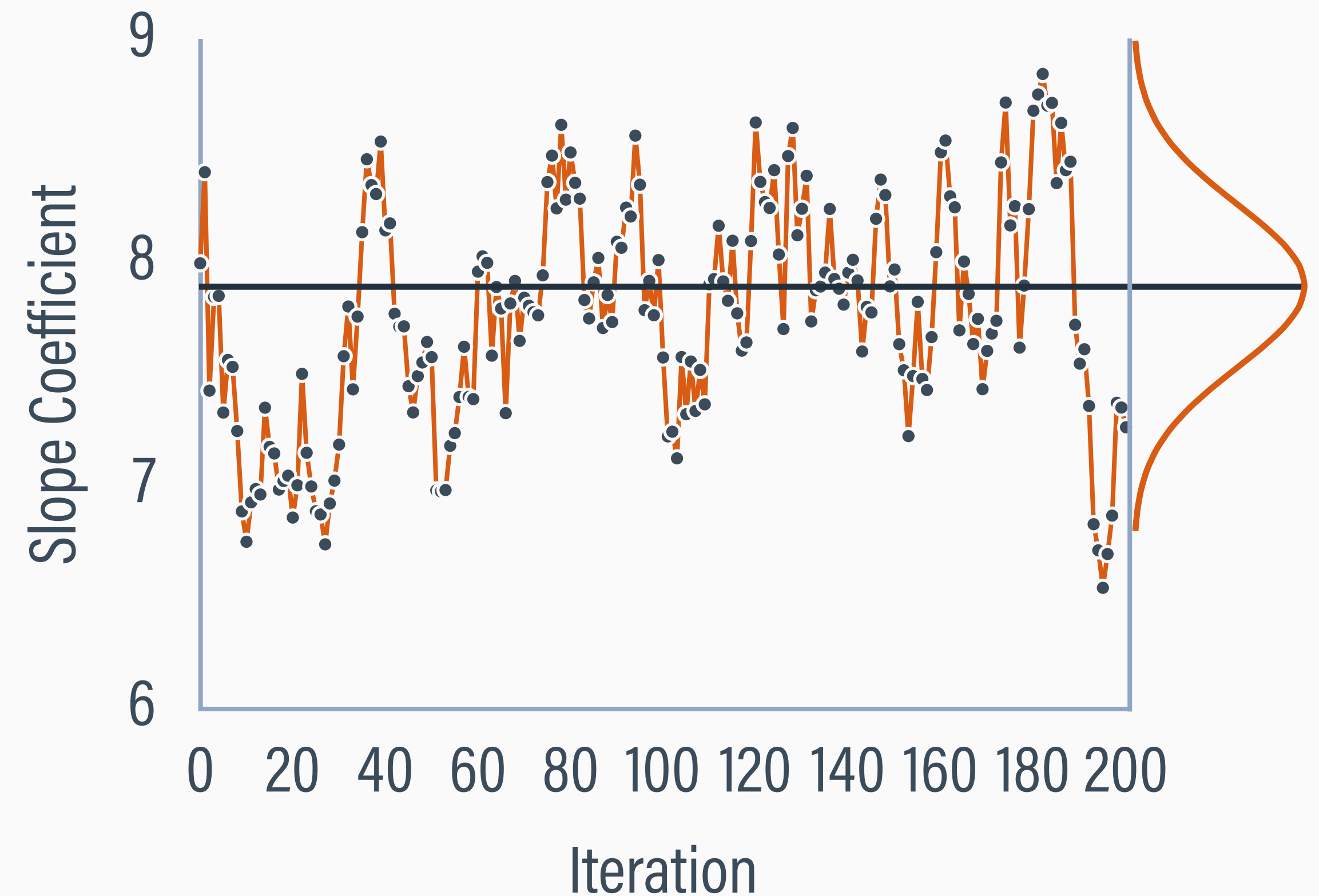
- » Estimate model parameters from the filled-in data
- » Impute missing values based on the model parameters

Repeat

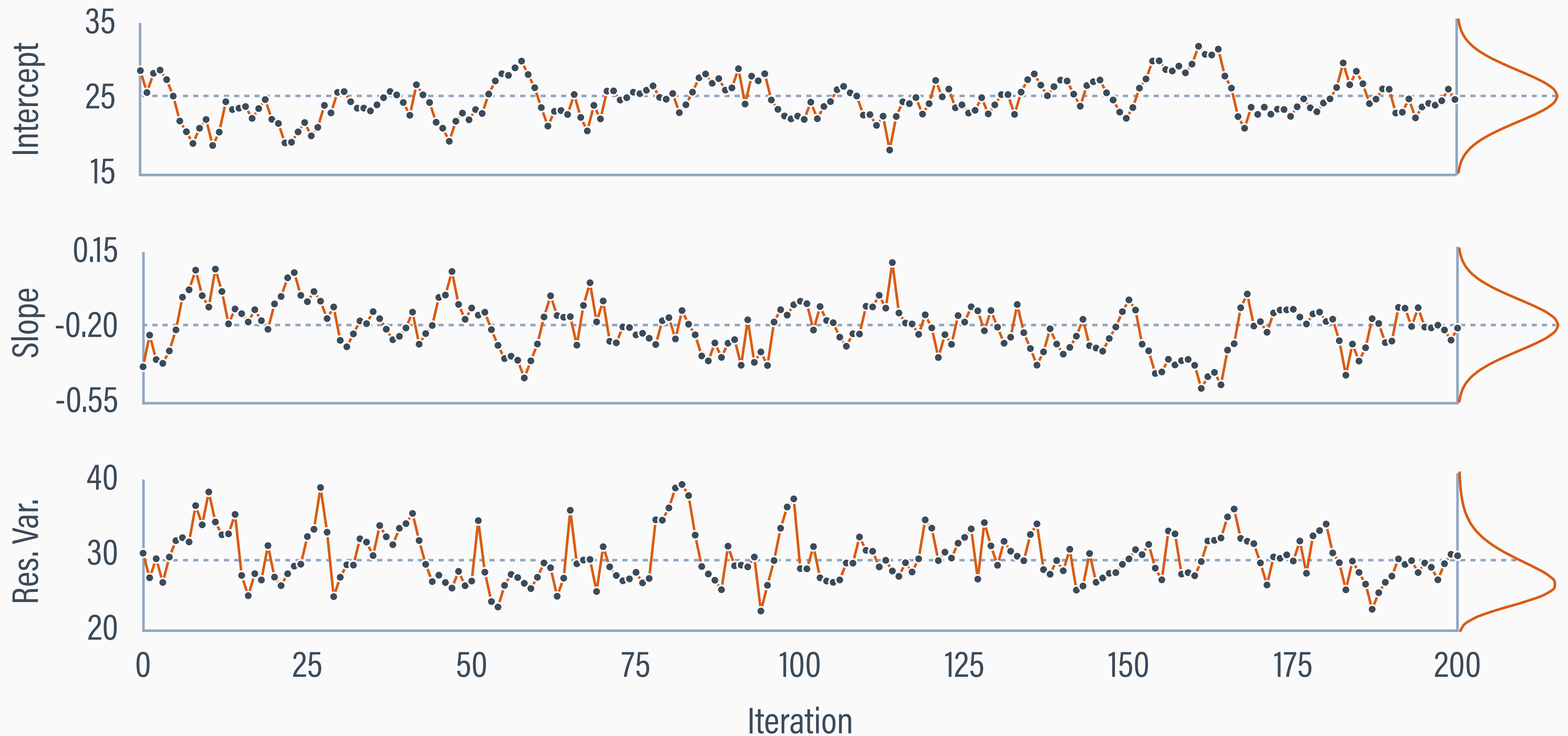
Summarize model parameters

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

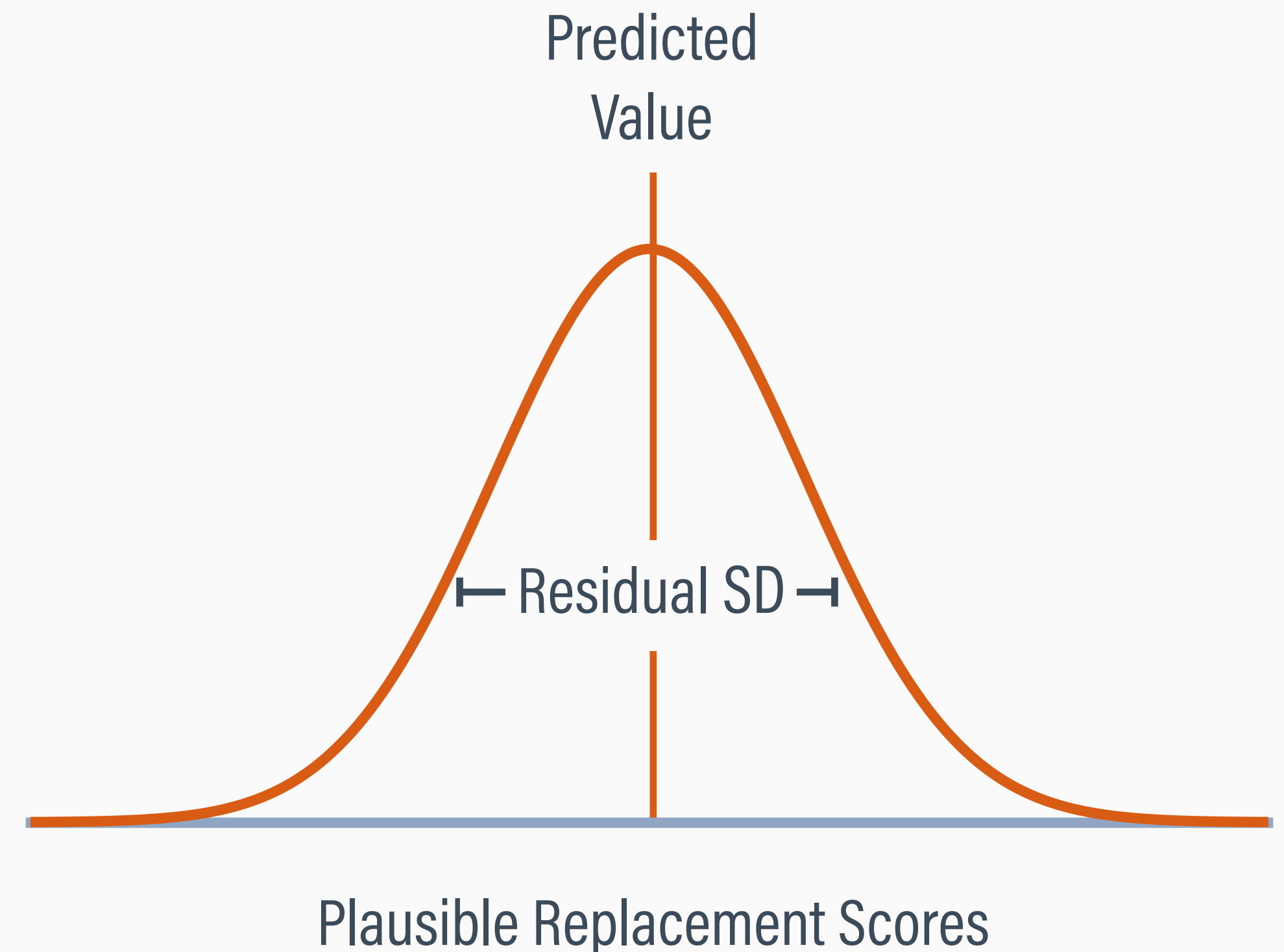


PARAMETERS FROM 200 MCMC CYCLES

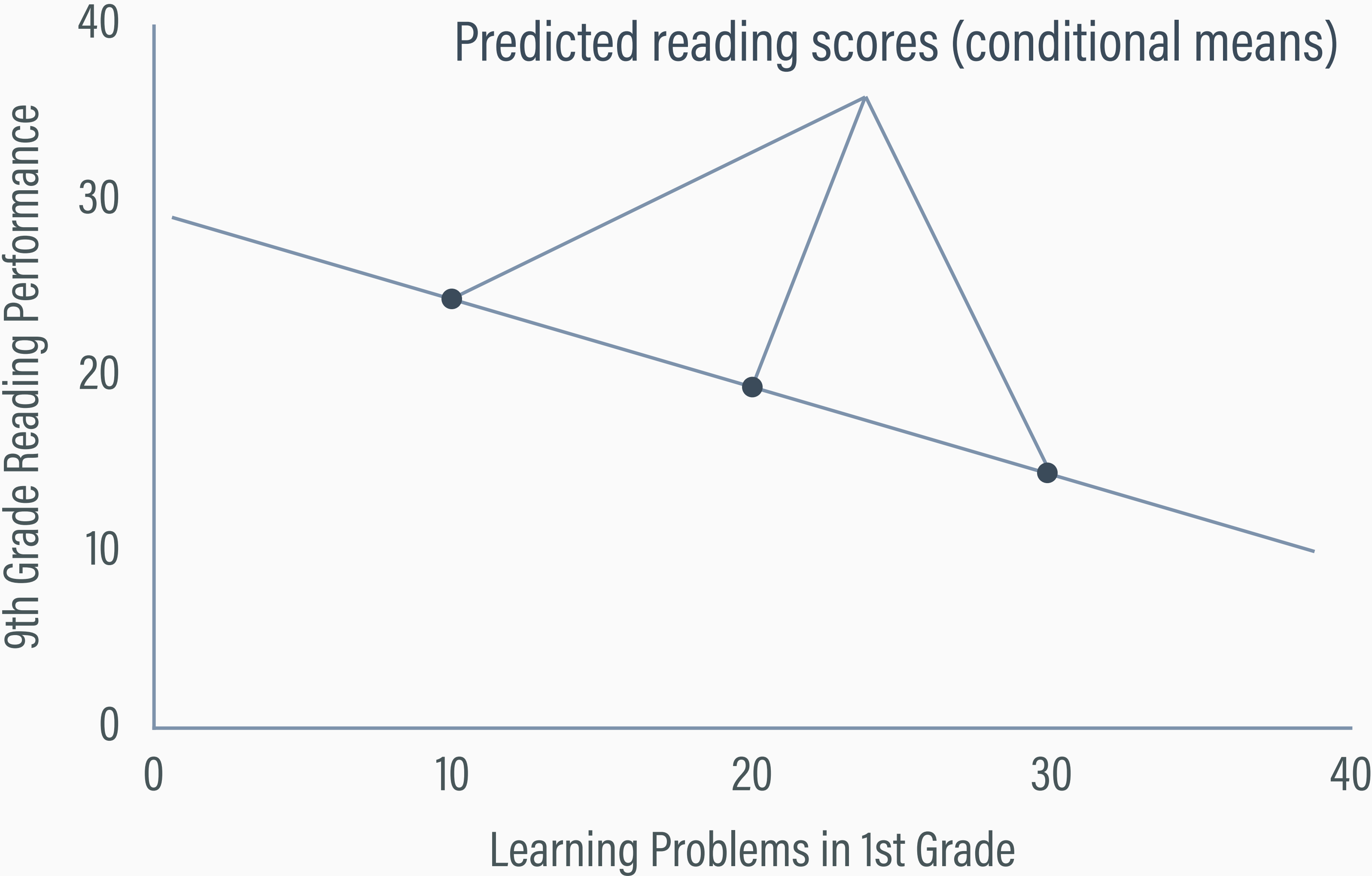


MISSING DATA IMPUTATION

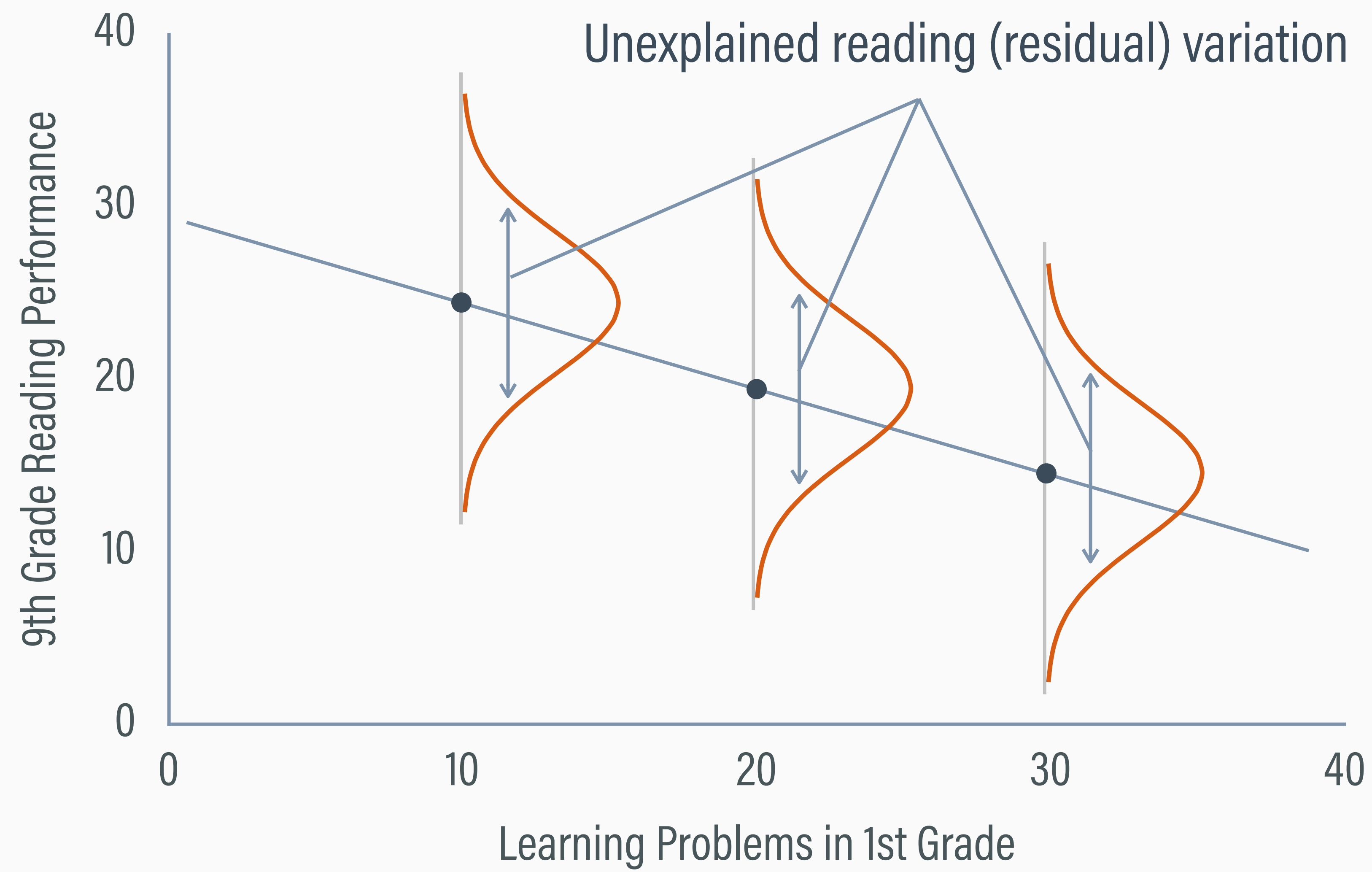
- Missing values are sampled from a distribution of plausible replacement scores
- The model parameters at each iteration define the distribution's shape
- Imputation = predicted value + computer-generated residual term



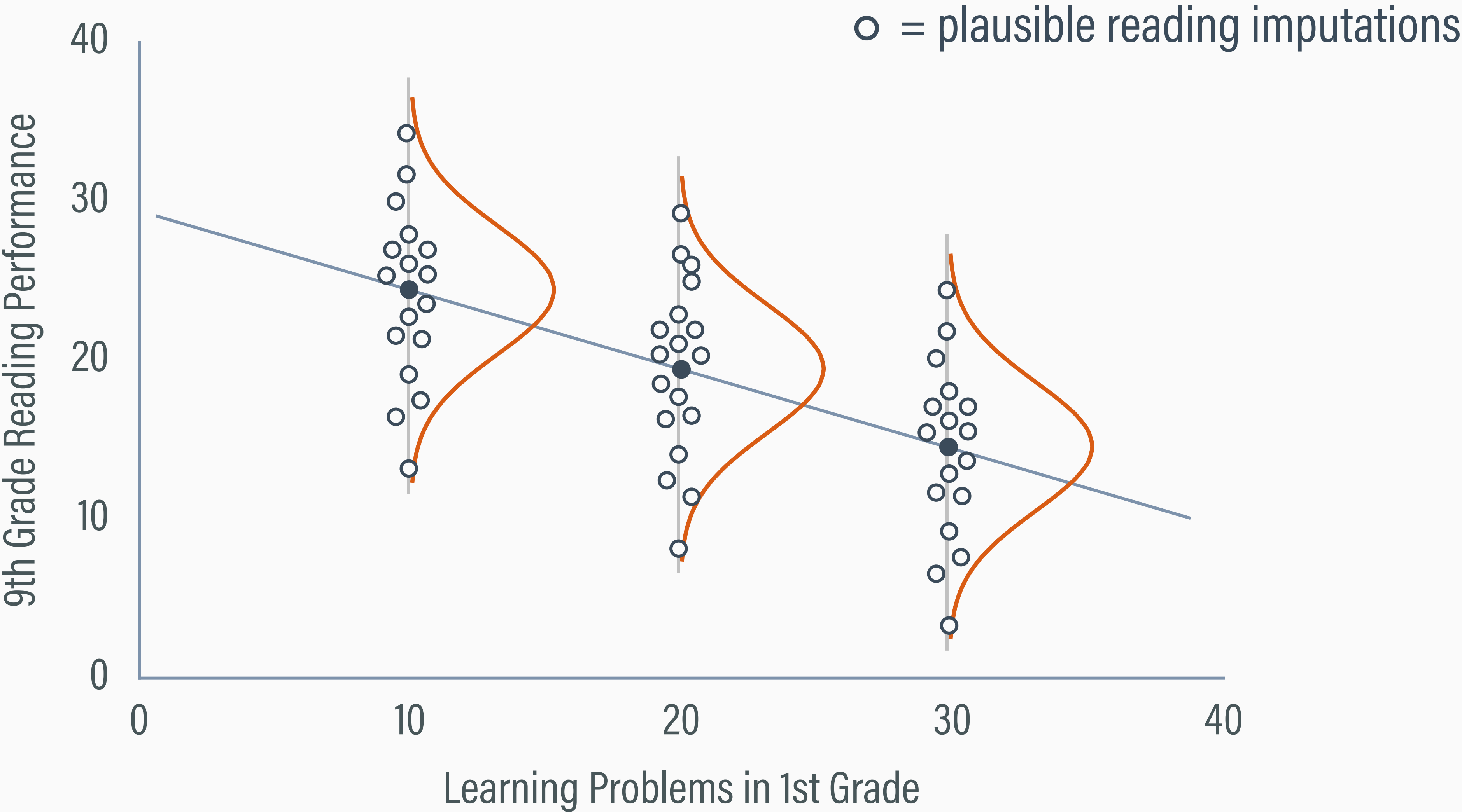
PREDICTED VALUES



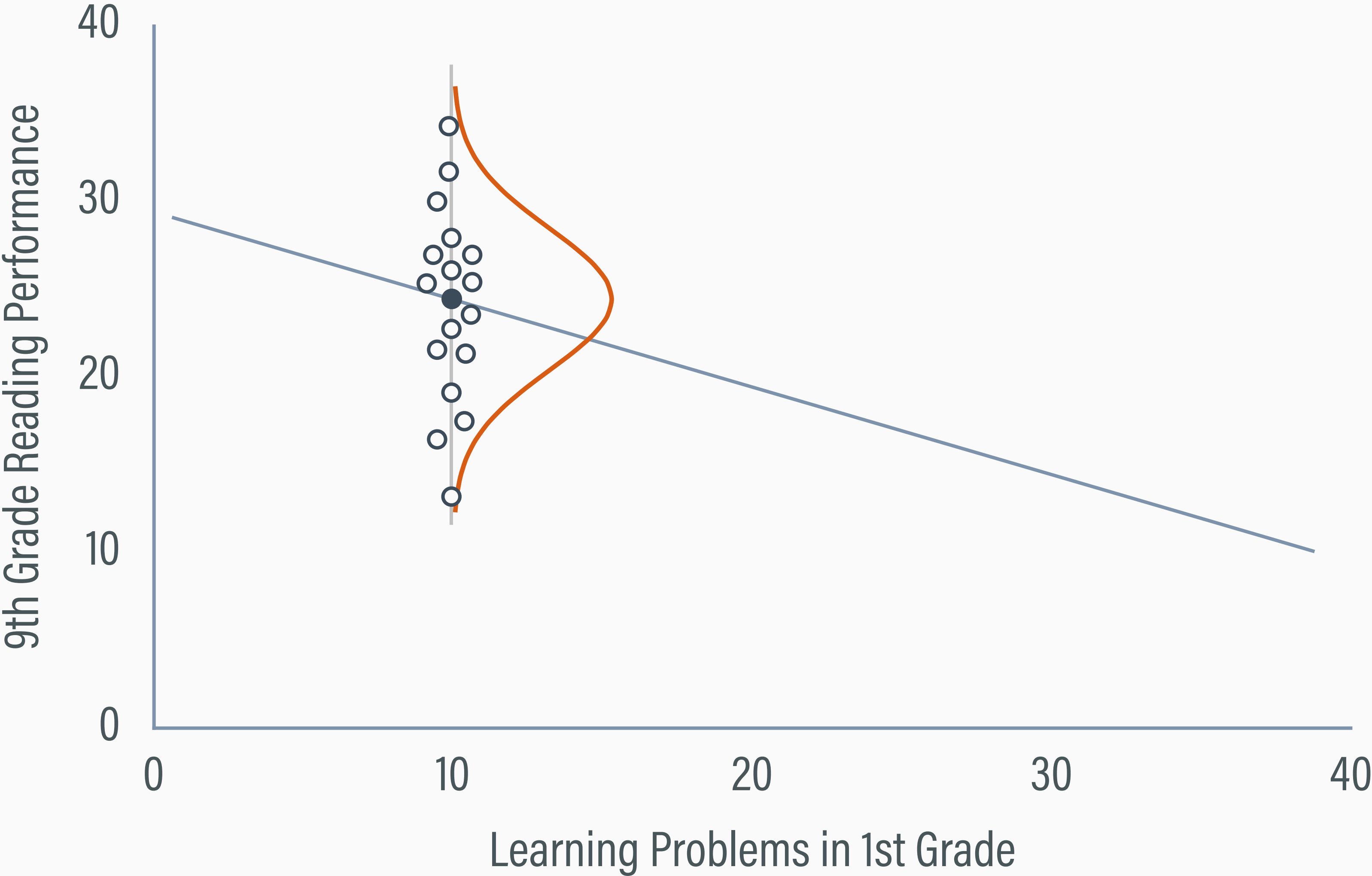
RESIDUAL VARIATION



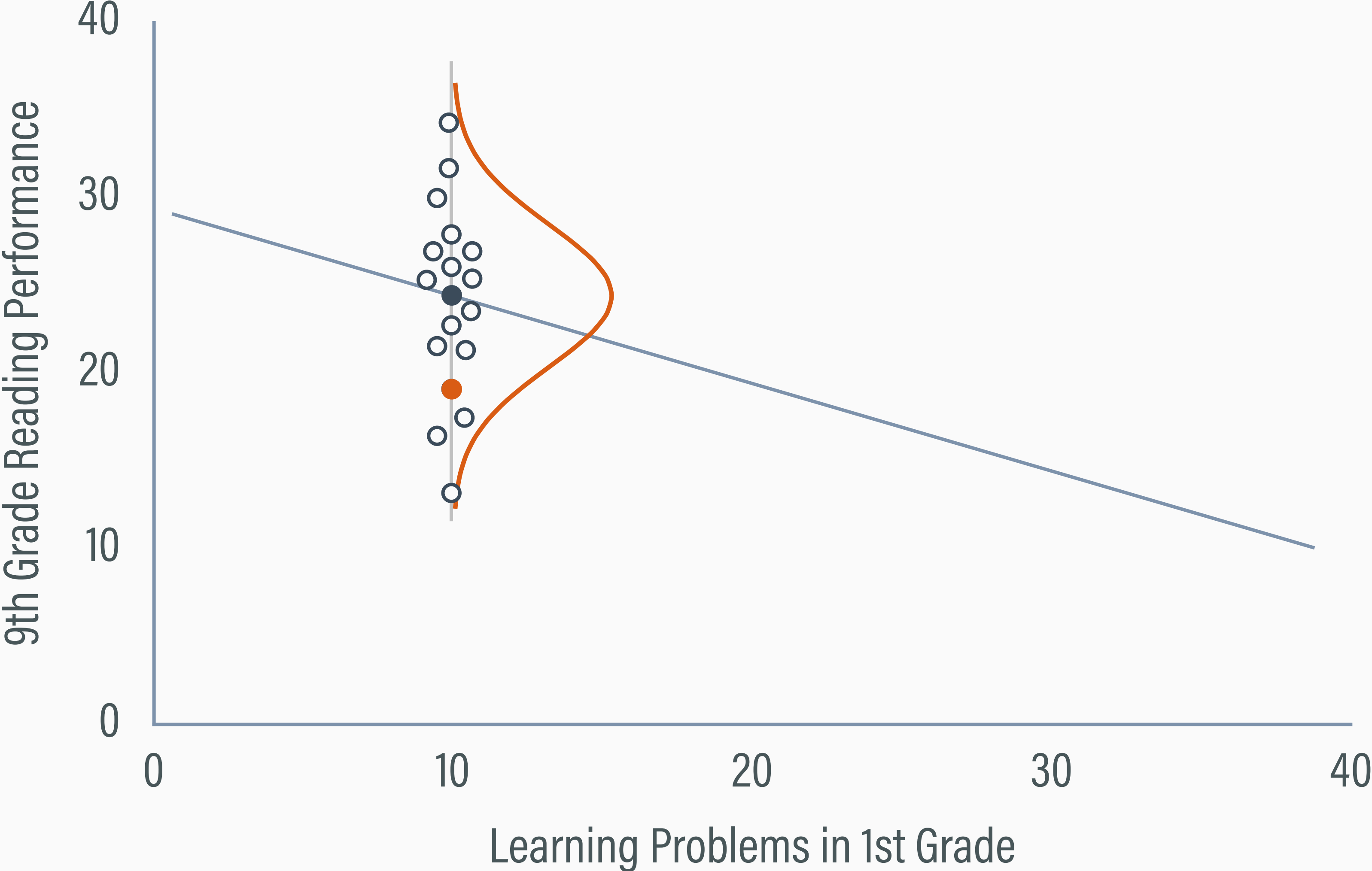
DISTRIBUTIONS OF IMPUTATIONS



IMPUTATION FOR LOW LEARNING PROBLEMS



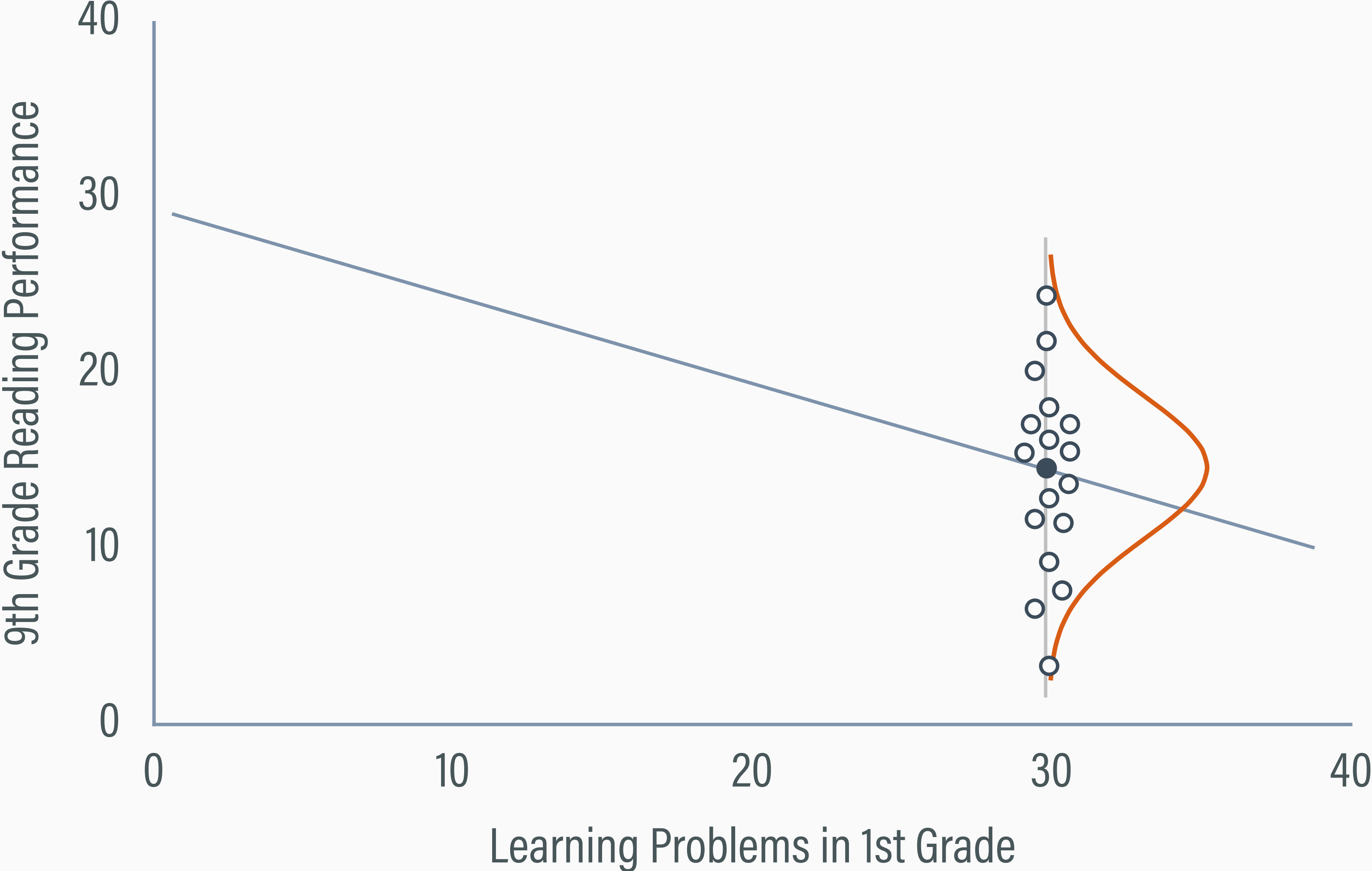
SAMPLE IMPUTATION AT RANDOM



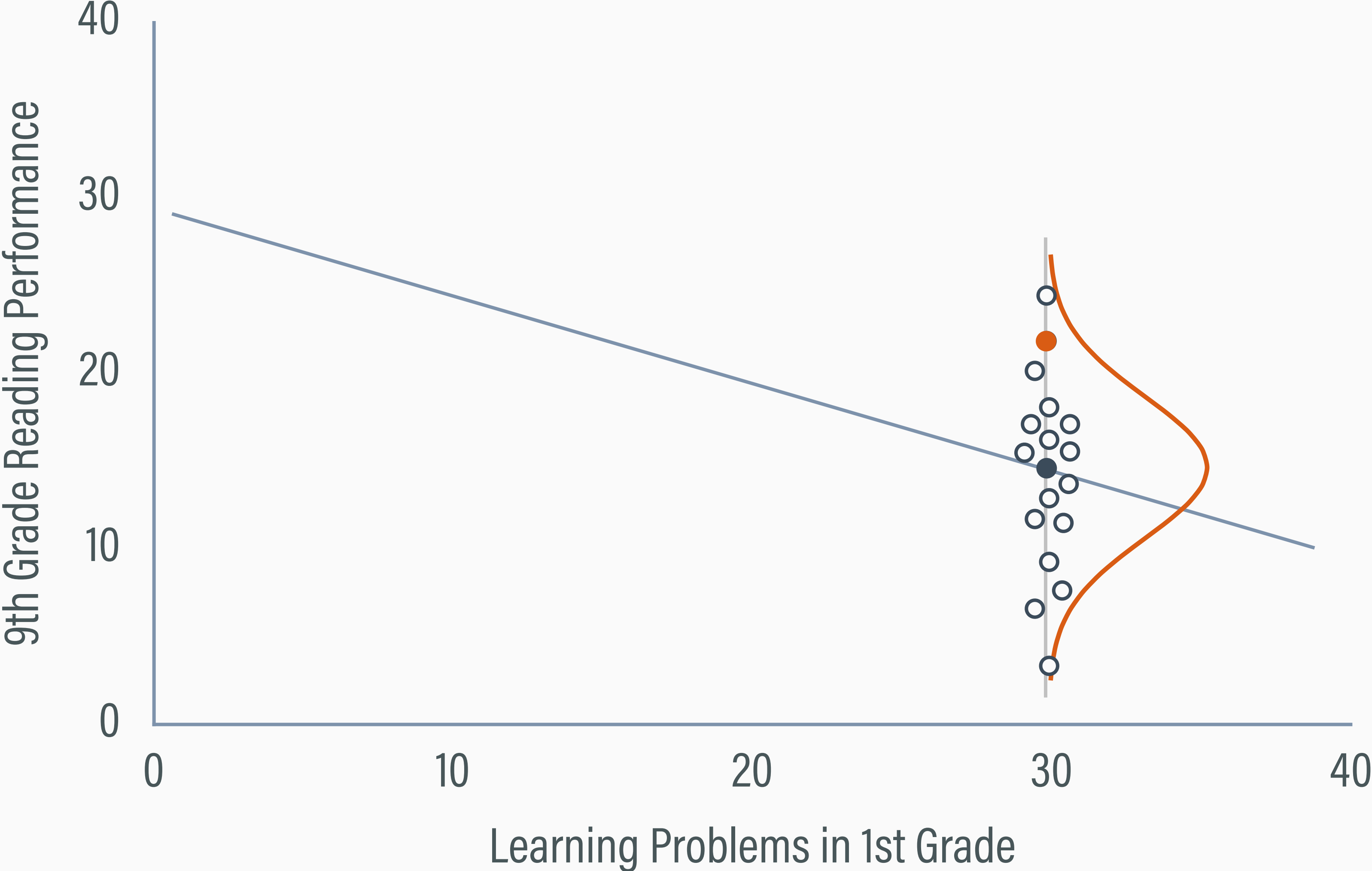
IMPUTATION = PREDICTION + NOISE



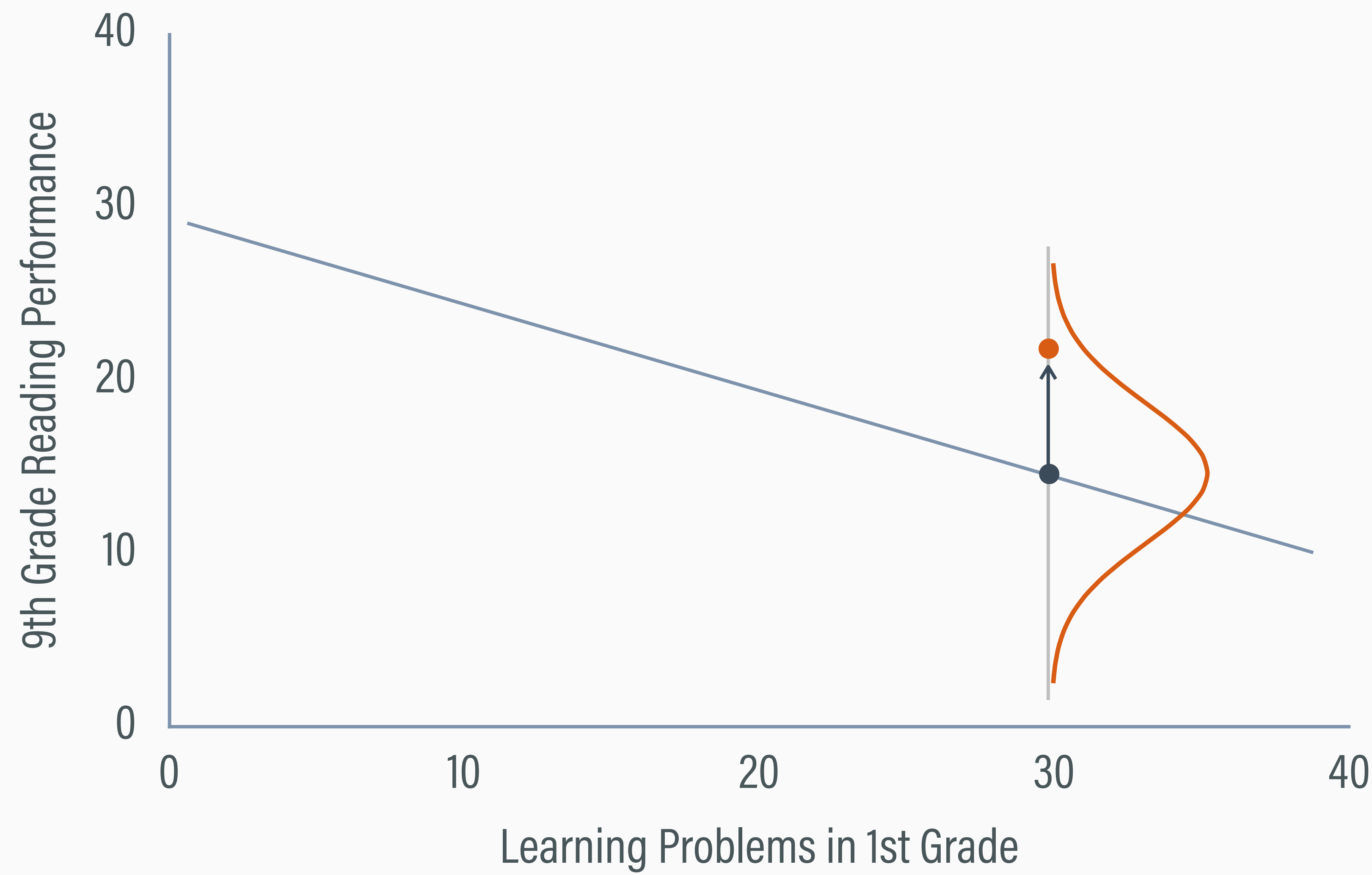
IMPUTATION FOR HIGH LEARNING PROBLEMS



SAMPLE IMPUTATION AT RANDOM



IMPUTATION = PREDICTION + NOISE

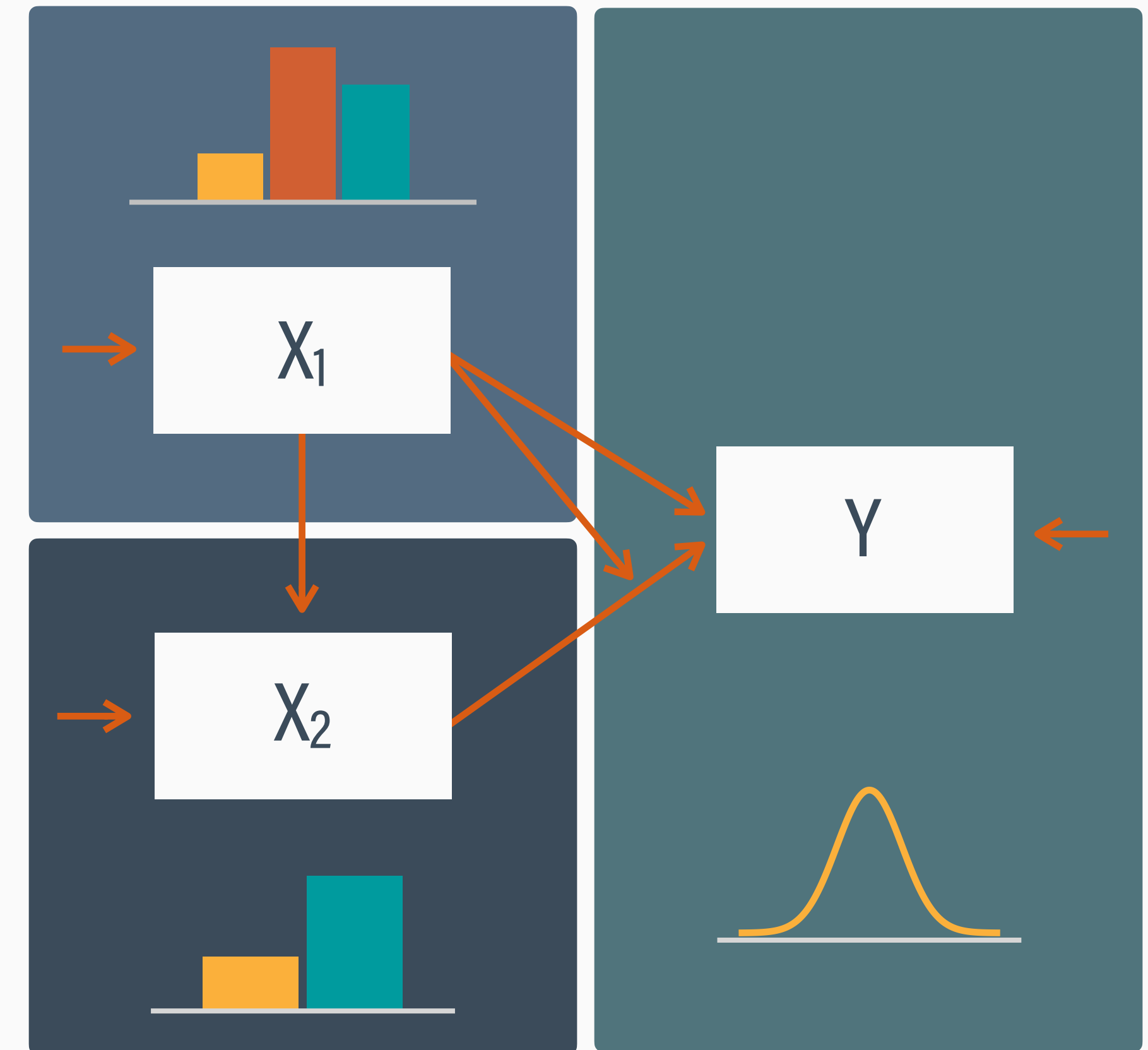


INCOMPLETE PREDICTORS

- 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

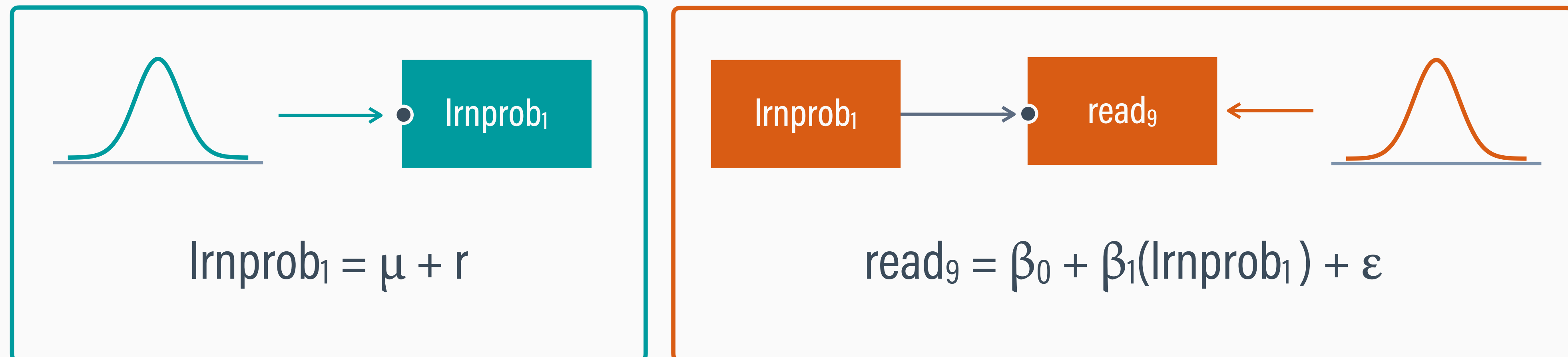
FACTORED REGRESSION SPECIFICATIONS

- 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)

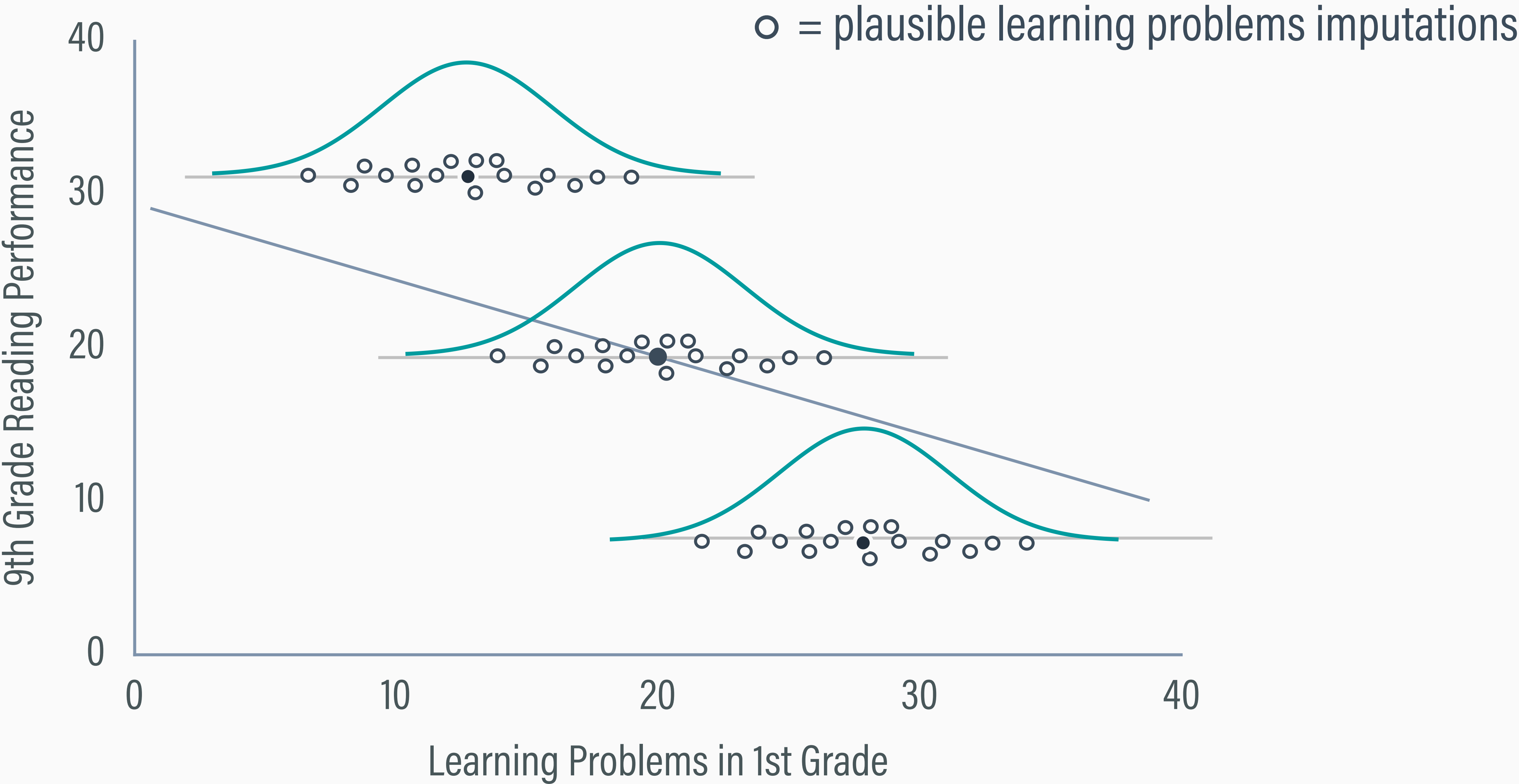


FACTORED SPECIFICATION

- 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

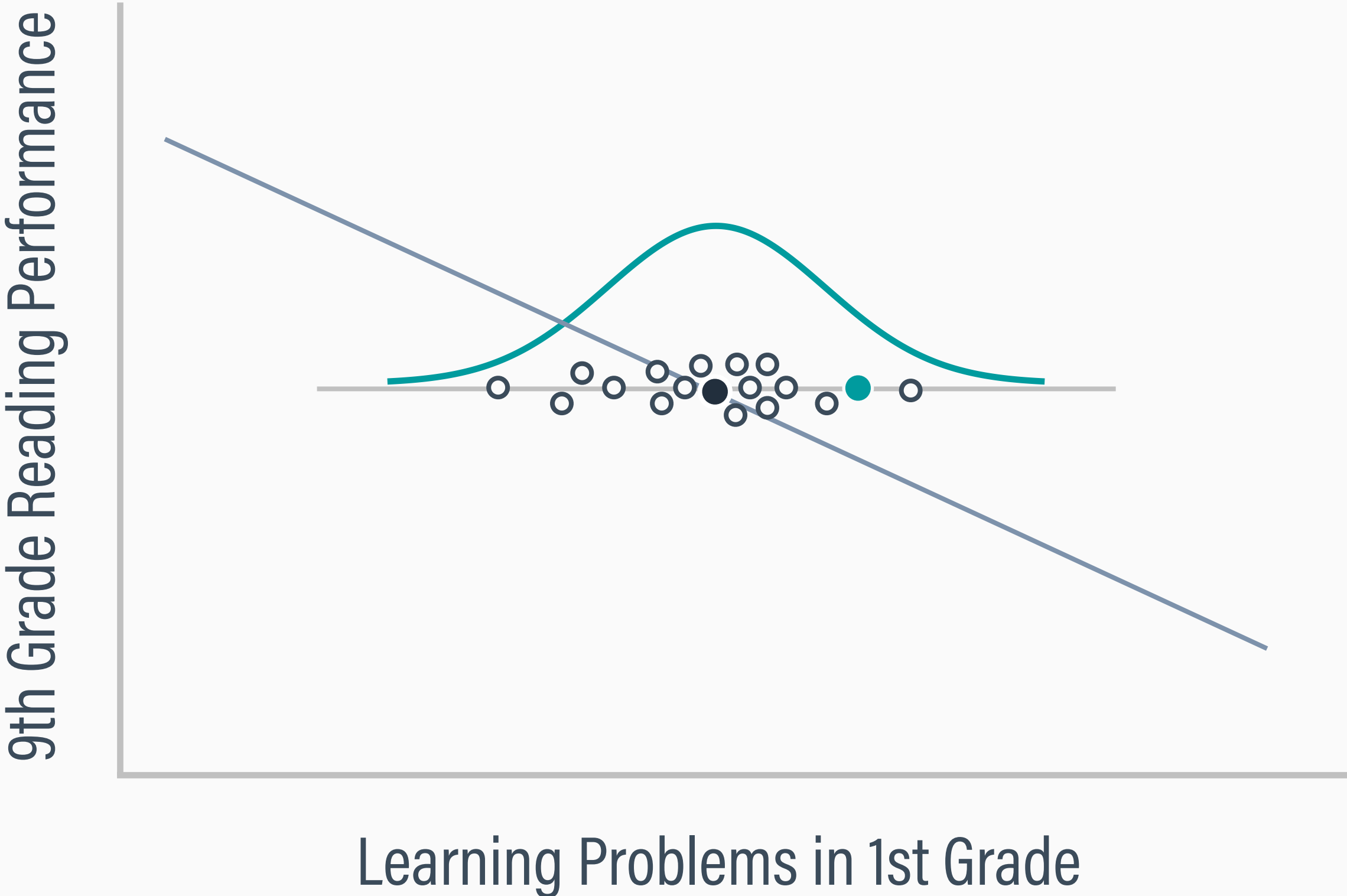


DISTRIBUTIONS OF IMPUTATIONS

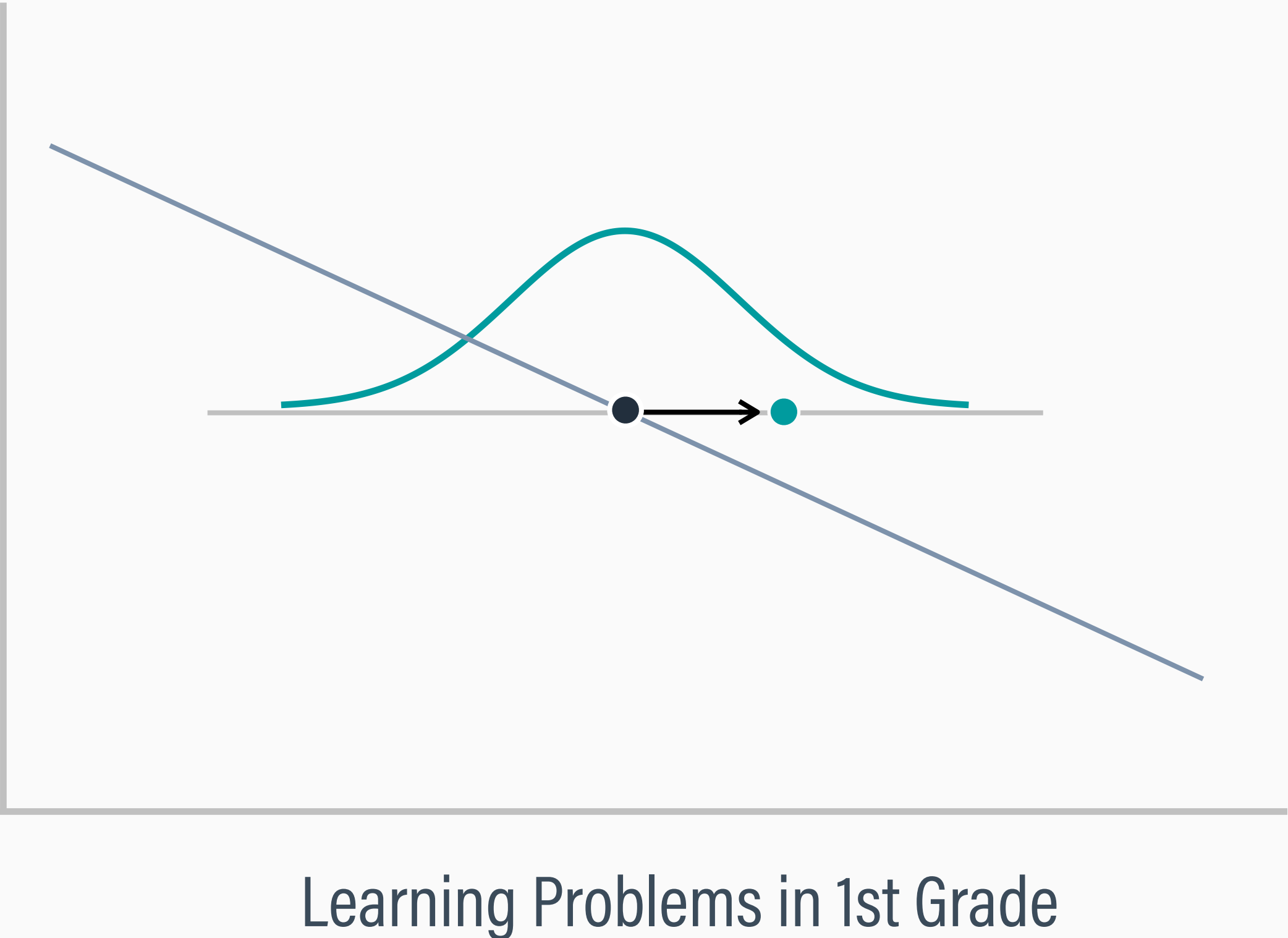


IMPUTATION EXAMPLE

Sample imputation at random

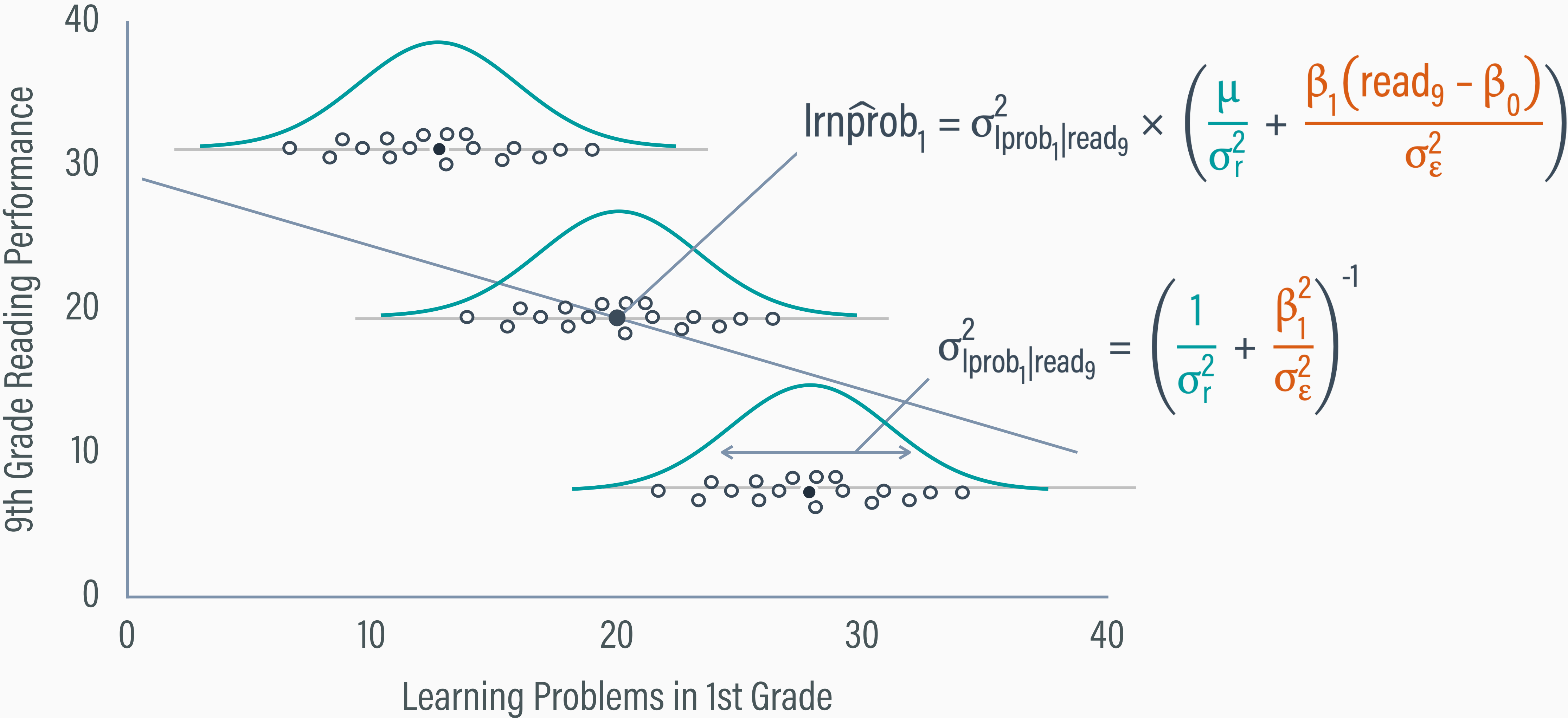


Imputation = predicted value + noise



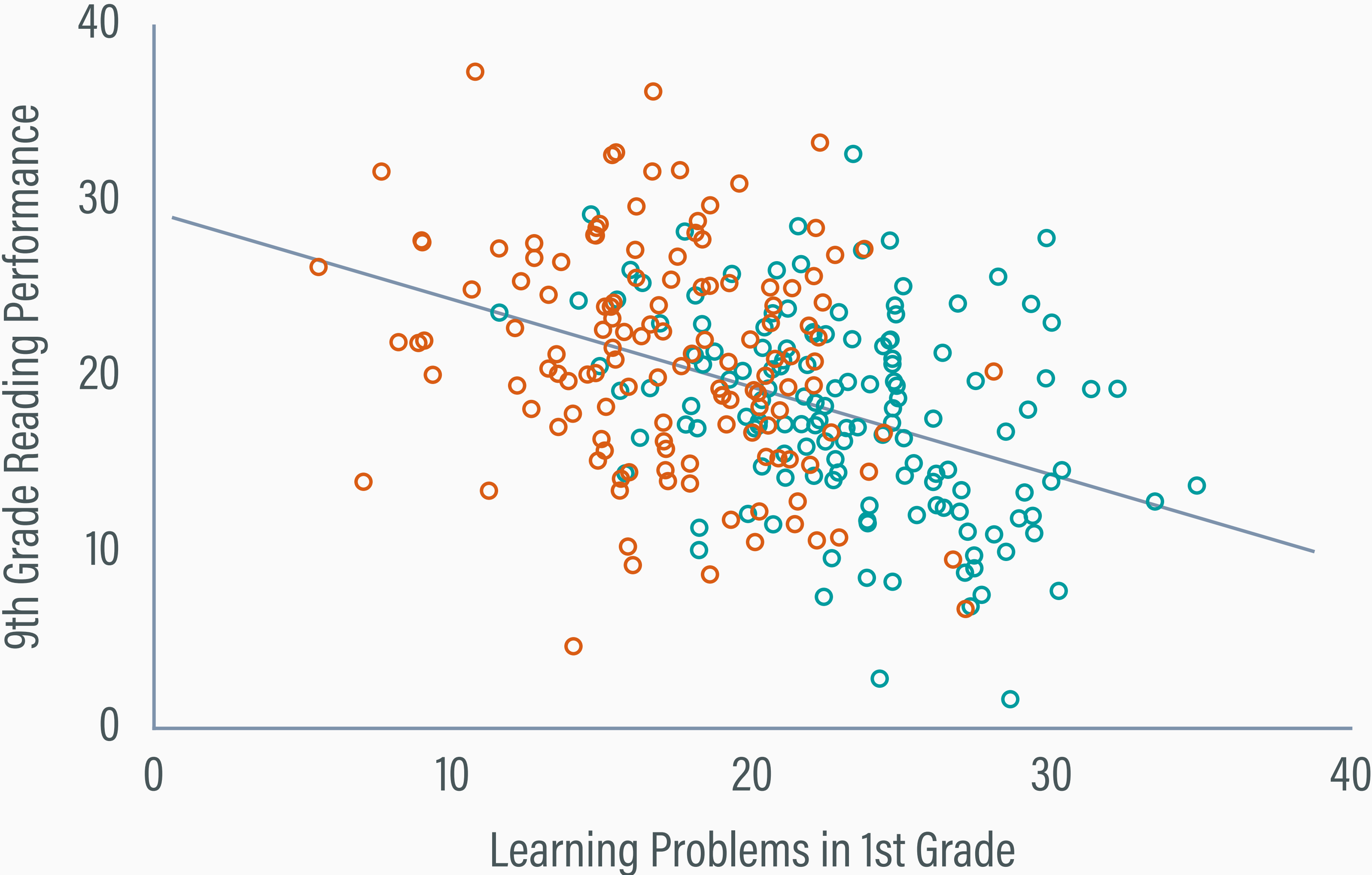
PREDICTED VALUES AND VARIATION

Multiple sets of model parameters define the mean and spread of the imputations



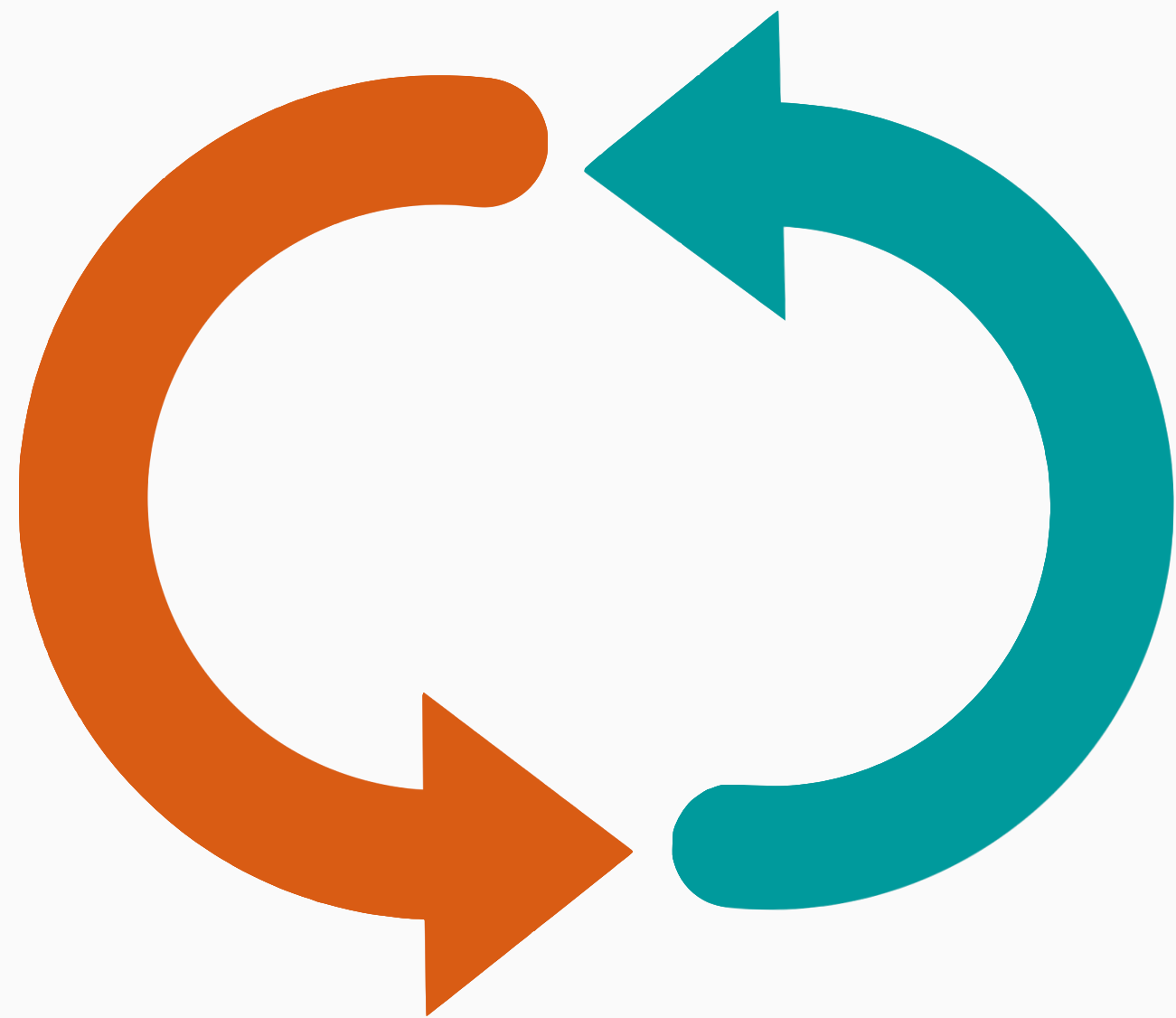
FILLED-IN DATA FROM ONE ITERATION

- Cases with imputed scores
- Cases with complete data



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

BLIMP STUDIO SCRIPT

DATA: reading.dat;

VARIABLES: id atrisk lnrprob1 read1 read9;

MISSING: 999;

MODEL: read9 ~ lnrprob1;

BURN: 10000;

ITER: 10000;

SEED: 90291;

DATA AND VARIABLES

```
DATA: reading.dat;           # data in same directory as the script
VARIABLES: id atrisk lnrprob1 read1 read9;  # name data columns
MISSING: 999;              # missing value code

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;

COMPUTATIONAL DETAILS

DATA: reading.dat;

VARIABLES: id atrisk lnrprob1 read1 read9;

MISSING: 999;

MODEL: read9 ~ lnrprob1;

BURN: 10000;

number of warm-up iterations

ITER: 10000;

number of parameter values for the analysis

SEED: 90291;

integer seed for Monte Carlo simulation

RBLIMP SCRIPT

```
model1 <- rblimp(  
  data = reading,  
  model = 'read9 ~ lnp1',  
  seed = 90291,  
  burn = 10000,  
  iter = 10000)  
output(model1)
```

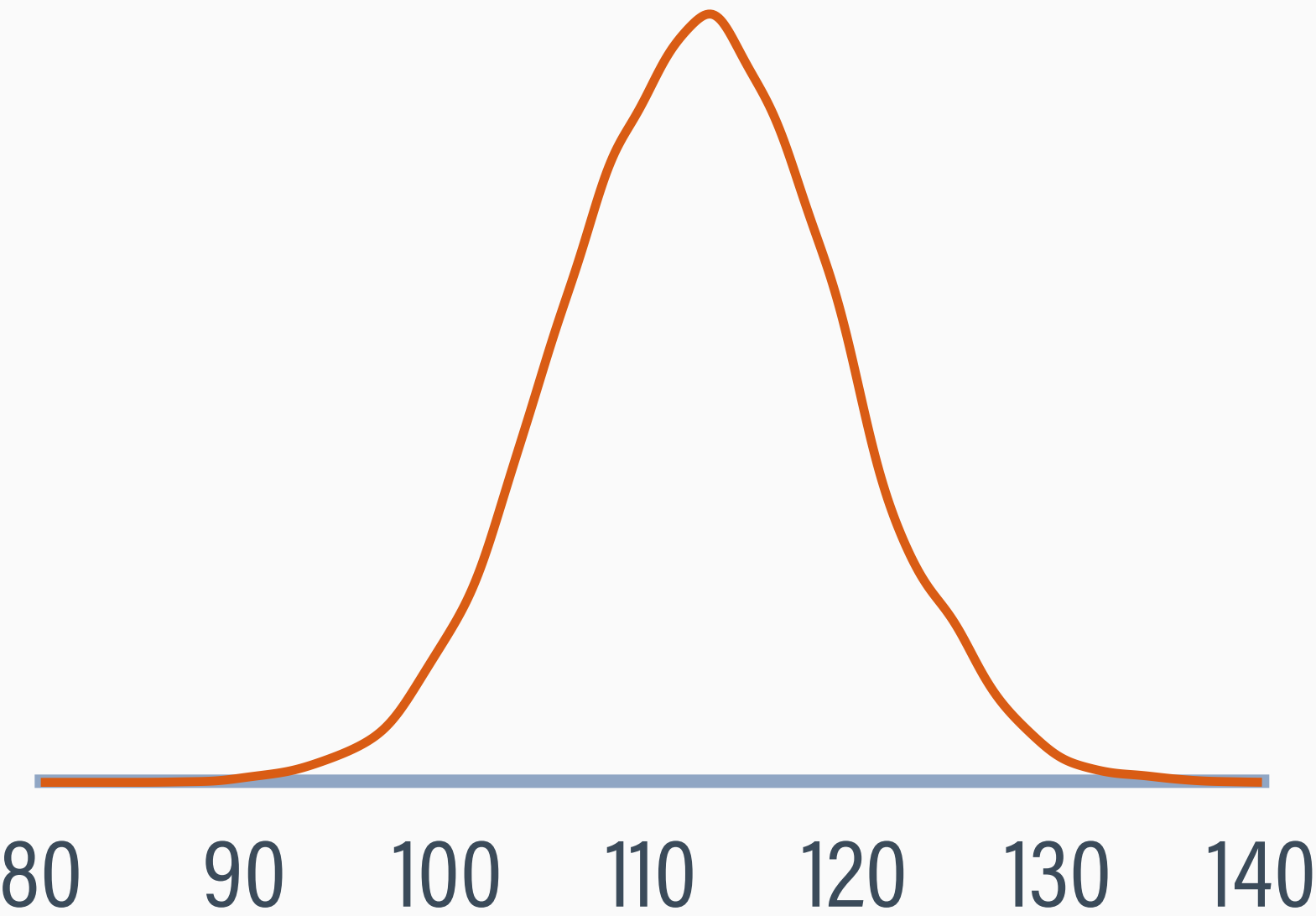
R data frame
regression model
integer seed for Monte Carlo simulation
number of warm-up iterations
number of parameter values for the analysis
print output

PARAMETER (POSTERIOR) DISTRIBUTIONS

Median = 112.48

Std. Dev. = 6.81

95% CI = (99.09, 125.87)

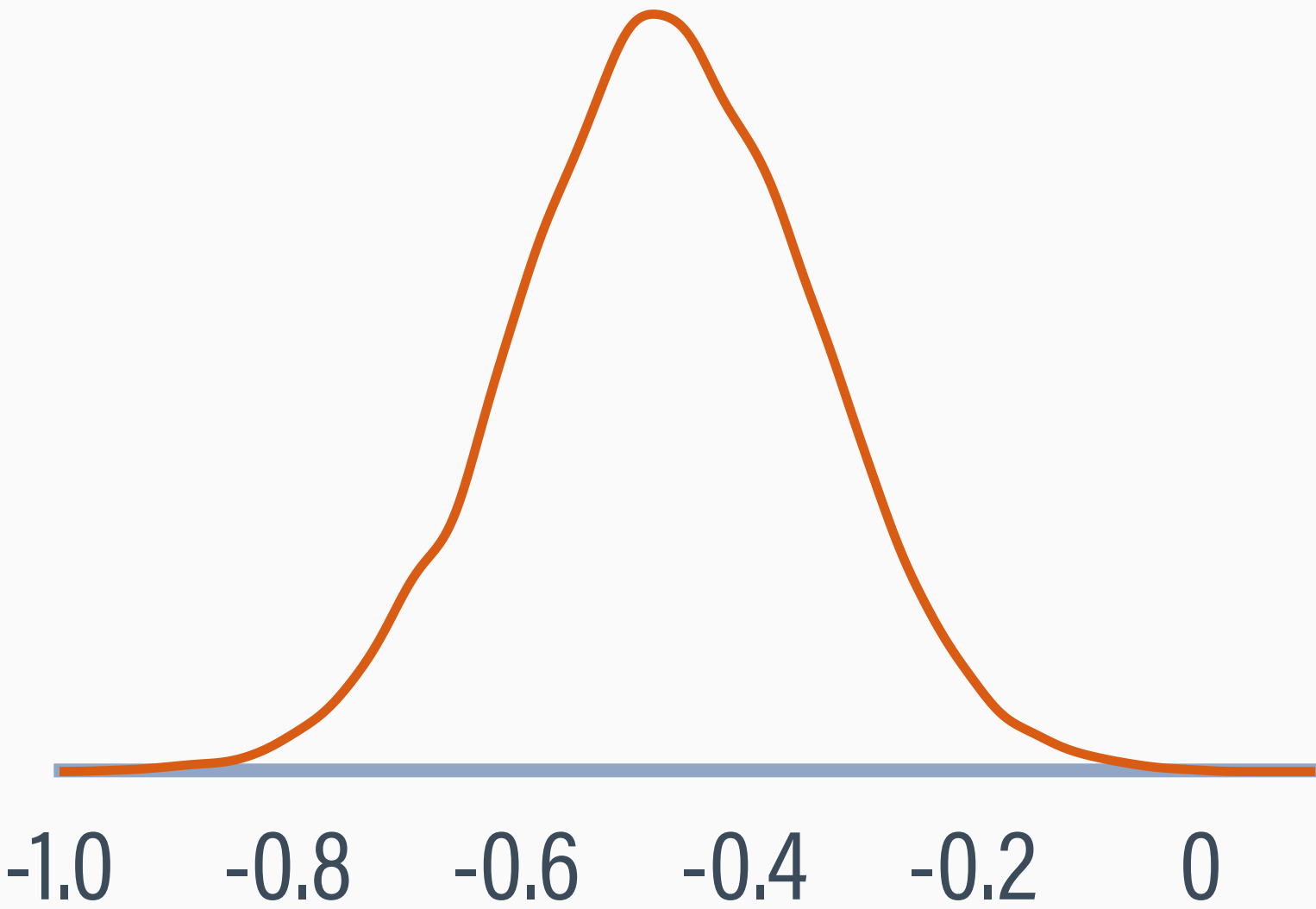


Intercept

Median = -0.47

Std. Dev. = 0.13

95% CI = (-0.73, -0.21)

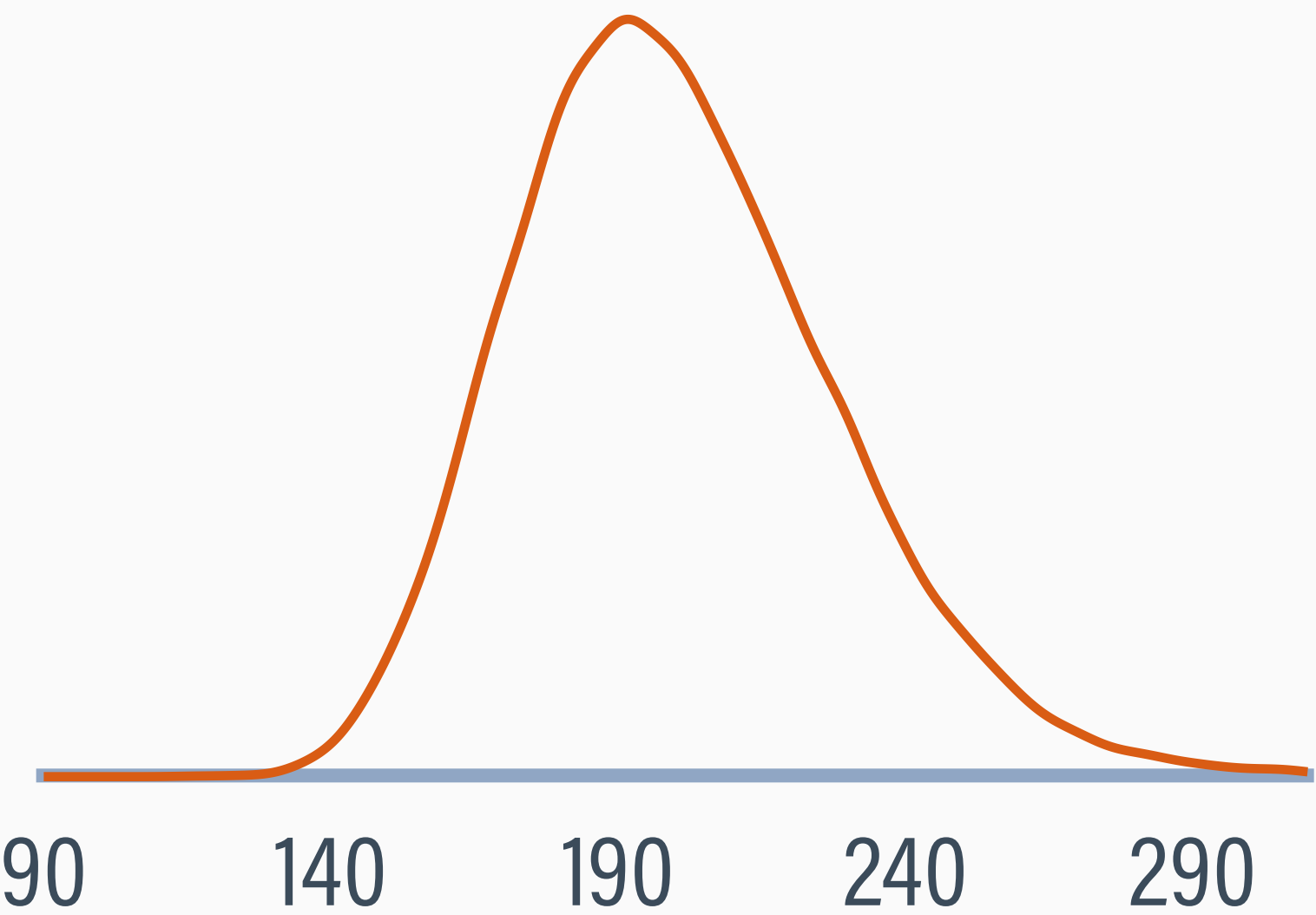


Grade 1 Learning Problems Slope

Median = 197.63

Std. Dev. = 27.62

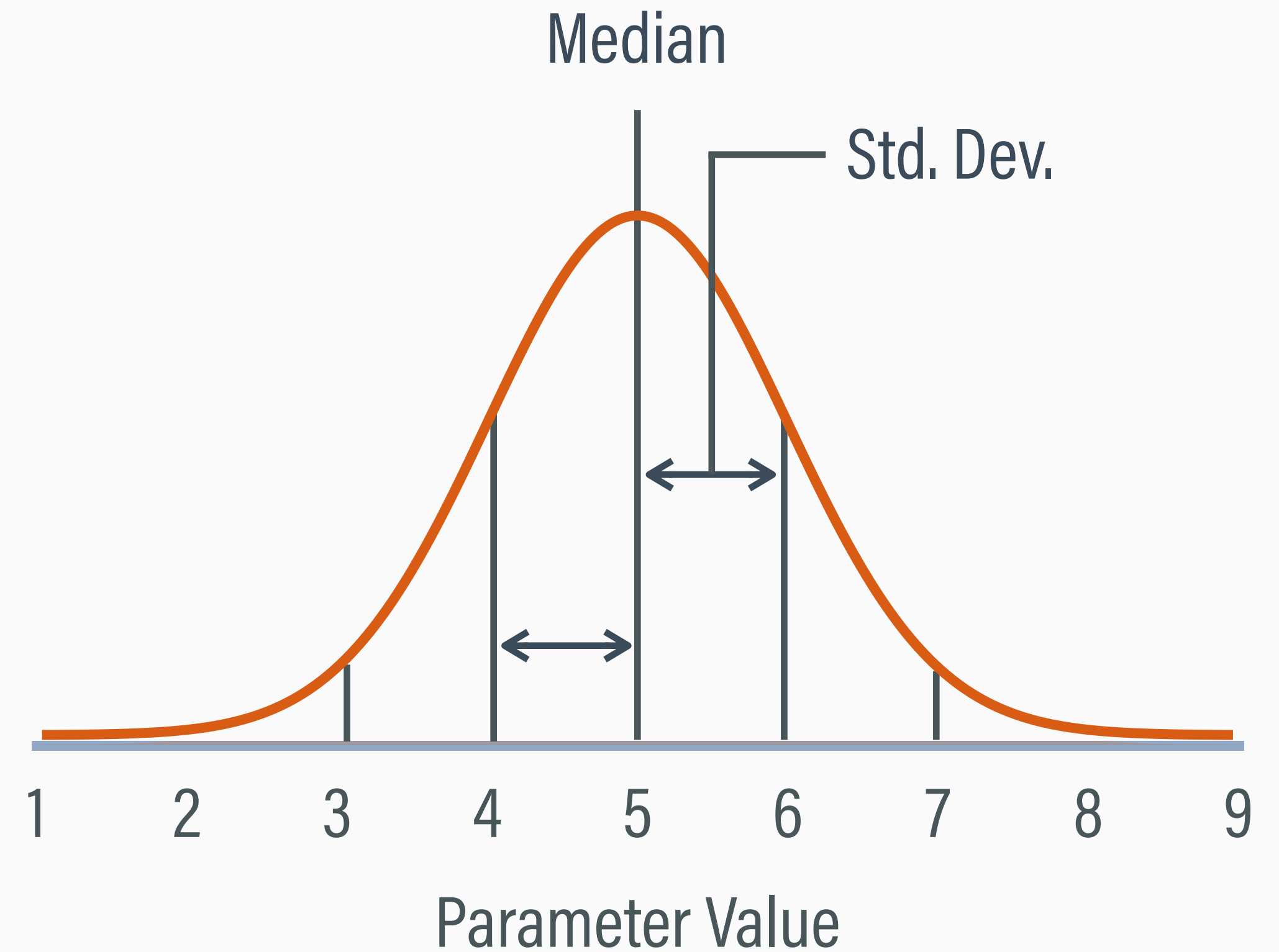
95% CI = (153.26, 260.29)



Residual Var.

ESTIMATES AND “BAYESIAN STANDARD ERRORS”

- The posterior median and standard deviation quantify the most likely parameter value and uncertainty
- Analogous to a point estimate and standard error but no repeated sampling



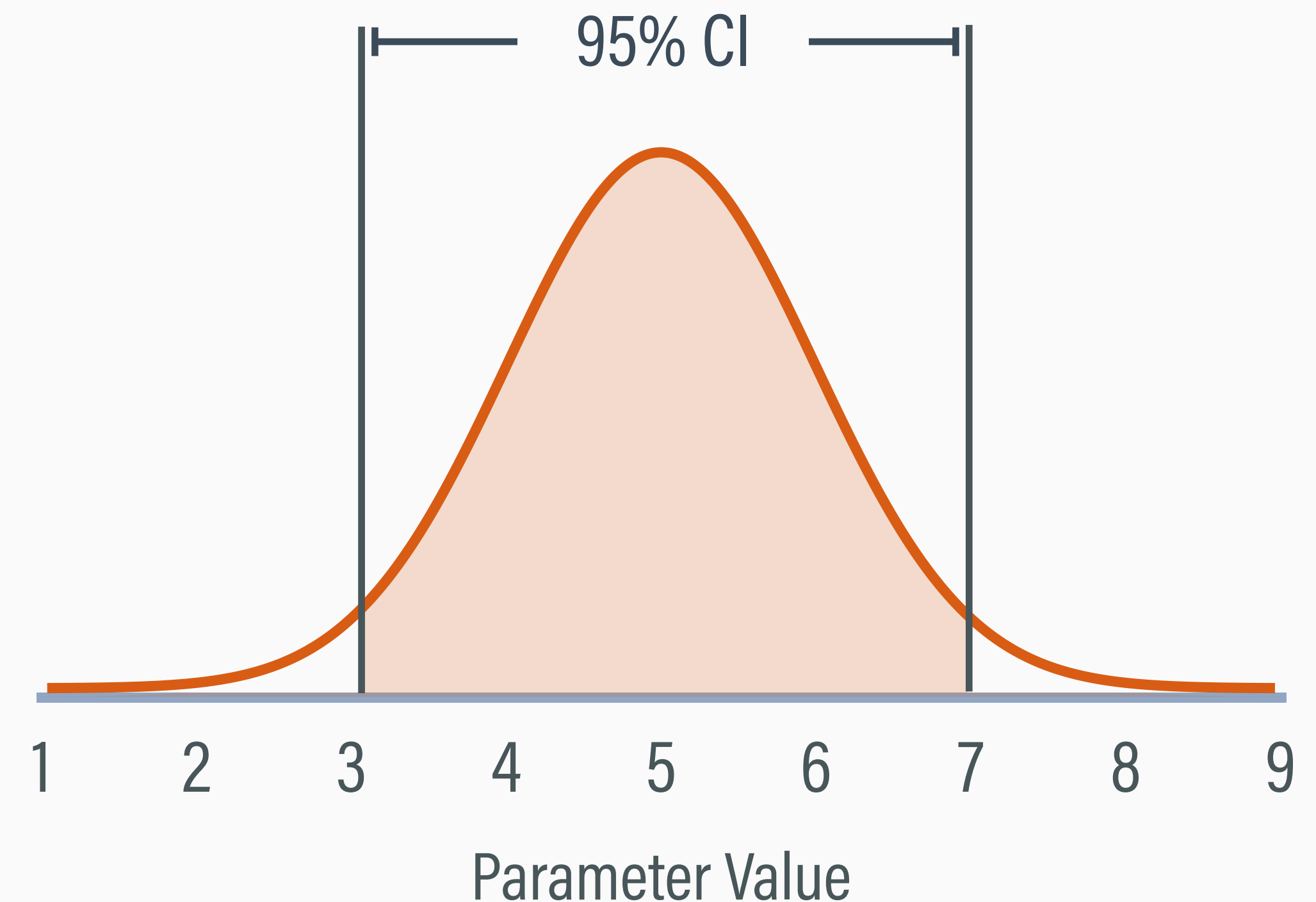
BLIMP OUTPUT

Outcome Variable: read9

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	197.634	27.615	153.259	260.292	---	---	6683.964
Coefficients:							
Intercept	112.482	6.812	99.090	125.866	272.342	0.000	6015.963
lnnprob1	-0.470	0.131	-0.728	-0.212	12.770	0.000	5573.516
Standard Deviations:							
Residual SD	14.058	0.966	12.380	16.134	---	---	6658.209
Standardized Coefficients:							
lnnprob1	-0.338	0.085	-0.490	-0.159	15.358	0.000	5516.808
Proportion Variance Explained							
by Coefficients	0.114	0.056	0.025	0.240	---	---	5579.911
by Residual Variation	0.886	0.056	0.760	0.975	---	---	5579.911

95% CREDIBLE INTERVALS

- The 95% credible interval gives limits spanning 95% of the parameter's range
- Akin to a confidence interval, but references a range of highly plausible parameter values



Outcome Variable: read9

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	197.634	27.615	153.259	260.292	---	---	6683.964
Coefficients:							
Intercept	112.482	6.812	99.090	125.866	272.342	0.000	6015.963
lnnprob1	-0.470	0.131	-0.728	-0.212	12.770	0.000	5573.516
Standard Deviations:							
Residual SD	14.058	0.966	12.380	16.134	---	---	6658.209
Standardized Coefficients:							
lnnprob1	-0.338	0.085	-0.490	-0.159	15.358	0.000	5516.808
Proportion Variance Explained							
by Coefficients	0.114	0.056	0.025	0.240	---	---	5579.911
by Residual Variation	0.886	0.056	0.760	0.975	---	---	5579.911

ESTIMATOR COMPARISON

The two estimators are effectively numerically equivalent even with a small N!!!

Parameter	MCMC			FIML		
	Median	SD	95% CI	Est.	SE	95% CI
Intercept	112.48	6.81	(99.09, 125.87)	112.43	6.58	(99.55, 125.32)
Learning Problems	−0.47	0.13	(−0.73, −0.21)	−0.47	0.13	(−0.72, −0.22)
Residual variance	197.63	27.62	(153.26, 260.29)	189.94	25.24	(140.46, 239.41)
R ²	.11	.06	(.03, .24)	.12	.06	—

MCMC AS COMPUTATIONAL FREQUENTISM

- Researchers adopting a computational frequentism view can use MCMC results as surrogates for frequentist inference (Levy & McNeish, 2021)
- In this scenario, MCMC is a flexible way to estimate frequentist quantities when FIML solutions are unavailable (e.g., missing data)



BLIMP OUTPUT

Outcome Variable: read9

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	197.634	27.615	153.259	260.292	---	---	6683.964
Coefficients:							
Intercept	112.482	6.812	99.090	125.866	272.342	0.000	6015.963
lnnprob1	-0.470	0.131	-0.728	-0.212	12.770	0.000	5573.516
Standard Deviations:							
Residual SD	14.058	0.966	12.380	16.134	---	---	6658.209
Standardized Coefficients:							
lnnprob1	-0.338	0.085	-0.490	-0.159	15.358	0.000	5516.808
Proportion Variance Explained							
by Coefficients	0.114	0.056	0.025	0.240	---	---	5579.911
by Residual Variation	0.886	0.056	0.760	0.975	---	---	5579.911

PROBLEMS MCMC ADDRESSES, REVISITED

- Mixed variable types (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

MODERATED REGRESSION MODEL

- Does the diagnostic utility of 1st-grade reading on 9th-grade reading depend on whether a student experiences learning problems in 1st grade?

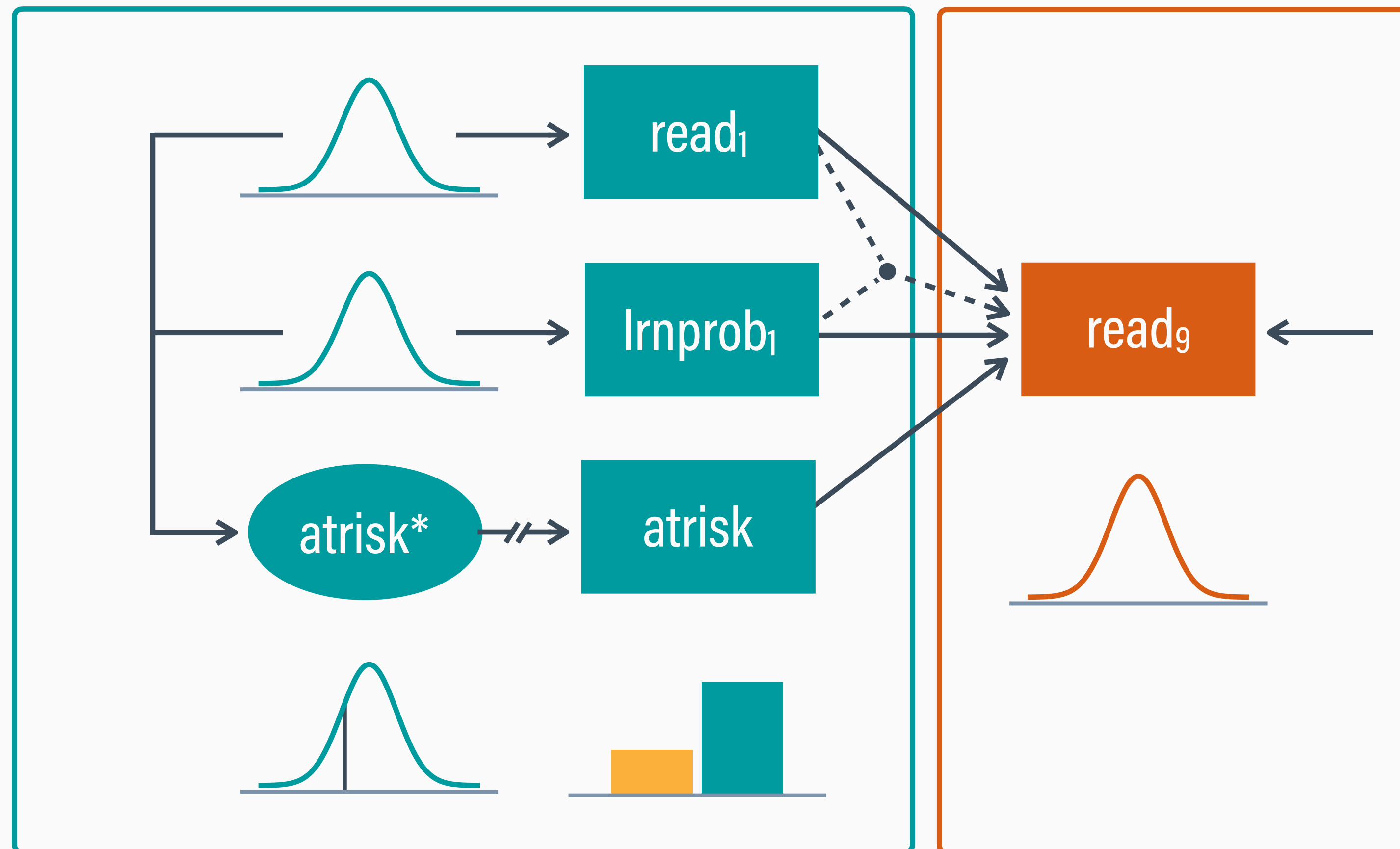
$$\text{read}_9 = \beta_0 + \beta_1(\text{read}_1) + \beta_2(\text{lrnprob}_1) + \beta_3(\text{read}_1)(\text{lrnprob}_1) + \beta_4(\text{atrisk}) + \varepsilon$$

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

FACTORED SPECIFICATION

Incomplete Predictor Model

Outcome Model

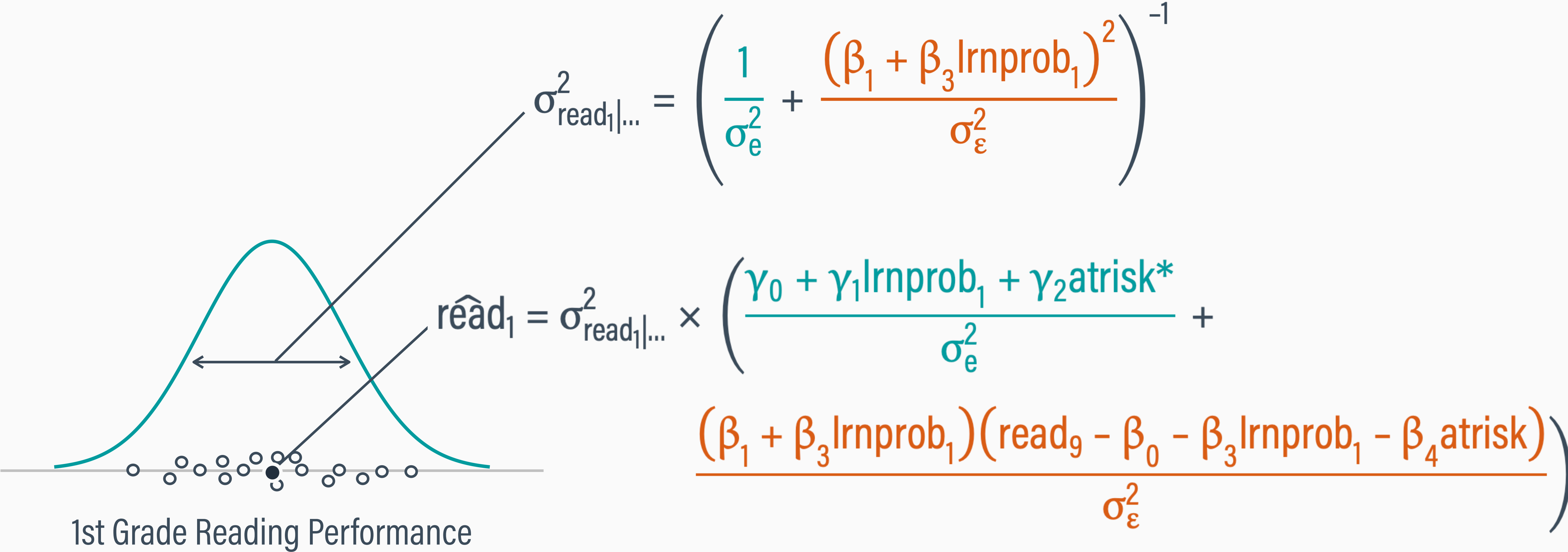


INCOMPLETE PREDICTORS

- Products are deterministic functions of lower-order predictors in the focal model rather than unique variables themselves
- With an interaction, a factored specification produces non-normal (heteroscedastic) predictor imputations that accommodate the nonlinear term in the focal model
- Multivariate normal methods introduce bias in this application

PREDICTED VALUES AND VARIATION

Imputations for interacting predictors are heteroscedastic (depends on one's simple slope)



BLIMP STUDIO SCRIPT

DATA: reading.dat;

VARIABLES: id atrisk lnrprob1 read1 read9;

MISSING: 999;

NOMINAL: atrisk;

CENTER: read1 lnrprob1; # center continuous predictors

MODEL: read9 ~ read1 lnrprob1 read1*lnrprob1 atrisk; # product in focal model

SIMPLE: read1 | lnrprob1; # conditional effects of read1 at different levels of lnrprob1

BURN: 10000;

ITER: 10000;

SEED: 90291;

RBLIMP SCRIPT

```
model2 <- rblimp(  
  data = reading,  
  nominal = 'atrisk',           # define nominal variables + automatic dummy coding  
  center = 'read1 lnrprob1',   # center continuous predictors  
  model = 'read9 ~ read1 lnrprob1 read1*lnrprob1 atrisk', # product in focal model  
  simple = 'read1 | lnrprob1', # conditional effects (simple slopes)  
  seed = 90291,  
  burn = 10000,  
  iter = 10000)  
output(model2)                 # print output  
simple_plot(read9 ~ read1 | lnrprob1, model2) # plot simple slopes  
jn_plot(read9 ~ read1 | lnrprob1, model2)   # Johnson-Neyman regions of significance
```

BLIMP OUTPUT

Outcome Variable: read9

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	83.095	12.307	63.746	111.827	---	---	4424.378
Coefficients:							
Intercept	89.316	1.752	85.821	92.718	2599.229	0.000	1732.732
read1	0.523	0.050	0.424	0.624	108.997	0.000	3908.295
lrnprob1	-0.424	0.097	-0.615	-0.233	19.142	0.000	2155.092
atrisk.1	-2.124	1.886	-5.835	1.520	1.284	0.257	5651.625
read1*lrnprob1	0.017	0.005	0.007	0.027	11.192	0.001	3025.910
Standard Deviations:							
Residual SD	9.116	0.662	7.984	10.575	---	---	4411.531
Standardized Coefficients:							
...							
Proportion Variance Explained							
by Coefficients	0.628	0.050	0.518	0.712	---	---	4420.718
by Residual Variation	0.372	0.050	0.288	0.482	---	---	4420.718

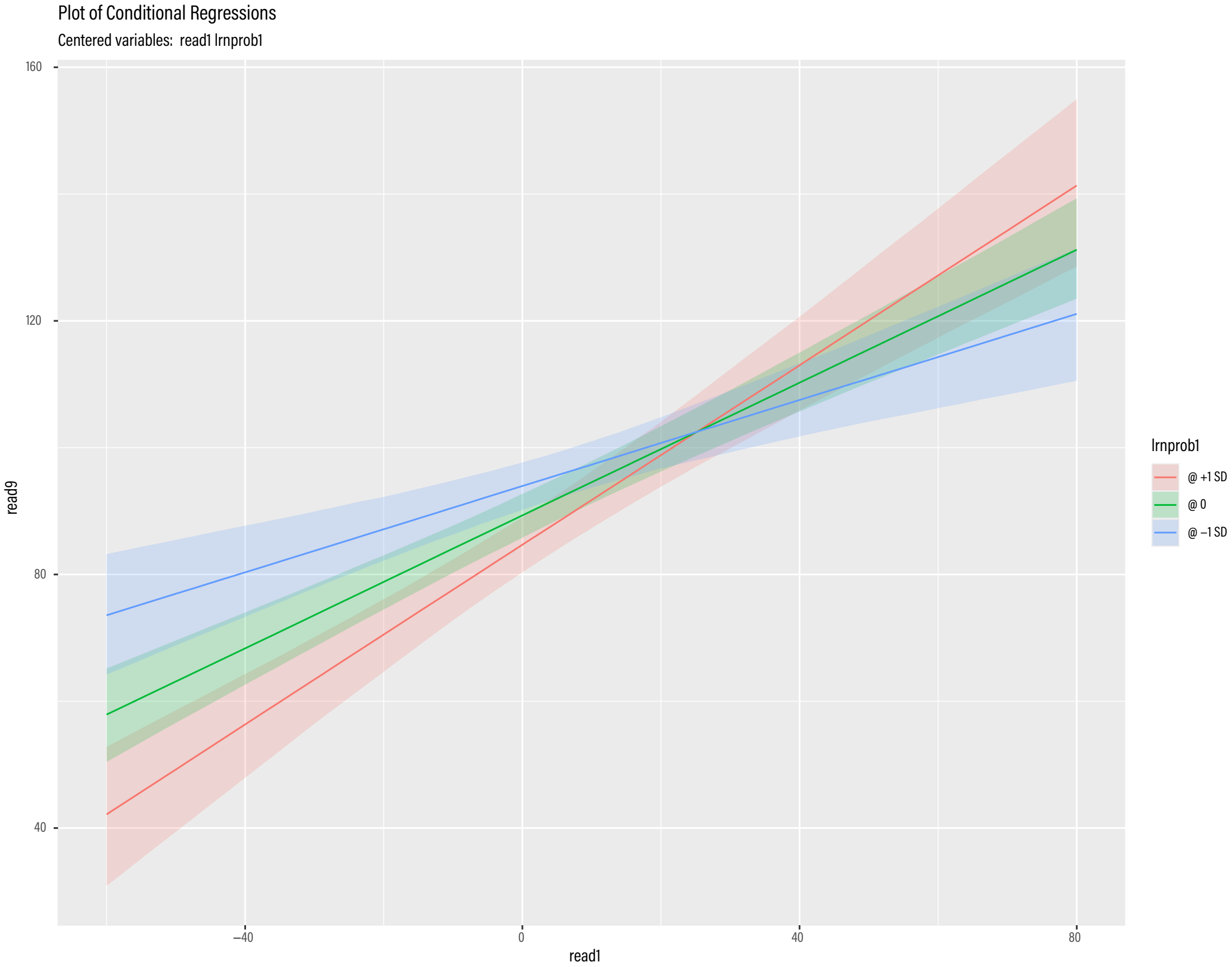
CONDITIONAL EFFECTS OUTPUT

Conditional Effects	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff

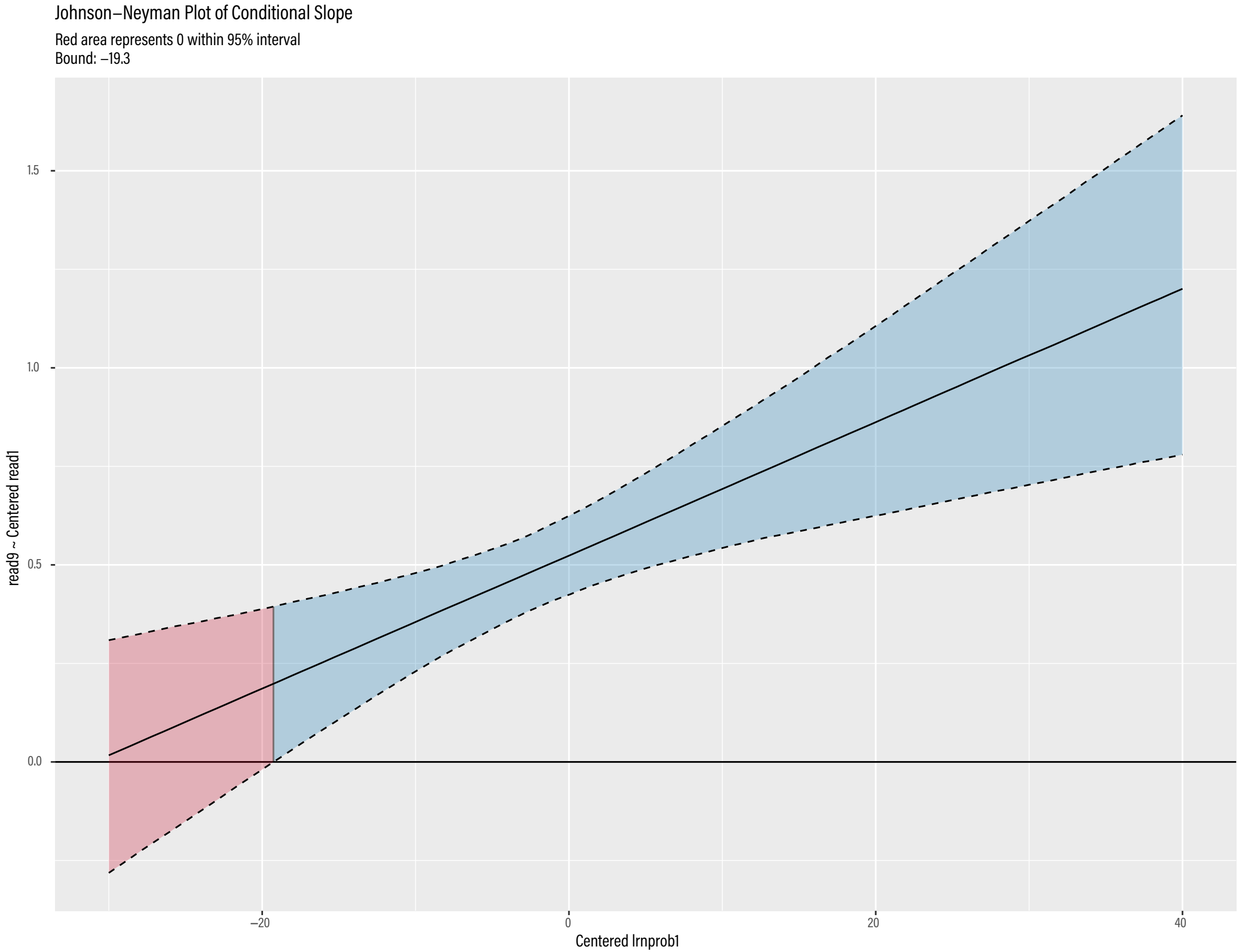
read1 lnrnprob1 @ +1 SD							
Intercept	84.692	2.211	80.319	89.026	1467.138	0.000	1532.751
Slope	0.709	0.083	0.550	0.876	73.692	0.000	3170.278
read1 lnrnprob1 @ 0							
Intercept	89.316	1.752	85.821	92.718	2599.229	0.000	1732.732
Slope	0.523	0.050	0.424	0.624	108.997	0.000	3908.295
read1 lnrnprob1 @ -1 SD							
Intercept	93.951	1.913	90.167	97.671	2410.733	0.000	2582.344
Slope	0.340	0.067	0.205	0.470	25.269	0.000	4334.720

RBLIMP GRAPHICAL OUTPUT

Conditional Effects (Simple Slopes)



Johnson–Neyman Regions of Significance



OUTLINE

1

Big Three Missing Data Methods

2

Introduction to MCMC Estimation

3

Application: Moderated Regression

4

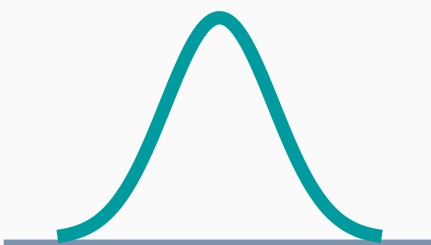
Application: Structural Equation Model

FACTORED SEM

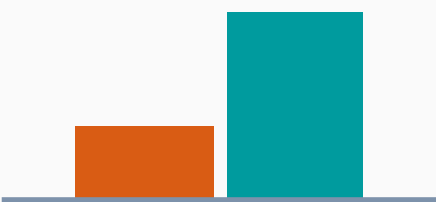
- Blimp's SEM functionality allows combinations of linear and generalized linear models with up to three levels
- Accommodates diverse variable types: normal or skewed continuous, binary, ordinal, nominal, count, and two-part
- Supports complex data features, including interactions, nonlinear effects, heteroscedasticity, dynamic effects, and functions of predictors (e.g., sum scores, residualized predictors)

MODEL AND VARIABLE TYPES

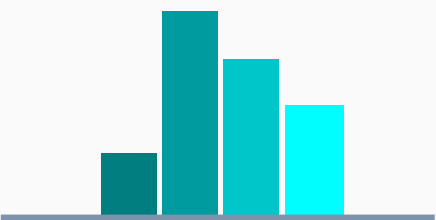
Exogenous Predictors
(Multivariate Submodel)



Normal
(Manifest)



Binary

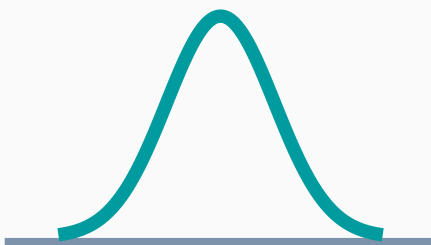


Ordinal

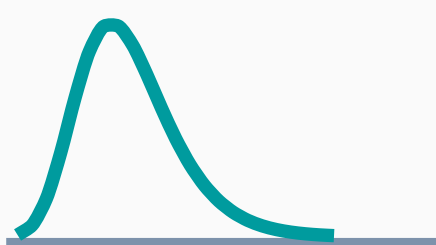


Nominal

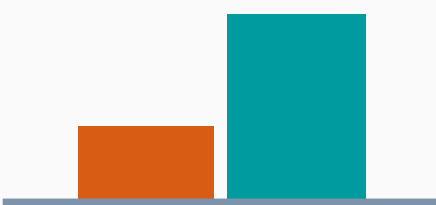
Dependent Variables
(Univariate Submodels)



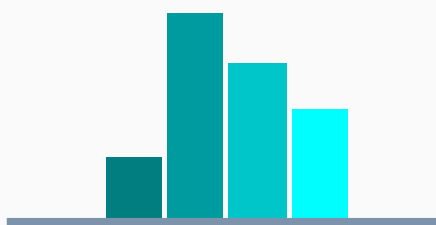
Normal
(Manifest or Latent)



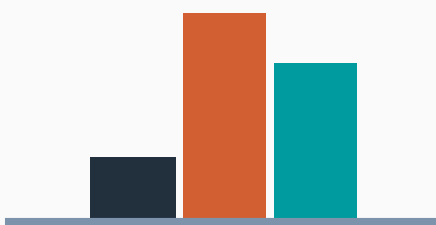
Skewed



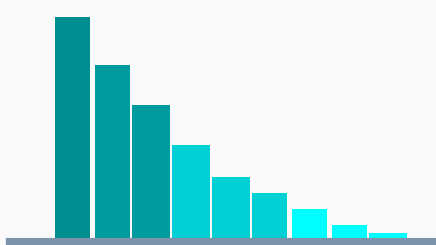
Binary



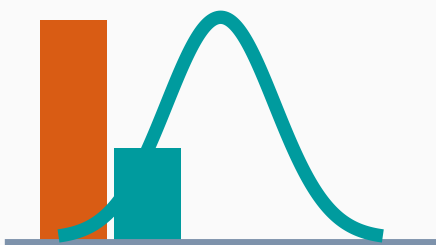
Ordinal



Nominal

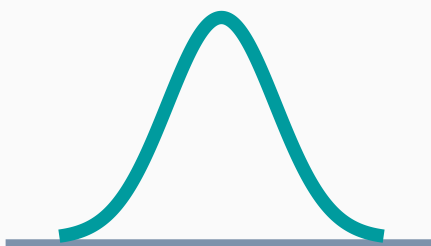


Count

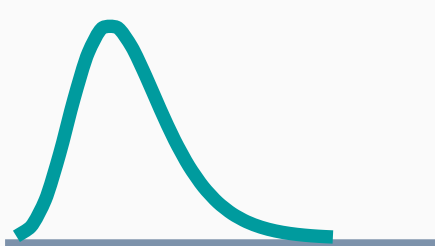


Two-Part

Dependent Variables
(Multivariate Submodels)



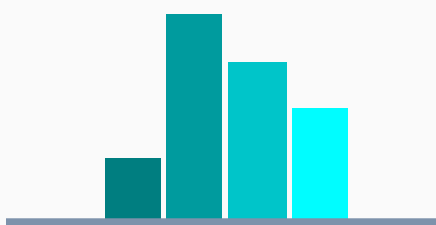
Normal



Skewed



Binary



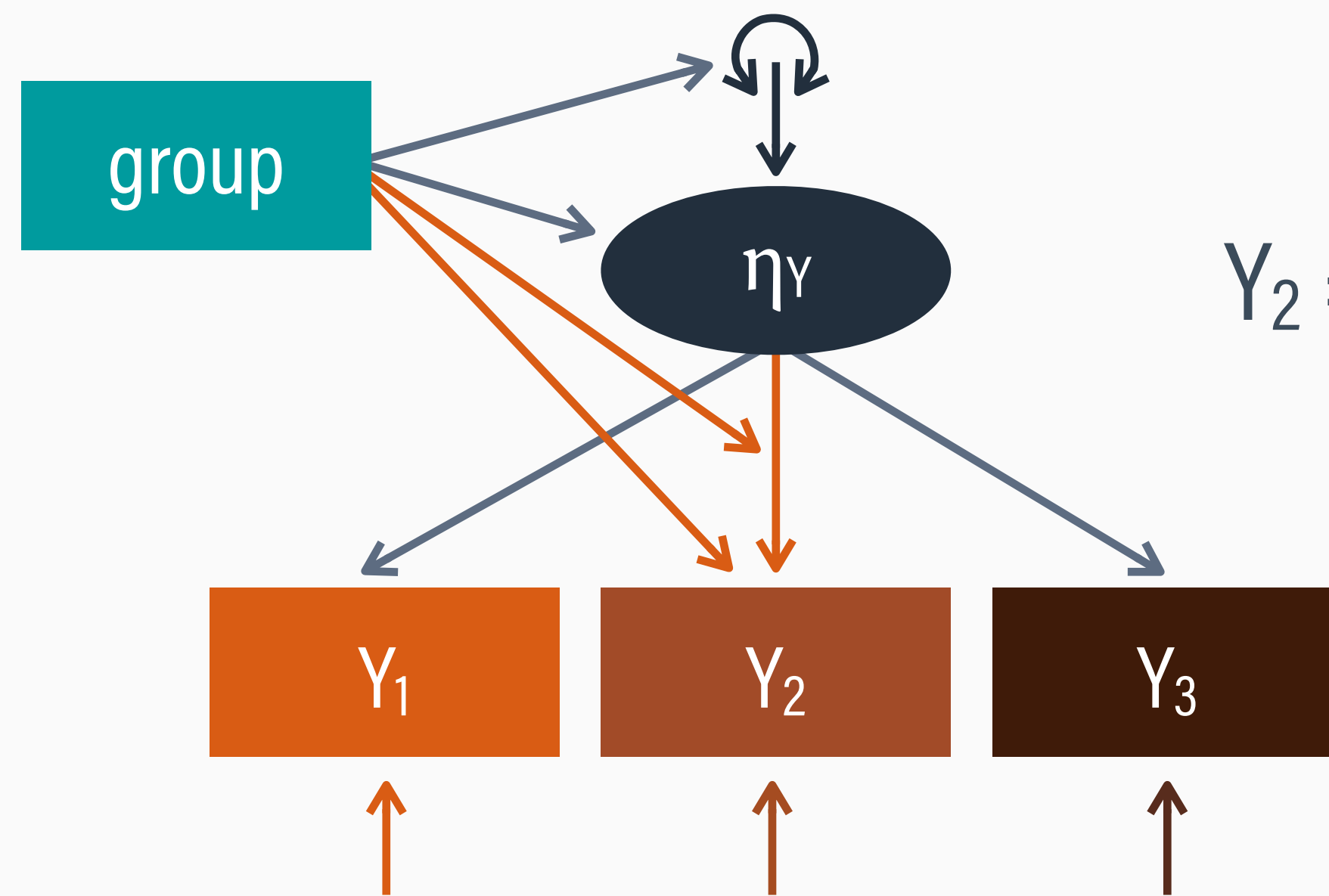
Ordinal

LATENT VARIABLES AS MISSING DATA

- Blimp treats latent variables as missing data to be imputed; factors are just continuous variables with 100% missing data
- Imputing latent variables supports two-way and higher-order interactions, including various latent-by-manifest interactions, and other types of curvilinear and nonlinear effects
- Like other continuous variables, latent variables can be normal, skewed, and heteroscedastic

MODERATED NONLINEAR FACTOR MODEL

- Treating latent variables as missing data accommodates a variety of latent-by-manifest variable interactions



$$Y_2 = \beta_0 + \beta_1(\text{group}) + \beta_2(\eta_Y) + \beta_3(\eta_Y)(\text{group}) + \varepsilon$$

BLIMP SCRIPT

DATA: mnlf.dat;

VARIABLES: group y1 y2 y3 y4;

MISSING: 999;

ORDINAL: y1 y2 y3 y4;

invokes probit link for ordinal indicators

NOMINAL: group;

multicategorical variable with automatic dummy coding

LATENT: eta;

MODEL:

eta -> y1 y3 y4;

measurement model

y2 ~ eta group group*eta;

measurement model with latent-by-nominal interaction (Dif)

eta ~ group;

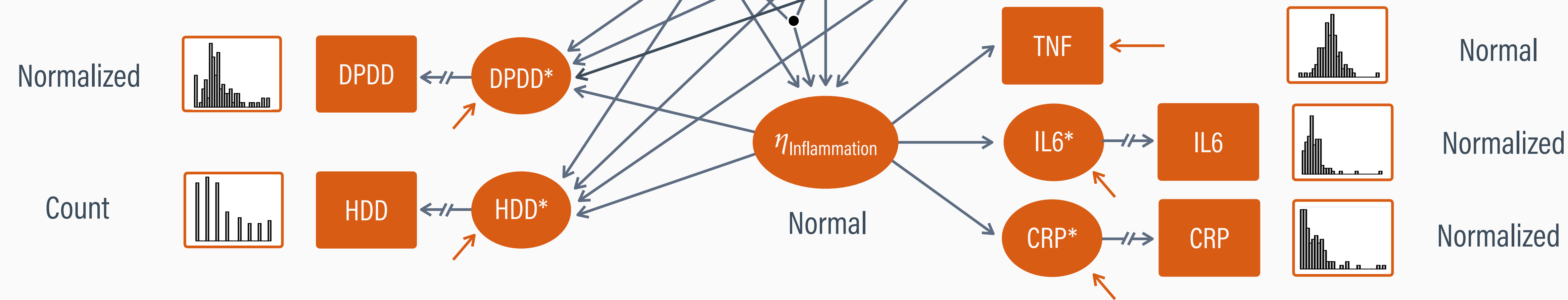
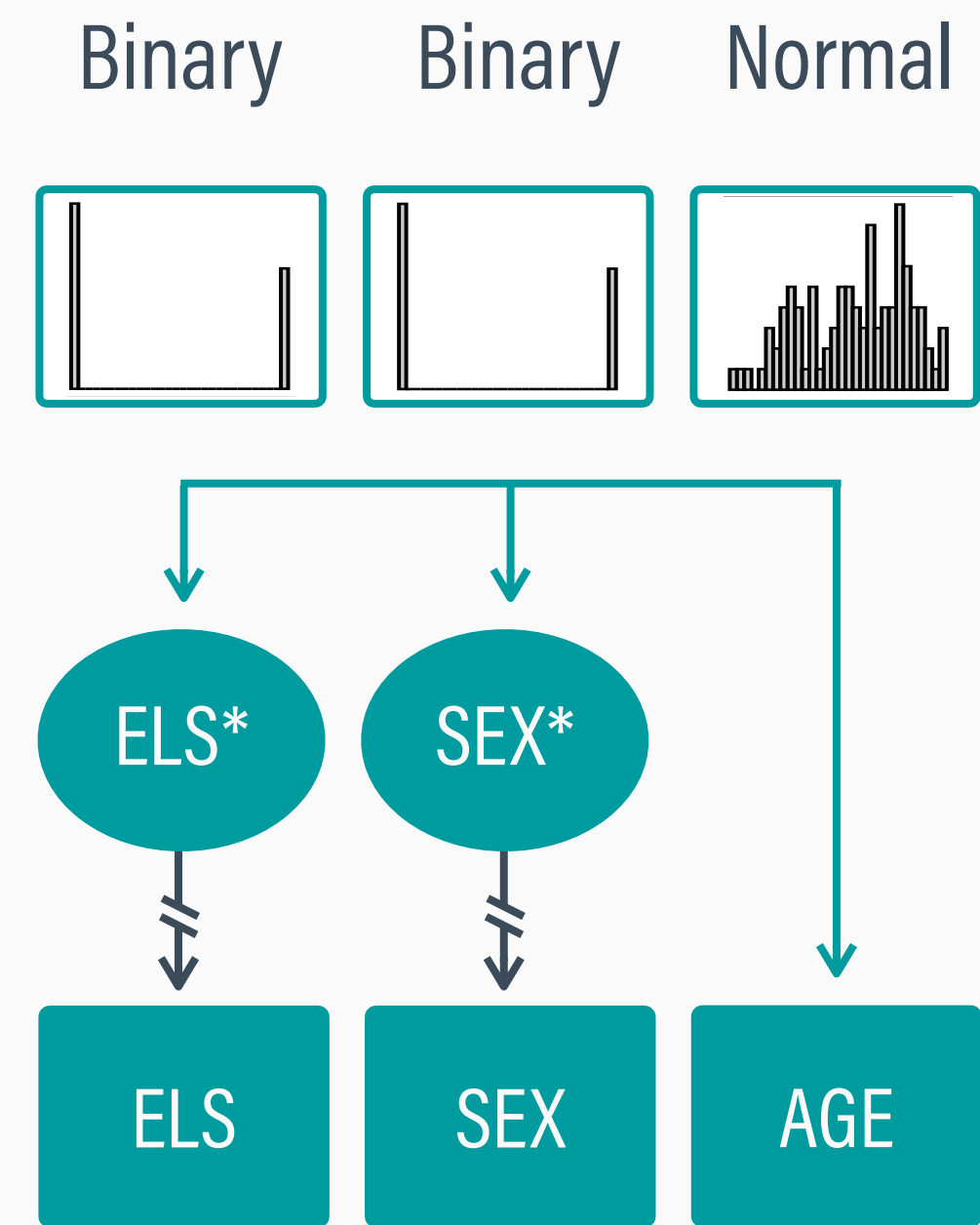
structural model

BURN: 20000;

ITERATIONS: 20000;

SEED: 90291;

LATENT VARIABLE MEDIATION MODEL



BLIMP SCRIPT

DATA: fsem.dat;

VARIABLES: els sex age tnf il6 crp dpdd hdd;

MISSING: 999;

NOMINAL: els sex;

COUNT: hdd;

LATENT: inflam;

MODEL:

inflam -> yjt(il6) yjt(crp) tnf;

measurement model with skewed indicators

inflam ~ els sex els*sex;

latent mediator model with interaction

yjt(dpdd) ~ nflam els sex;

skewed outcome model

hdd ~ nflam els sex;

count outcome model

BURN: 20000;

ITERATIONS: 20000;

SEED: 90291;



For more information go to

WWW.APPLIEDMISSINGDATA.COM