AN INTRODUCTION TO MISSING DATA ANALYSES FOR EDUCATION RESEARCH

Craig Enders, UCLA

Brian Keller, University of Missouri

Remus Mitchell, UCLA

COURSE MATERIAL DOWNLOAD

WWW.APPLIEDMISSINGDATA.COM/BLIMP-PAPERS

APPLIED MISSING DATA home analysis examples blimp **blimp papers** videos centerstat workshop quantitude podcast

Workshops and Training

Enders, C., Keller, B., & Mitchell, R. (2025, April). *An introduction to missing data analyses for educational research*. Professional development workshop presented at the annual meeting of the American Educational Research Association. Denver, CO.

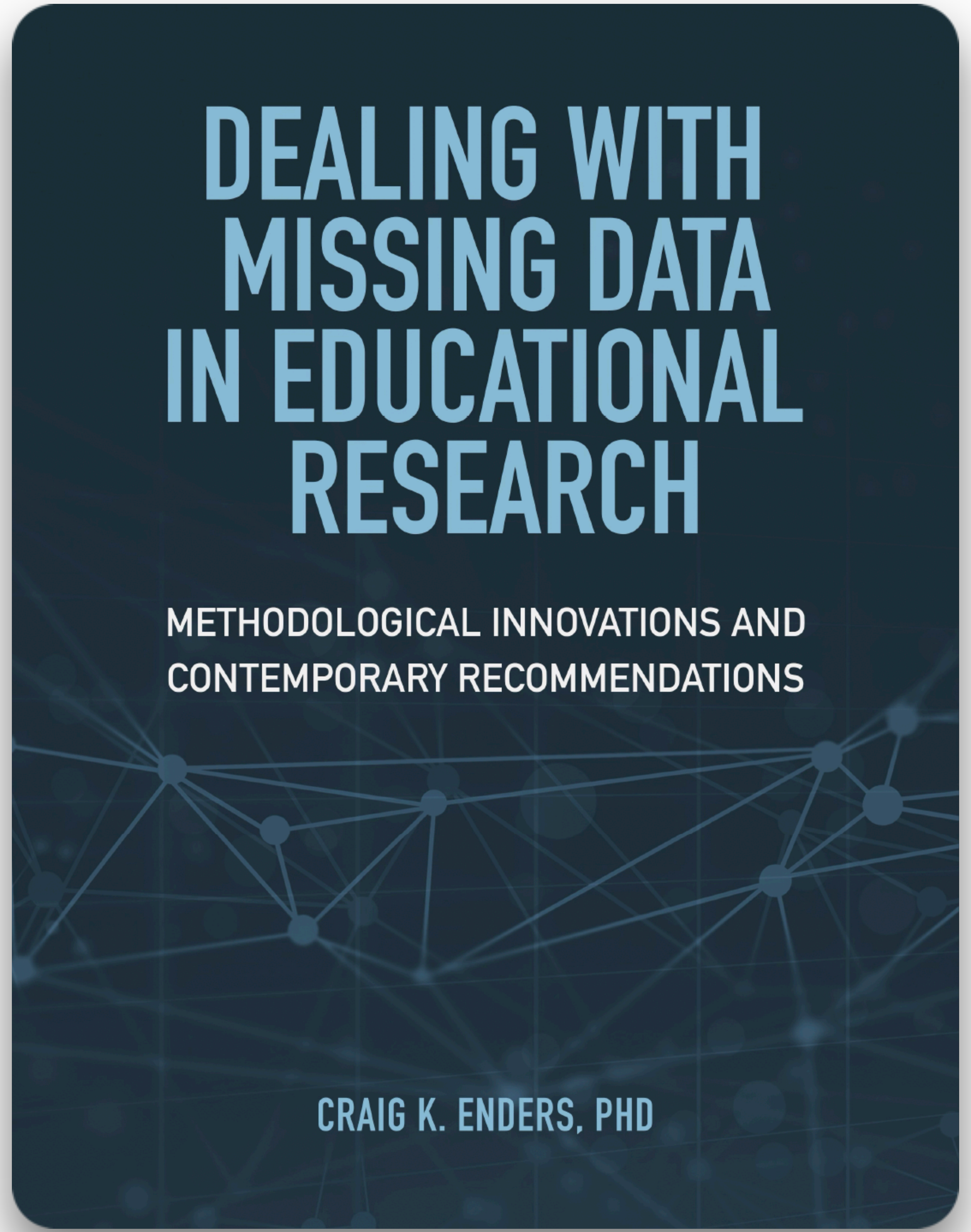
DOWNLOAD WEBINAR MATERIALS

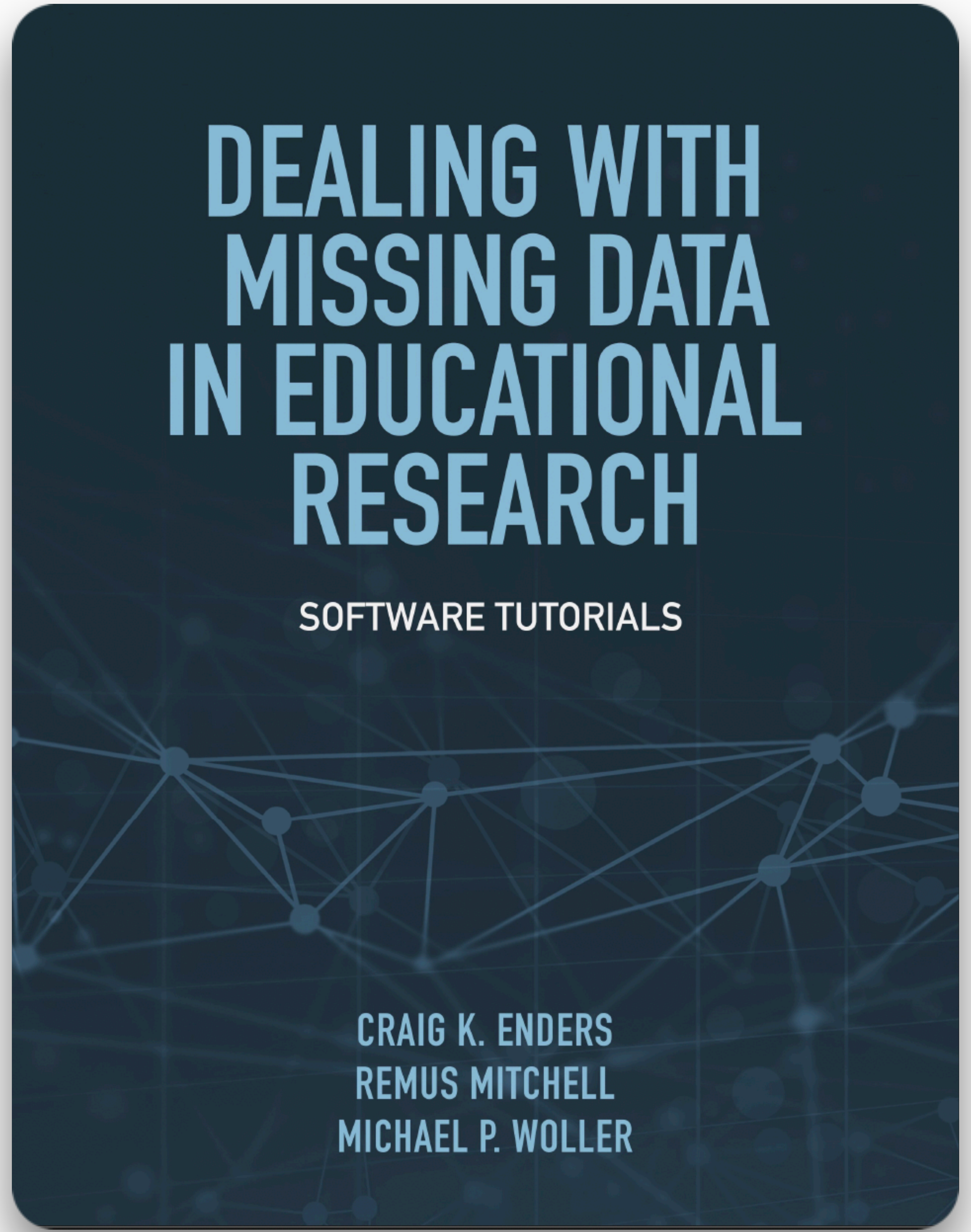
COURSE MATERIALS

- AERA Analysis Examples
  - Blimp Studio Analysis 1 Script.imp
  - Blimp Studio Analysis 2 Script.imp
  - Blimp Studio Analysis 3 Script.imp
  - Blimp Studio Analysis 4 Script.imp
  - rBlimp Analysis Scripts.R
  - reading.dat
- AERA Missing Data Workshop.pdf
- IES Missing Data Toolkit
  - Dealing With Missing Data Analysis Scripts
  - Dealing With Missing Data in Educational Research - Software Tutorials.pdf
  - Dealing With Missing Data in Educational Research.pdf

IES MISSING DATA TOOLKIT

IES Institute of Education Sciences





IES Institute of Education Sciences

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 for updates, as new videos are continually added.

INSTALLING BLIMP

IES Institute of Education Sciences

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.

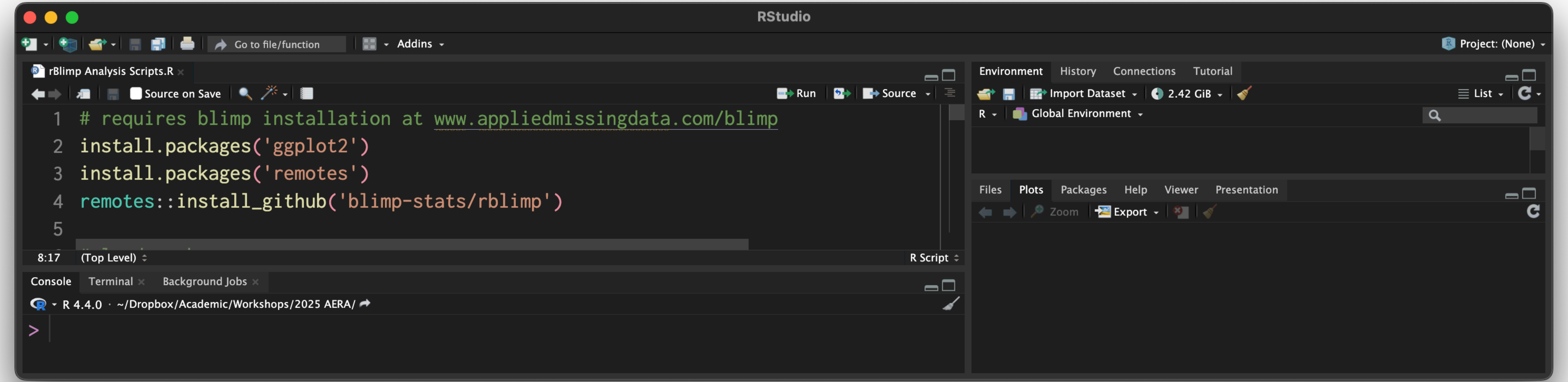
Download Now

User's Guide

BLIMP

INSTALLING RBLIMP (OPTIONAL)

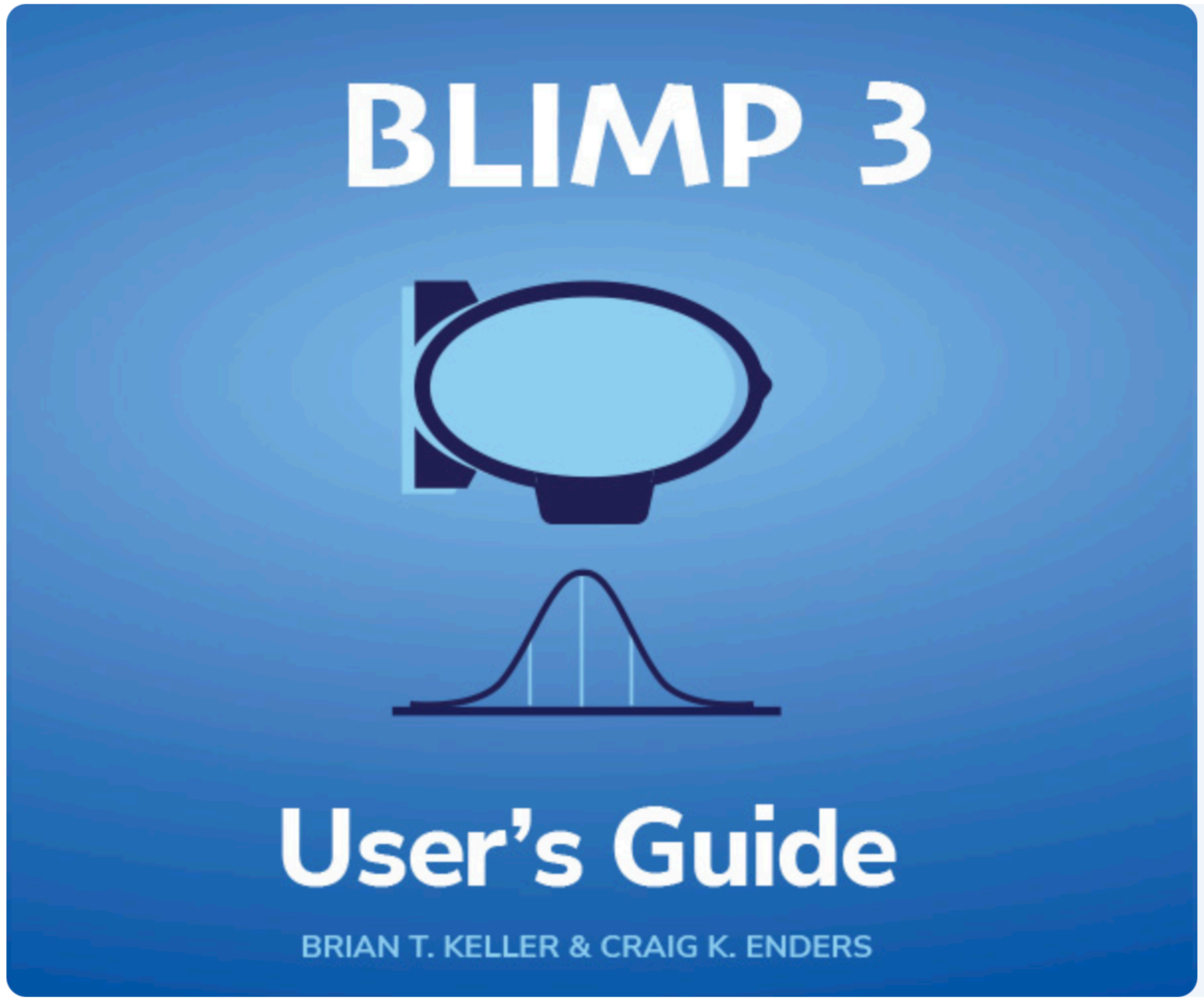
- Package installation lines are at the top of the rBlimp analysis scripts file in the course materials



BLIMP USER GUIDE

IES Institute of Education Sciences

WWW.APPLIEDMISSINGDATA.COM/BLIMP



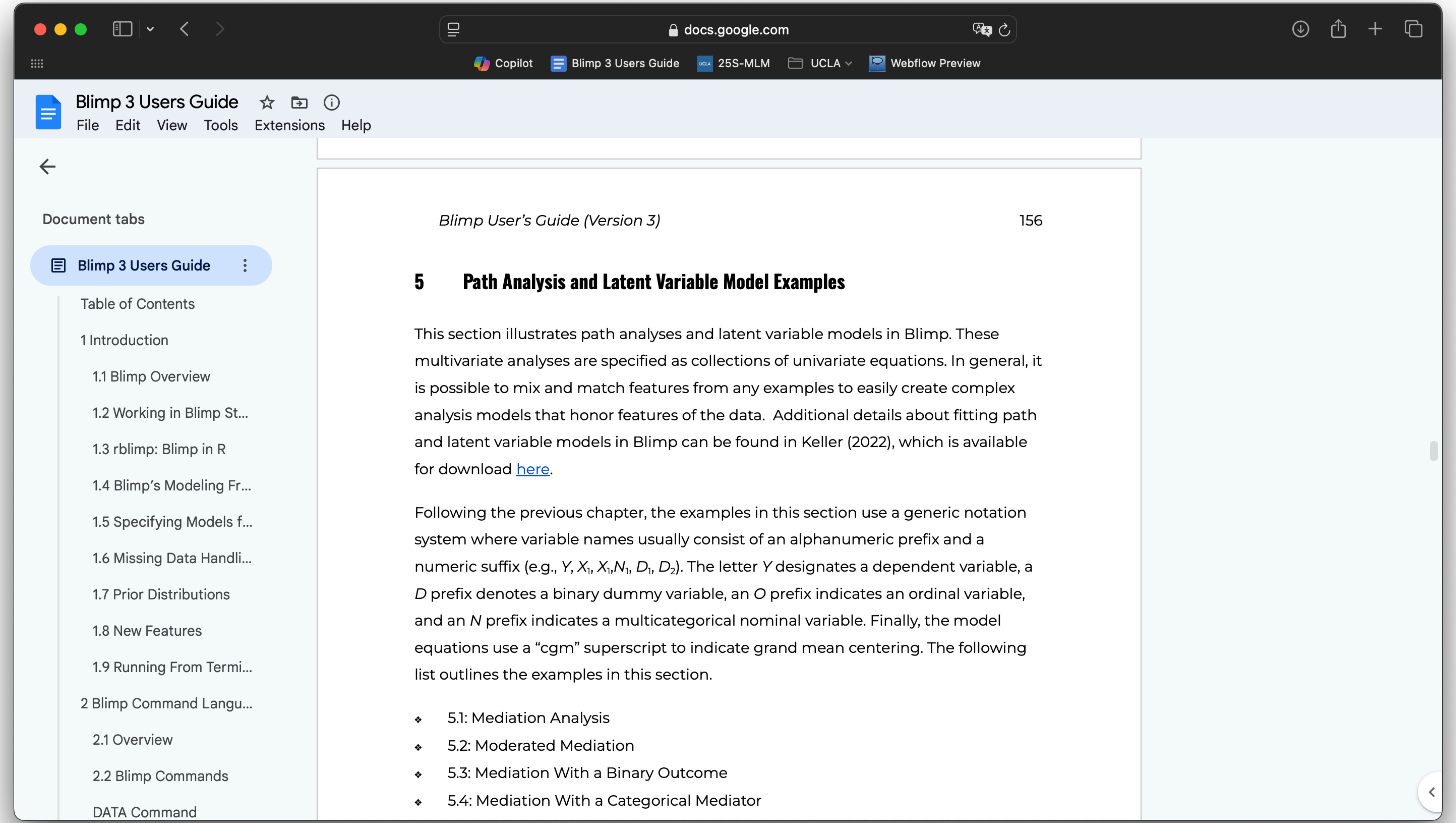
USER'S GUIDE

User Guide

Discussion Board (Coming Soon)

The User Guide provides an accessible overview of Blimp's simple scripting language, including dozens of new analysis examples covering regression models, path and latent variable models, psychometric models, multilevel models, and missing not at random models. All user examples are accessible and executable from Blimp Studio pull-down menus.

BLIMP USER GUIDE



Blimp User's Guide (Version 3) 156

5 Path Analysis and Latent Variable Model Examples

This section illustrates path analyses and latent variable models in Blimp. These multivariate analyses are specified as collections of univariate equations. In general, it is possible to mix and match features from any examples to easily create complex analysis models that honor features of the data. Additional details about fitting path and latent variable models in Blimp can be found in Keller (2022), which is available for download here.

Following the previous chapter, the examples in this section use a generic notation system where variable names usually consist of an alphanumeric prefix and a numeric suffix (e.g., Y, X_1, X_1N_1, D_1, D_2). The letter Y designates a dependent variable, a D prefix denotes a binary dummy variable, an O prefix indicates an ordinal variable, and an N prefix indicates a multicategorical nominal variable. Finally, the model equations use a "cgm" superscript to indicate grand mean centering. The following list outlines the examples in this section.

- 5.1: Mediation Analysis
- 5.2: Moderated Mediation
- 5.3: Mediation With a Binary Outcome
- 5.4: Mediation With a Categorical Mediator

OUTLINE

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. MCMC Estimation and Missing Data Imputation
4. Fitting Regression Models in Blimp
5. Incomplete Categorical Variables
6. Interaction Effects
7. Reporting Results From a Missing Data Analysis

OUTLINE

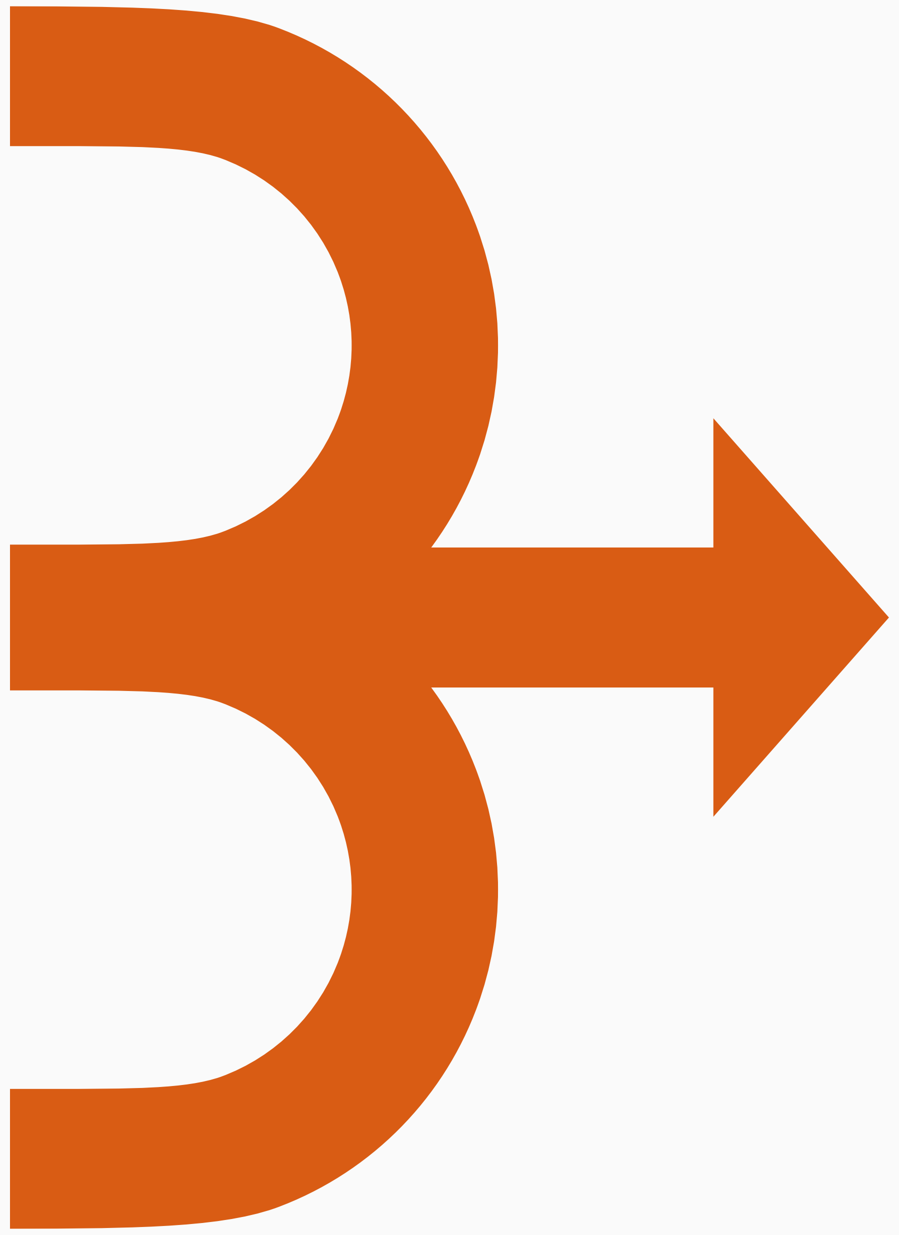
1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. MCMC Estimation and Missing Data Imputation
4. Fitting Regression Models in Blimp
5. Incomplete Categorical Variables
6. Interaction Effects
7. Reporting Results From a Missing Data Analysis

MODERN MISSING DATA METHODS

Maximum likelihood

Bayesian MCMC estimation

Multiple imputation



the
Big
Three

KEY ADVANTAGES OF BIG THREE

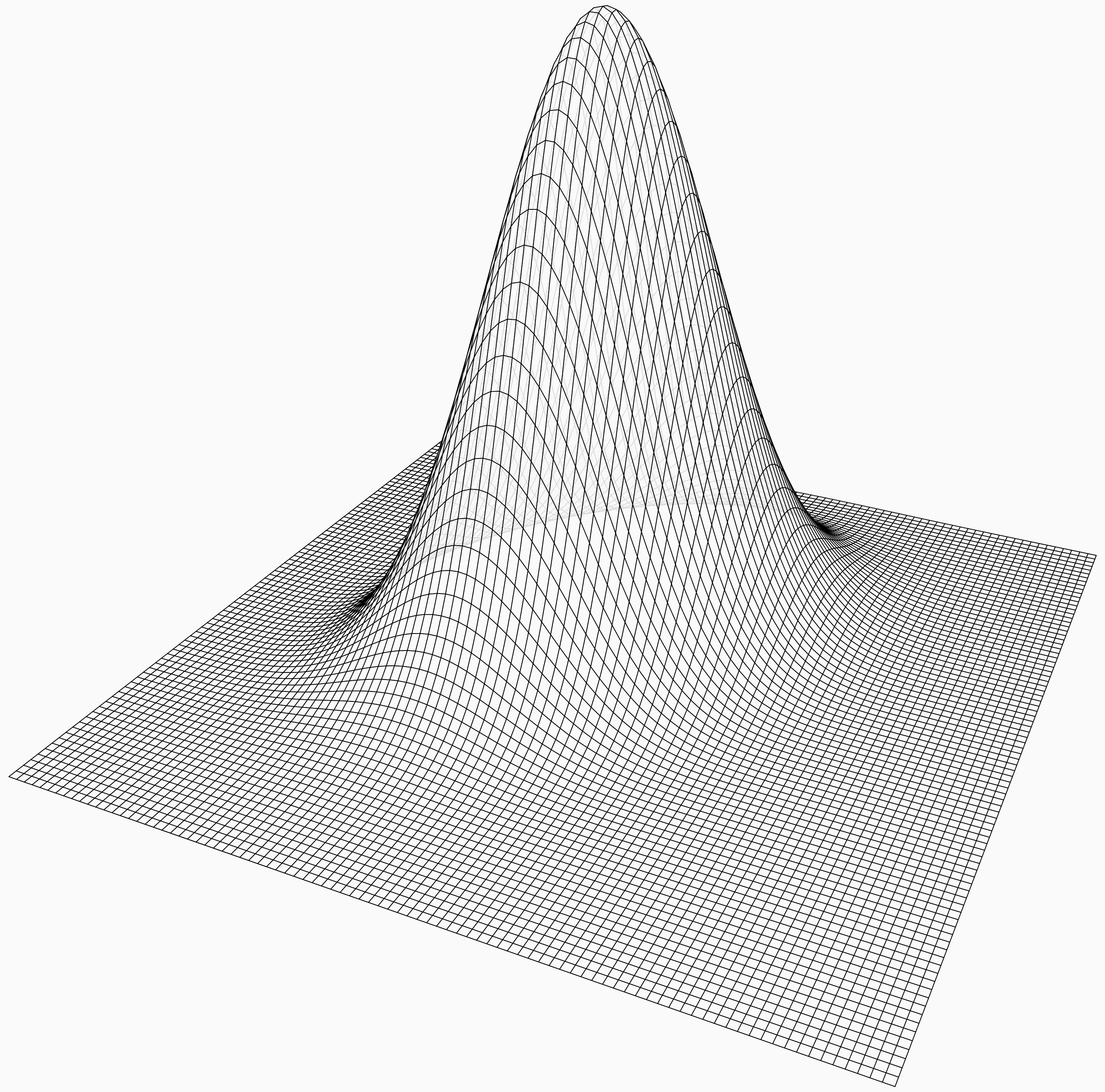
- Achieve unbiasedness with a more realistic assumption about the missing data process
- Allow for alternate assumptions about nonresponse process
- Maximize power
- Use all available data, no wasted resources

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
- How we represent those distributions—multivariate versus factored specifications—is what matters

MODELING FRAMEWORKS

Multivariate modeling

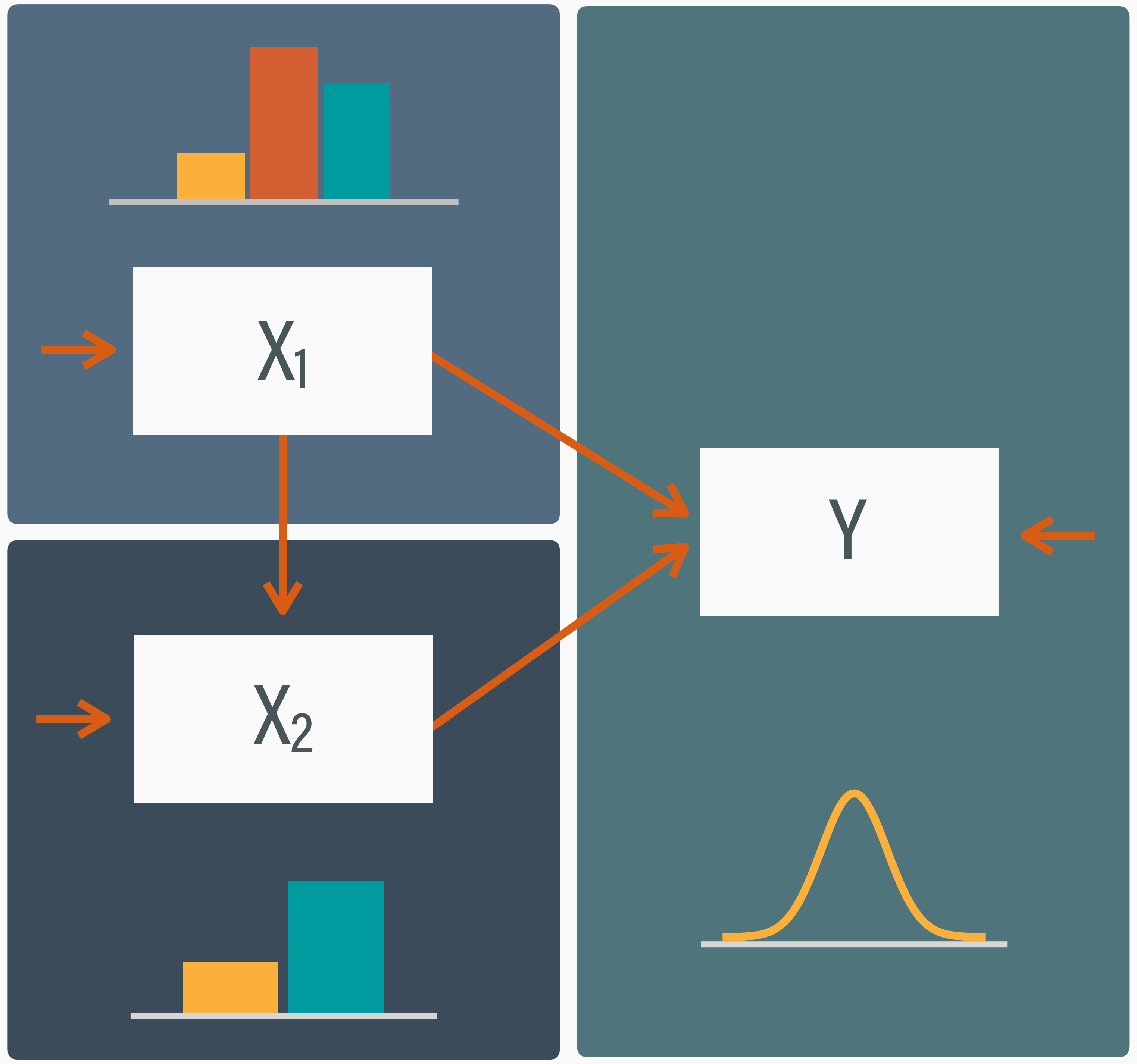


- Classic approaches often assume multivariate normality
- Most applications of maximum likelihood and multiple imputation

MODELING FRAMEWORKS

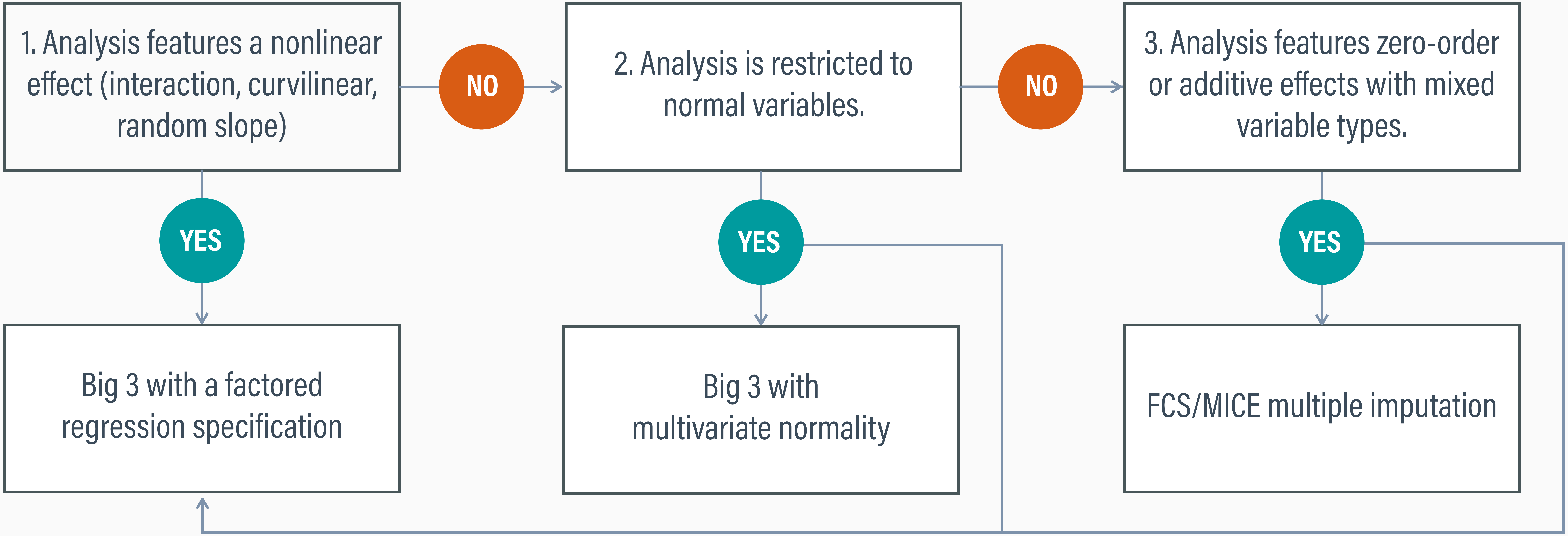
Multivariate modeling

Factored regression specification



- Factored regression invokes a unique model and distribution for each variable
- Each model can include terms that are at odds with multivariate normality (e.g., categorical variables, interactions, random slopes)

MISSING DATA DECISION TREE



WHY CHOOSE MCMC?

- MCMC readily handles complex missing data problems, including:
  - Mixed metrics (normal, ordinal, nominal, skewed, count, latent)
  - Nonlinear effects (interactions, curvilinear effects)
  - Multilevel data (random coefficients, interactions)
  - Latent variable modeling (interactions)
- FIML estimators with factored specifications are limited

OUTLINE

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. MCMC Estimation and Missing Data Imputation
4. Fitting Regression Models in Blimp
5. Incomplete Categorical Variables
6. Interaction Effects
7. Reporting Results From a Missing Data Analysis

HOW MUCH MISSING DATA IS TOO MUCH?

- The Big Three can tolerate substantial amounts of missing data
- The Big Three are increasingly better than ad hoc methods (e.g., deleting incomplete cases) as missingness increases
- The amount of missing data is less important than why the data are missing (the missingness process or mechanism)

RUBIN'S MISSING DATA MECHANISMS

- Missing data mechanisms (processes) describe different ways in which the data relate to nonresponse
- Missingness may be completely random or systematically related to different parts of the data
- Mechanisms function as statistical assumptions

PARTITIONING THE DATA

Complete = Observed + Missing; Indicators

Y_1	Y_2	Y_3		Y_1	Y_2	Y_3		Y_1	Y_2	Y_3	M_1	M_2	M_3
4	4	3		4	4	3					0	0	0
3	**3**	5		3	**NA**	5			**3**		0	**1**	0
7	1	6		7	1	6					0	0	0
2	1	6		**NA**	1	6		**2**			**1**	0	0
5	9	3	=	5	9	3	+				0	0	0
3	**2**	**2**		3	**NA**	**NA**			**2**	**2**	0	**1**	**1**
1	6	7		1	6	7					0	0	0
9	4	9		9	4	9					0	0	0
2	**5**	6		2	**NA**	6			**5**		0	**1**	0

MISSING COMPLETELY AT RANDOM

- The probability of missing values is completely unrelated to the data

$$f(M = 1 \mid data_{obs}, data_{mis}) = f(M = 1)$$

- MCAR is purely random missingness
- We don't care about this process or testing for it (e.g., Little's MCAR test)

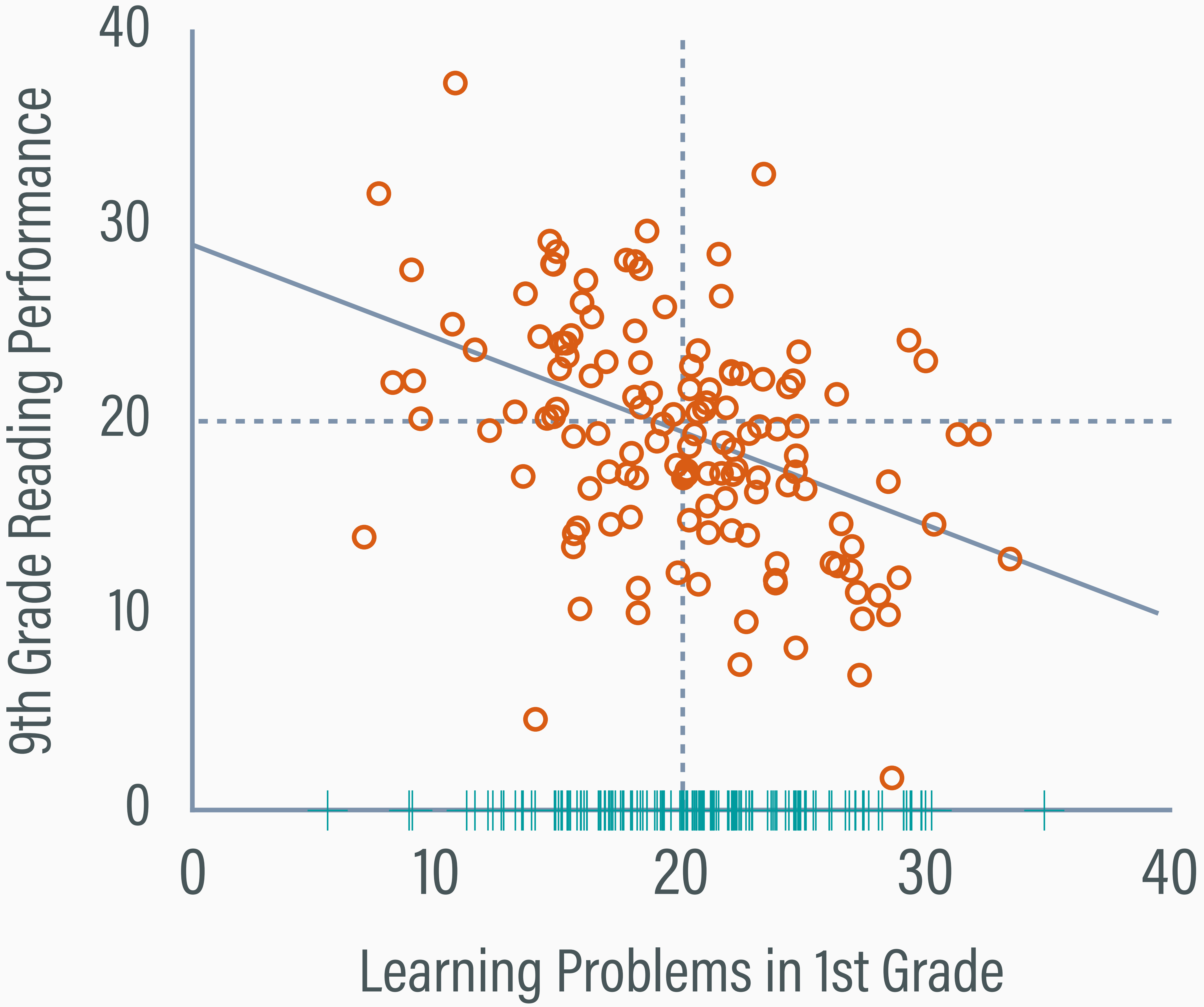
Missingness | Predictors of nonresponse

Indicators M_1	M_2	M_3	Observed Y_1	Y_2	Y_3	Missing Y_1	Y_2	Y_3
0	0	0	4	4	3			
0	1	0	3	NA	5		3	
0	0	0	7	1	6			
1	0	0	NA		6	2		
0	0	0	5	9	3			
0	1	1	3	NA	NA		2	2
0	0	0	1	6	7			
0	0	0	9	4	9			
0	1	0	2	NA	6		5	

RESEARCH SCENARIO

- Study investigating association between learning problems in 1st grade and reading performance in 9th grade

- Learning problems ratings are complete and reading scores are incomplete



MCAR EXAMPLE

- Missingness is unrelated to the observed learning problems measure and unrelated to the unseen reading scores
- Planned missing data design where 9th grade reading scores are collected from a random subset of the original sample in order to reduce data collection costs
- Unplanned missingness is unrelated to the data (e.g., scheduling conflicts, administrative errors, family relocation)

(CONDITIONALLY) MISSING AT RANDOM

- Systematic missingness related to the observed data but unrelated to the unseen latent data

$$f(M = 1 \mid data_{obs}, data_{mis}) = f(M = 1 \mid data_{obs})$$

- Most Big Three applications assume CMAR

Missingness			Predictors of nonresponse					
Indicators			Observed			Missing		
M_1	M_2	M_3	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
0	0	0	4	4	3			
0	1	0	3	NA	5		3	
0	0	0	7	1	6			
1	0	0	NA	1	6	2		
0	0	0	5	9	3			
0	1	1	3	NA	NA		2	2
0	0	0	1	6	7			
0	0	0	9	4	9			
0	1	0	2	NA	6		5	

CONDITIONALLY MAR EXAMPLE

- Missingness is related to the observed learning problems measure but unrelated to the unseen reading scores
- Students with high levels of learnings problems are more likely to have missing data due to increased dropout risk, disciplinary actions, or family or situational instability
- The Big Three assume a CMAR process by default

MISSING NOT AT RANDOM

- Systematic missingness related to the observed data and the unseen latent data

$$f(M = 1 \mid data_{obs}, data_{mis})$$

- The Big Three also allow MNAR processes (selection and pattern mixture models)

Missingness			Predictors of nonresponse					
Indicators			Observed			Missing		
M_1	M_2	M_3	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
0	0	0	4	4	3			
0	1	0	3	NA	5		3	
0	0	0	7	1	6			
1	0	0	NA	1	6	2		
0	0	0	5	9	3			
0	1	1	3	NA	NA		2	2
0	0	0	1	6	7			
0	0	0	9	4	9			
0	1	0	2	NA	6		5	

MNAR EXAMPLE

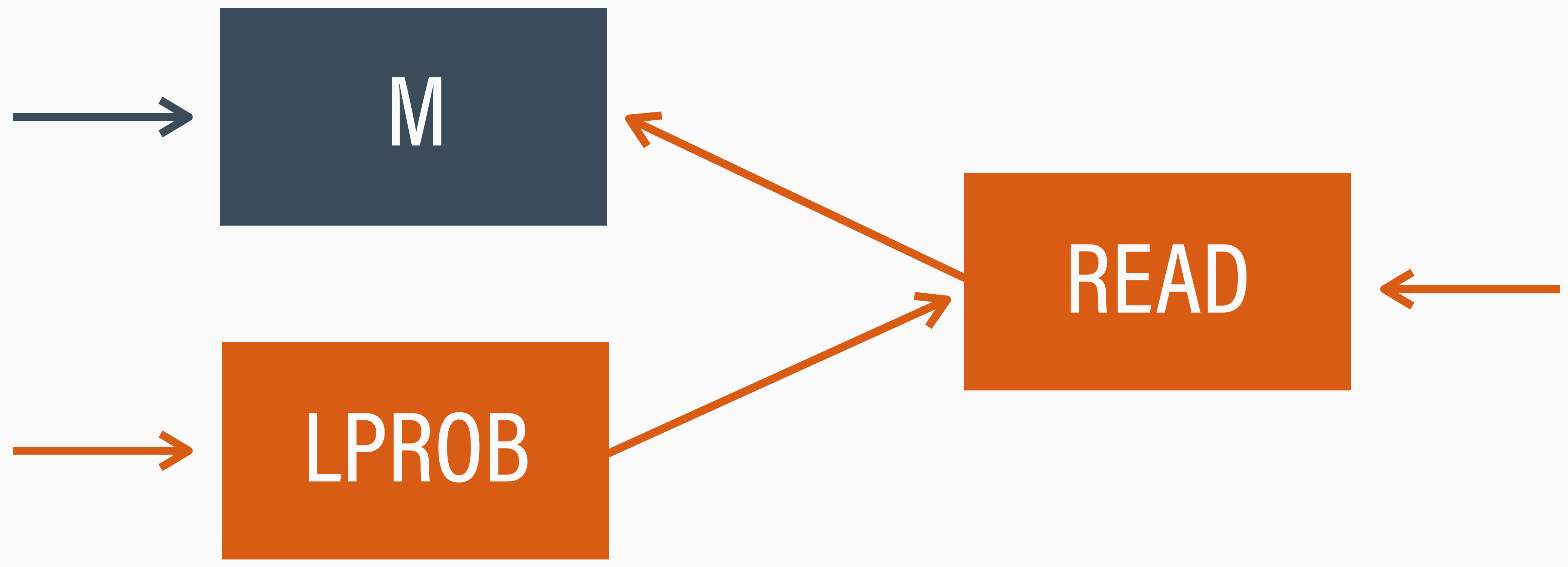
- Missingness is related to the observed learning problems measure and also related to the unseen reading scores

- Individuals with the low reading levels opt out because they feel discouraged or anxious about testing or because they were moved to specialized programs or alternative educational settings where standardized testing protocols differ

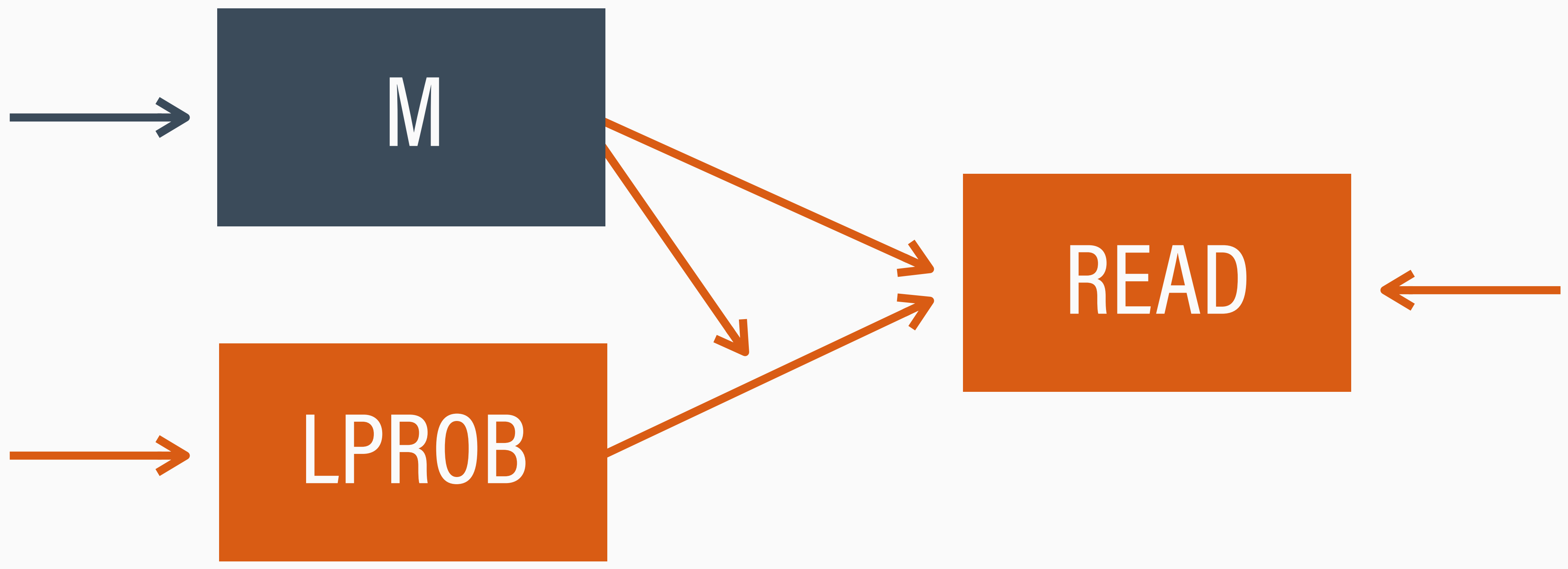
MNAR MODELING

- Missing not at random processes require an explicit model that incorporates the missing data indicator (M)

Selection Model



Pattern Mixture Model



TESTING THE CMAR ASSUMPTION

- The CMAR assumption is untestable because it stipulates no relation between missingness and the unseen scores
- We must rely on logical arguments about why the unseen scores should not be related to missingness
- When in doubt, conduct sensitivity analyses that compare the estimates from CMAR and MNAR assumptions

OUTLINE

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. MCMC Estimation and Missing Data Imputation
4. Fitting Regression Models in Blimp
5. Incomplete Categorical Variables
6. Interaction Effects
7. Reporting Results From a Missing Data Analysis

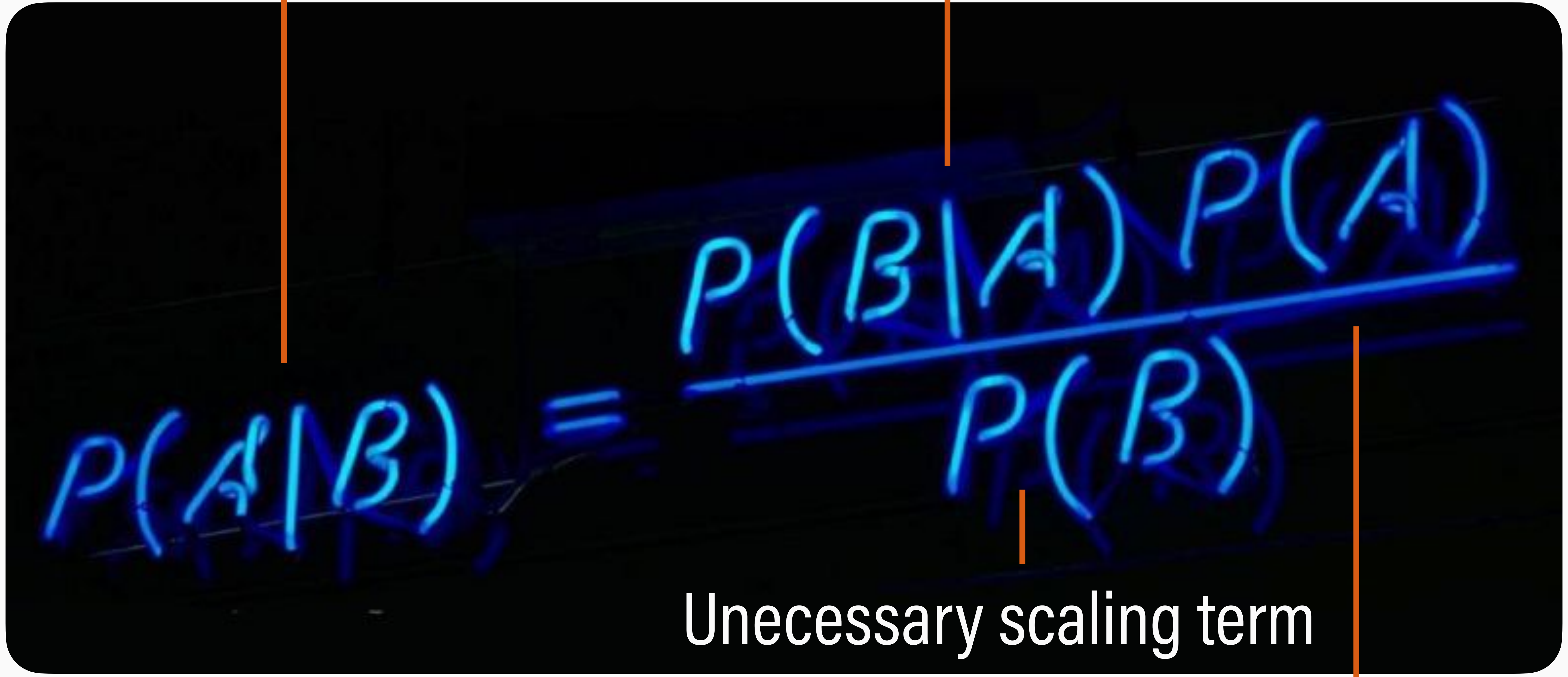
FREQUENTIST VS. BAYESIAN PARADIGMS

Frequentist	Bayesian
The parameter is a fixed quantity, estimates vary across different samples	Parameters are random variables with a distribution of plausible realizations
Statements about probability, precision, and confidence refer to estimates	Statements about probability, precision, and intervals refer to the parameter
Probability = long run frequency of outcomes across many samples	Probability = our degree of certainty about a parameter after analyzing data

BAYES' THEOREM

Posterior = parameters (A) given the data (B)

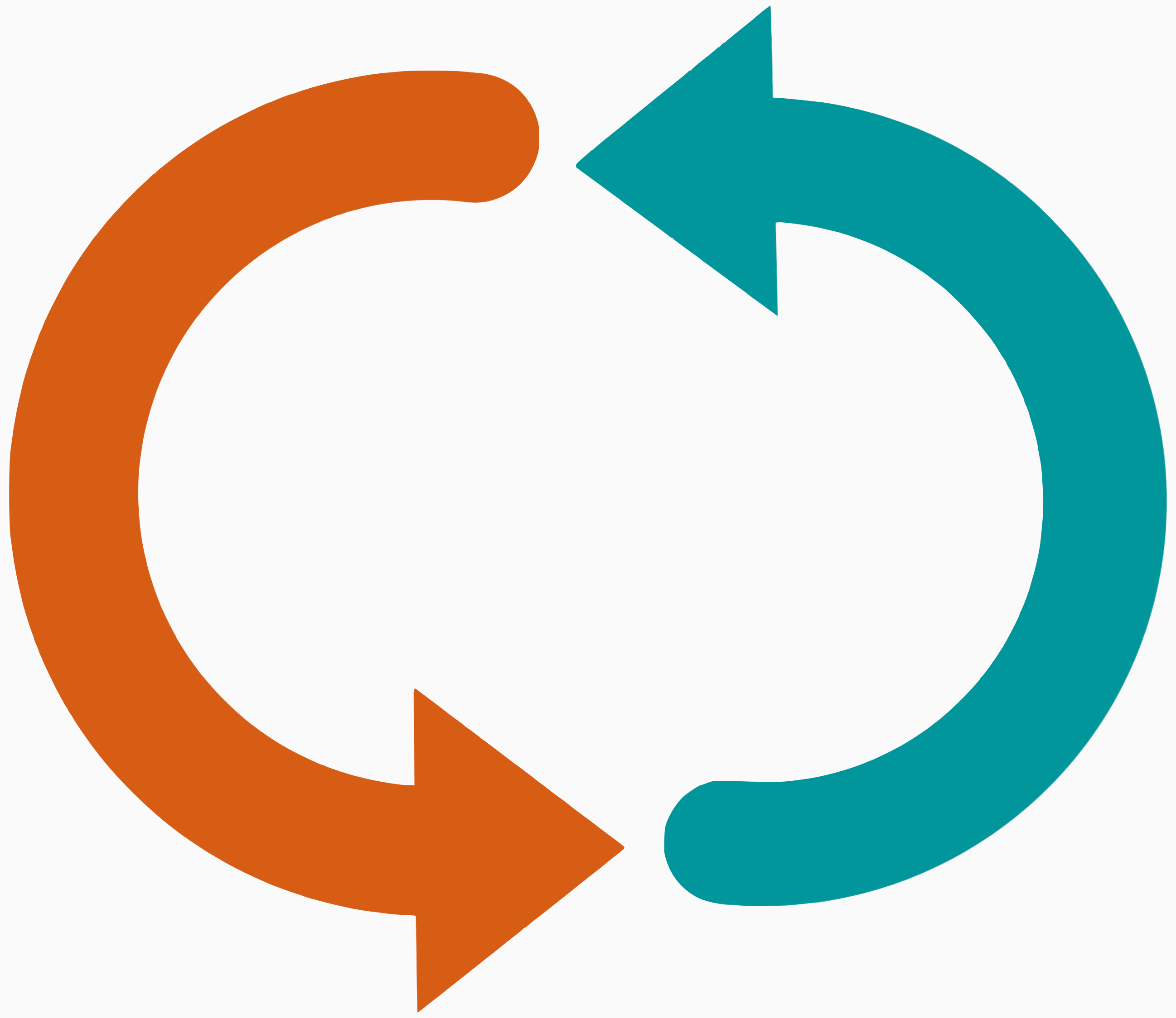
Frequentist likelihood = data (B) given the parameters (A)



Prior = a priori belief about parameters (A)

MCMC ESTIMATION

Estimate regression models



Impute missing values

Do for t = 1 to 10,000 iterations

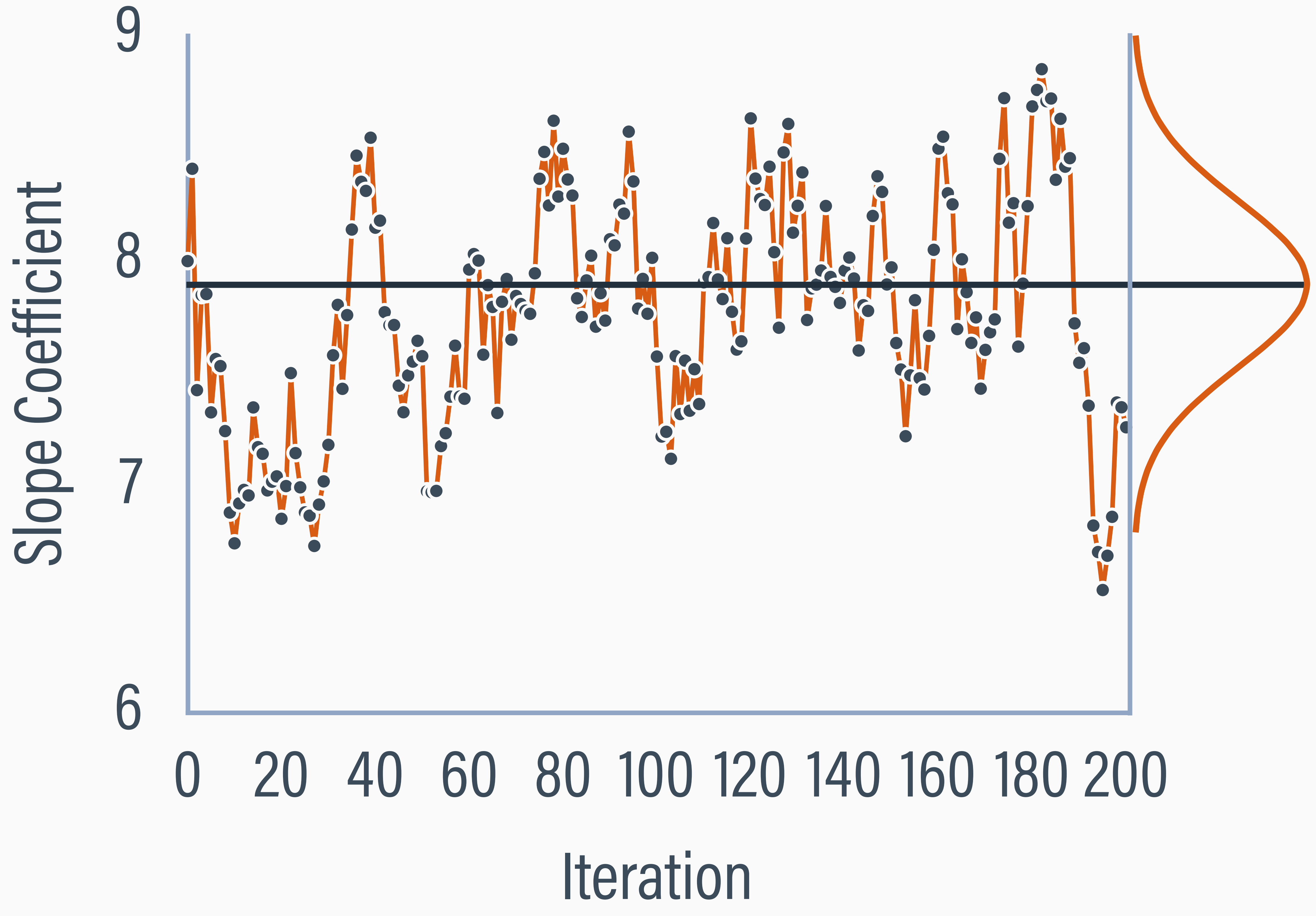
- Estimate model parameters, conditional on the filled-in data
- Impute missing values, conditional on the model parameters

Repeat

Summarize model parameters

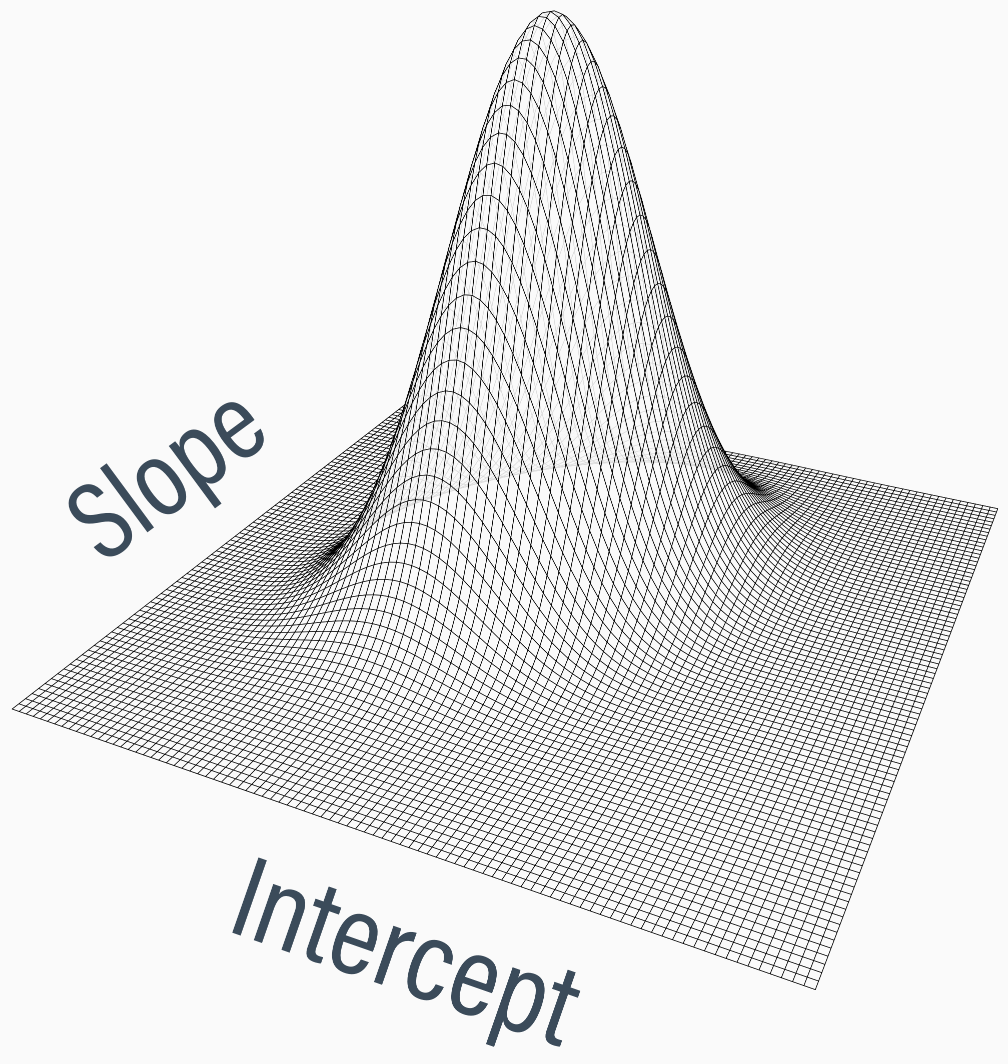
MEANING OF ESTIMATION

- MCMC uses computer simulation to “sample” parameters from a distribution
- Estimates continually vary across iterations in a random pattern
- Each iteration gives plausible parameter values that could have produced our data

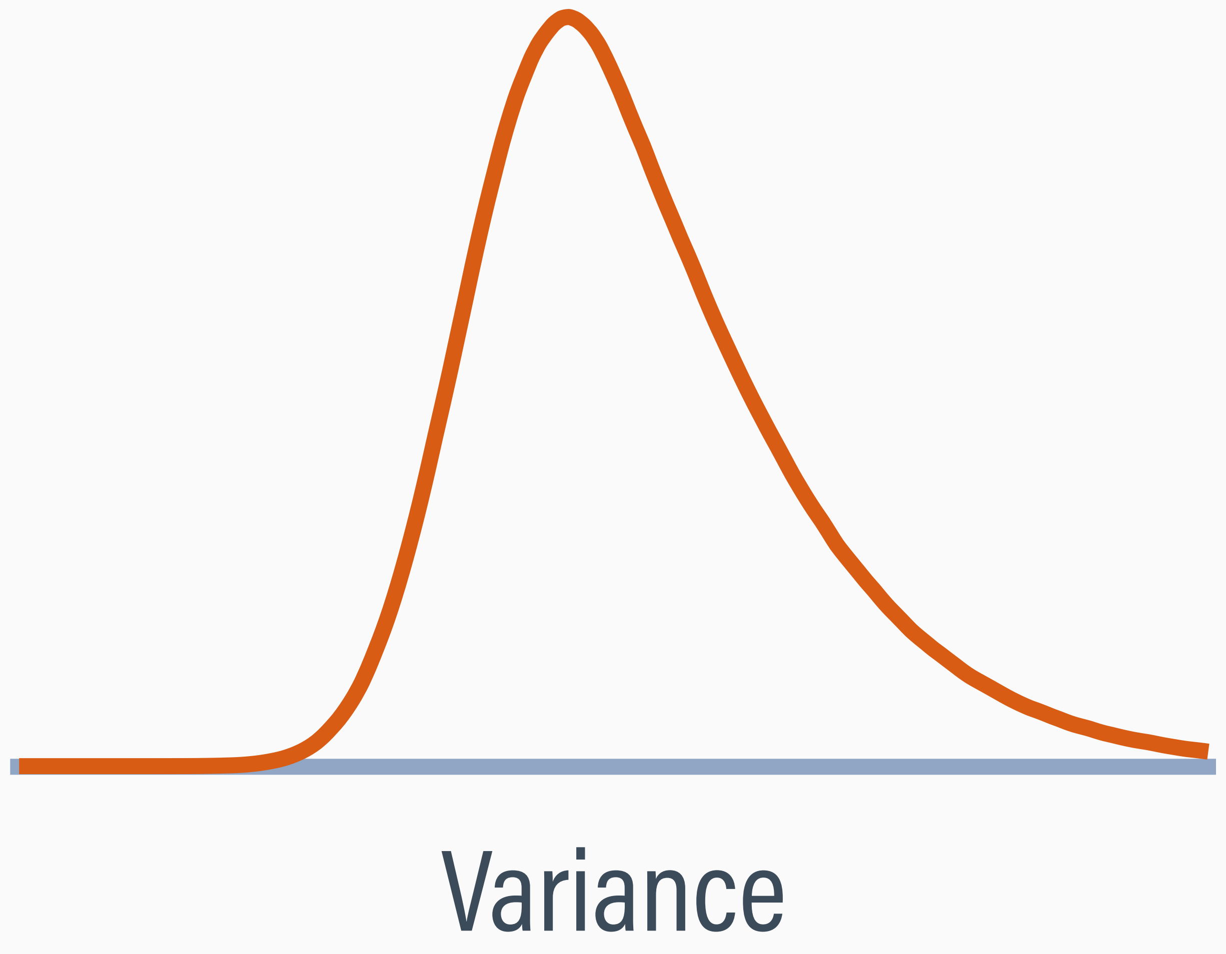


PARAMETER-GENERATING DISTRIBUTIONS

- MCMC draws coefficients from a multivariate normal distribution, with least-squares estimates defining shape



- MCMC draws variances from an inverse gamma distribution with its shape determined by the df and residual SS



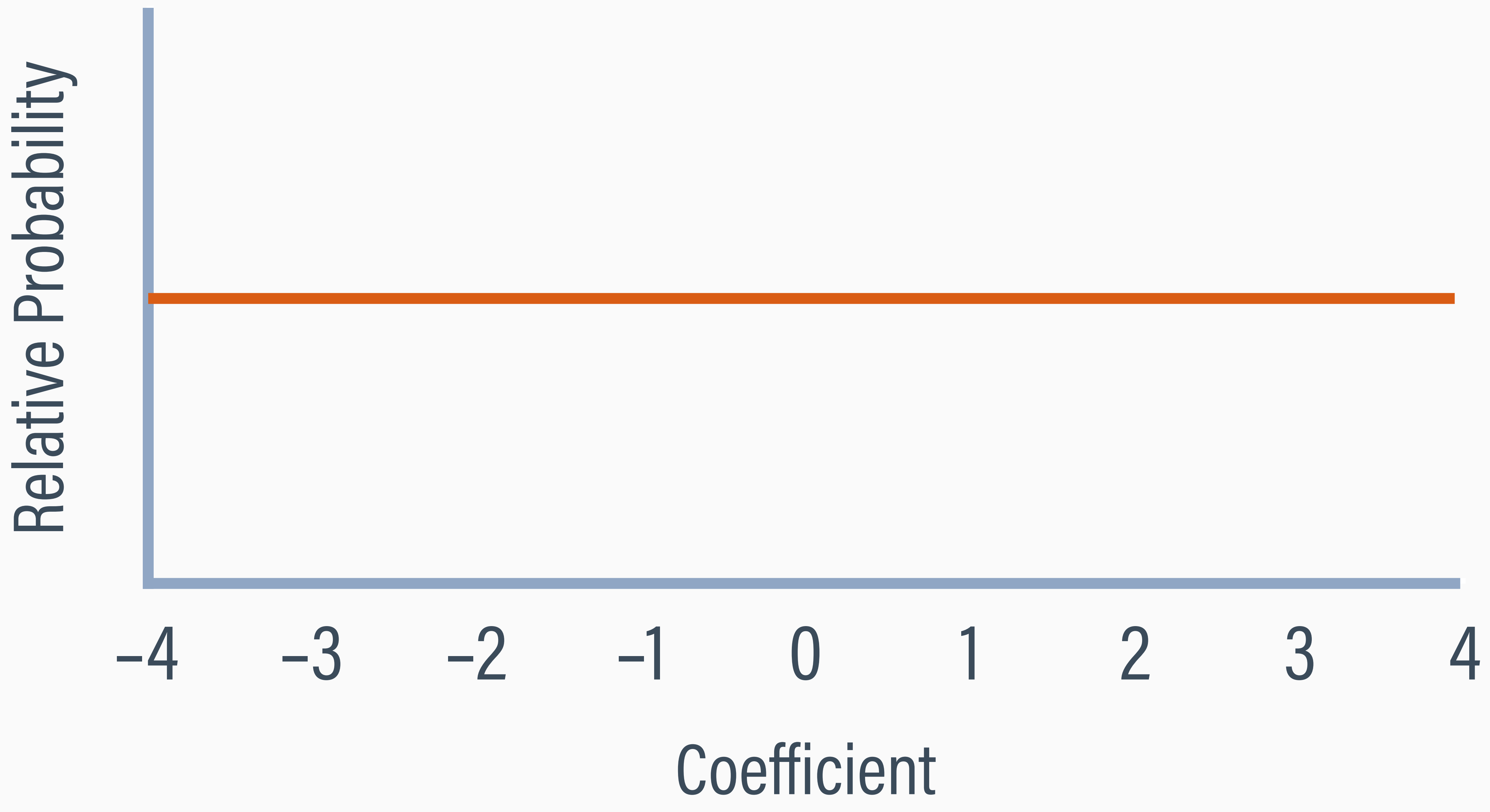
PARAMETERS FROM 200 MCMC CYCLES

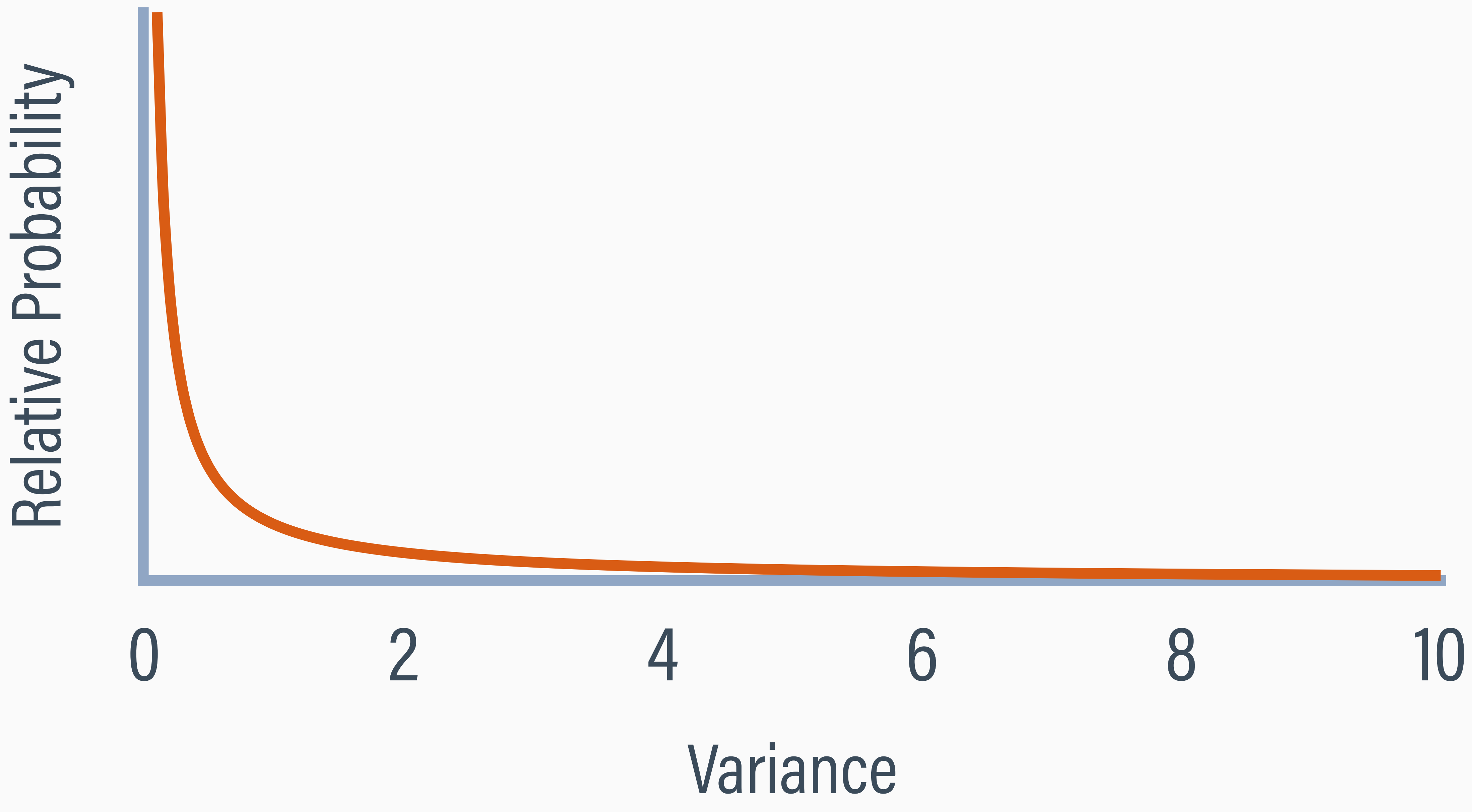
PRIOR DISTRIBUTIONS

- Bayesian analyses require prior distributions that encode our beliefs about the parameter values prior to analyzing the data
- Blimp adopts non-informative (diffuse) priors that impart as little information as possible (i.e., let the data do the talking)
- Prior distributions for variances can influence estimates when the N is small, so sensitivity analyses may be warranted (FIML variance estimates are also biased in the same scenarios)

PRIOR DISTRIBUTIONS

- A diffuse prior for means and coefficients conveys that all possible parameter values are equally likely a priori
- Diffuse priors for variances are slightly informative, and different options function like df adjustments in regression



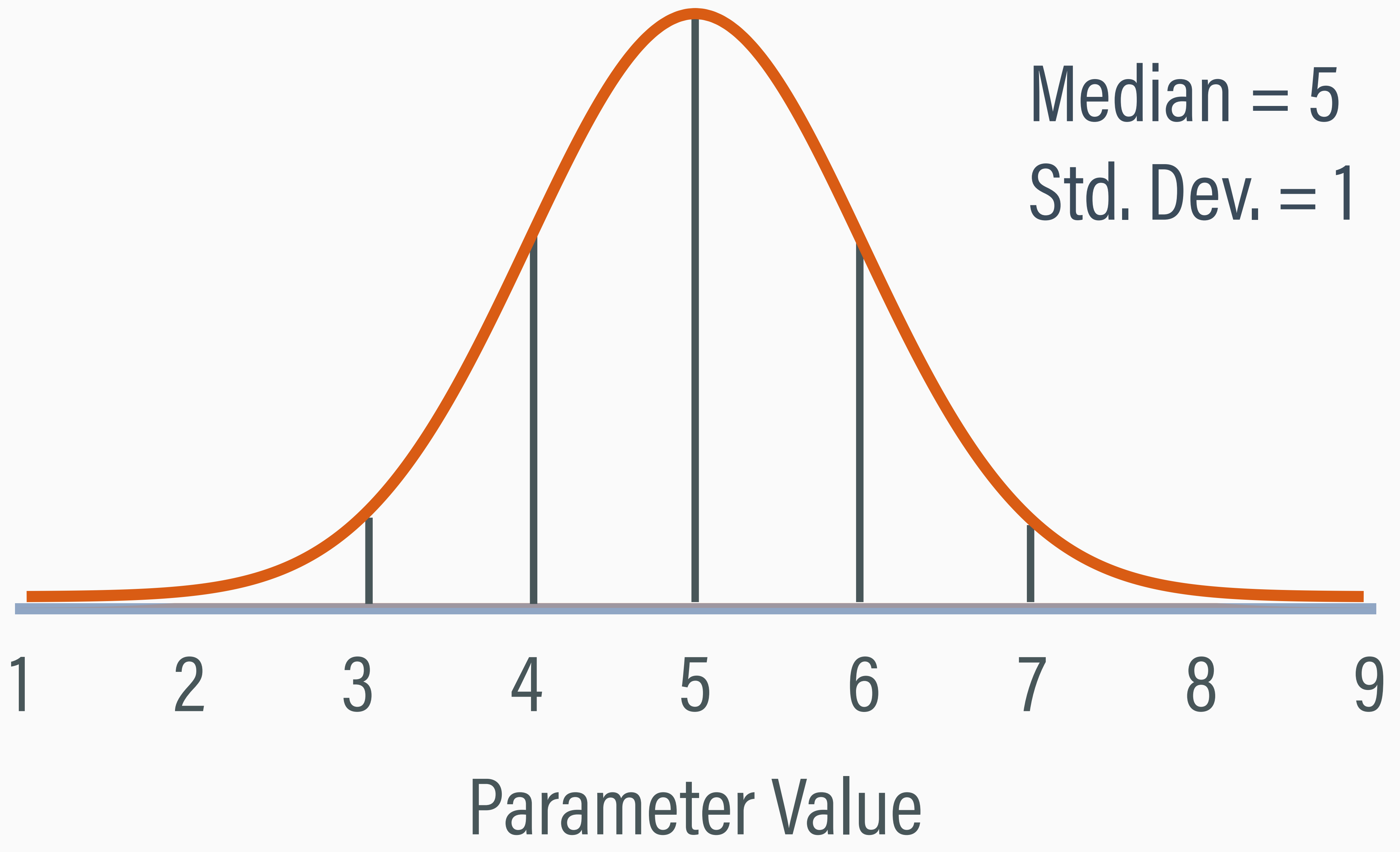


SUMMARIZING MCMC ESTIMATES

- MCMC iterates for thousands of cycles, and each cycle produces estimates based on one filed-in data set
- Bayesian estimation yields a distribution of parameters—called a posterior—that averages over thousands of imputations
- The posterior is a distribution of plausible parameter values that could have produced our particular data

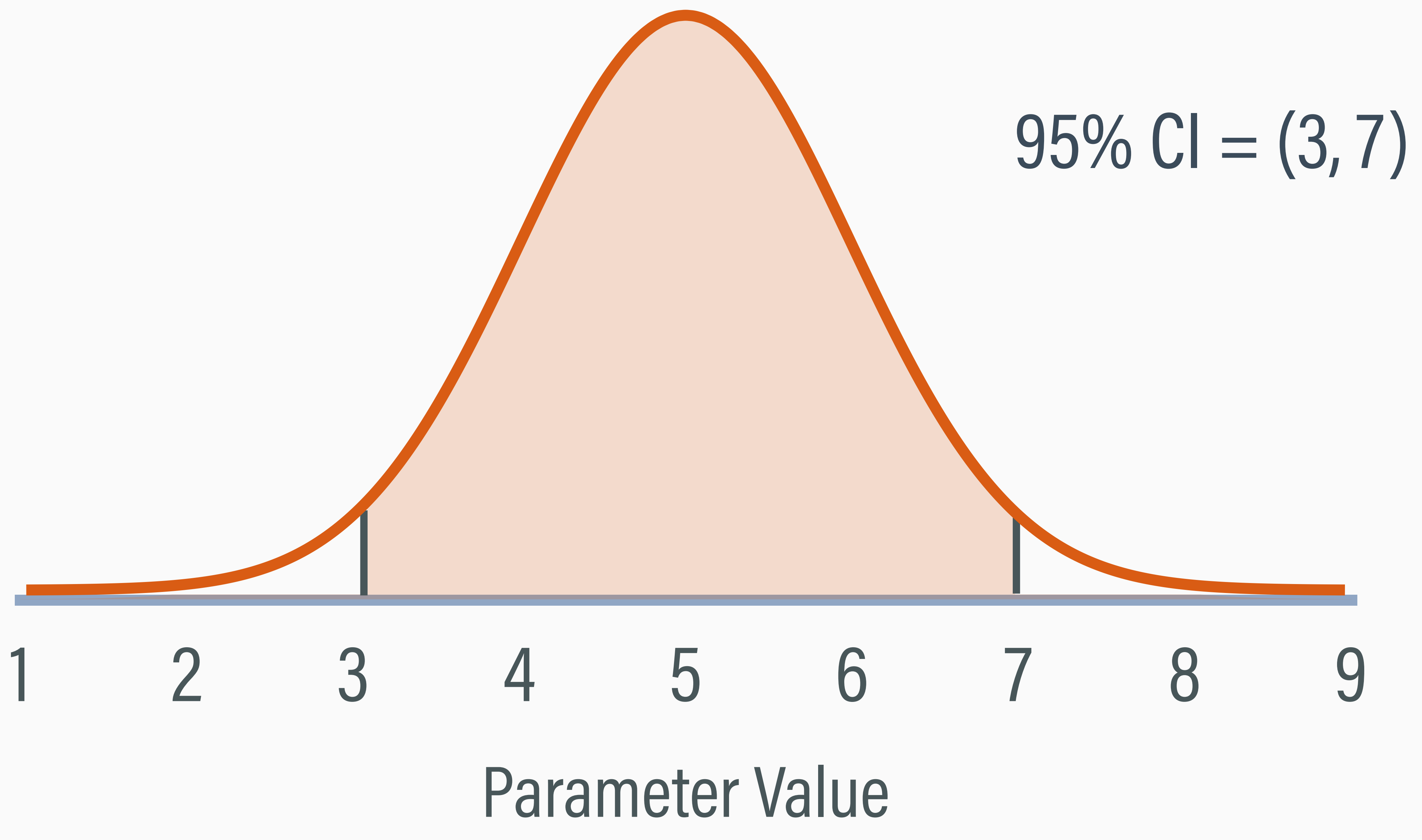
POSTERIOR MEDIAN AND STD. DEV.

- 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



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



SIMPLE REGRESSION ILLUSTRATION

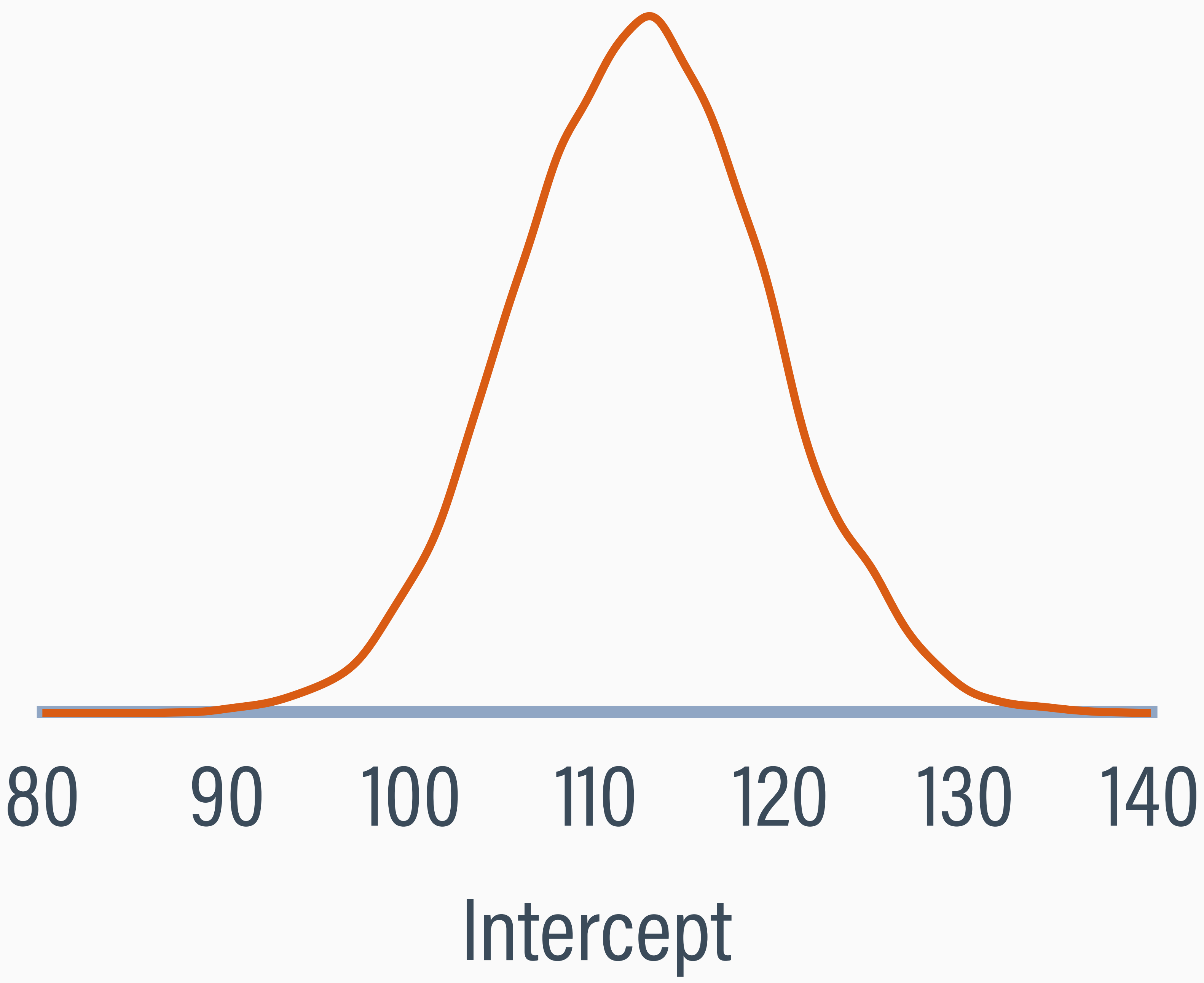
- Study that seeks to determine whether reading levels in 1st grade predict 9th grade reading achievement in middle school

$$read_9 = \beta_0 + \beta_1(lrnprob_1) + \varepsilon$$

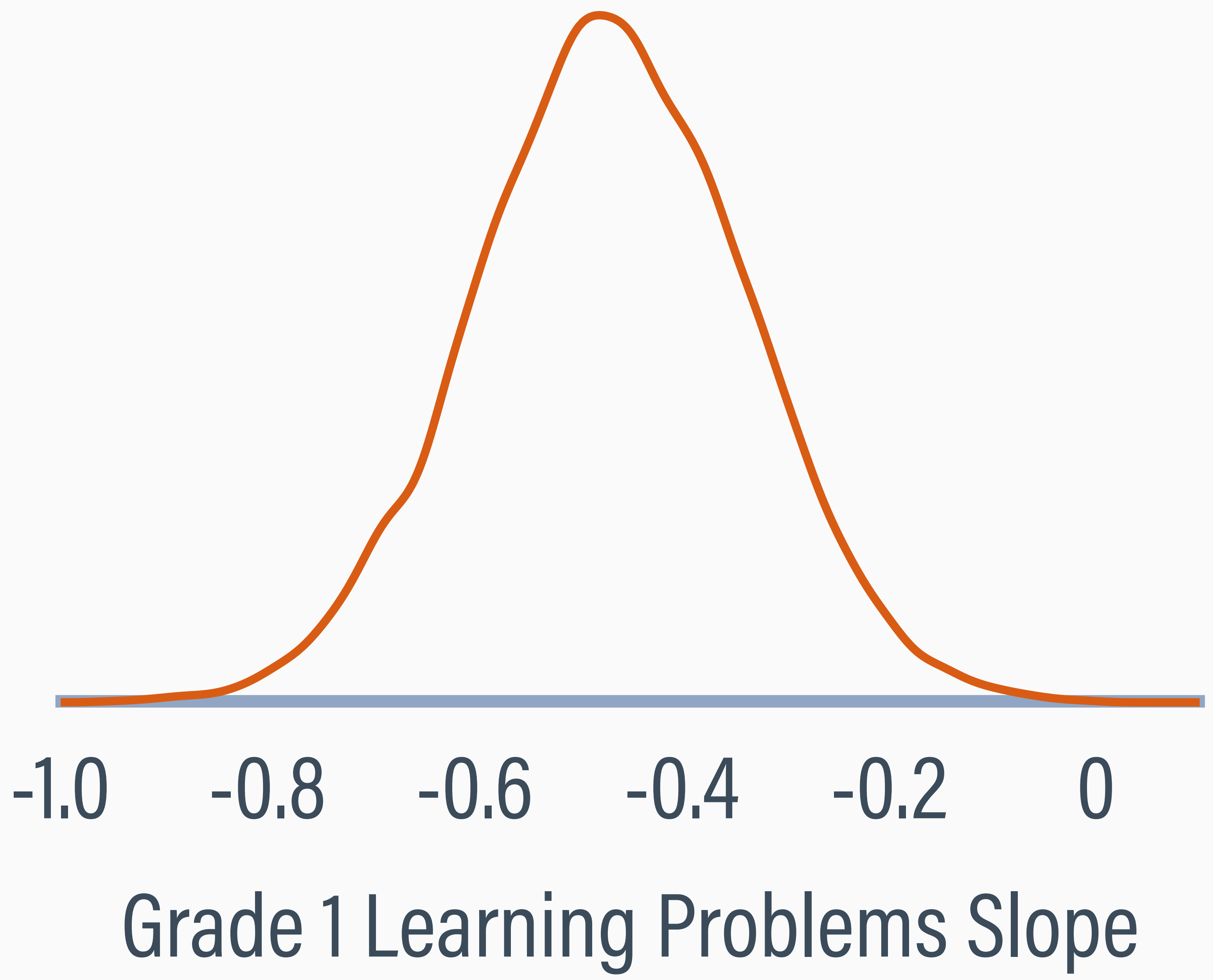
Variable	Definition	Missing %	Scale
atrisk	Emotional/behavioral risk code	2.2	0 = Low, 1 = Medium/high
lrnprob1	1st grade learning problems	2.2	Numeric (31 to 88)
read1	1st grade broad reading composite	6.5	Numeric (39 to 153)
read9	9th grade broad reading composite	17.4	Numeric (41 to 123)

POSTERIOR DISTRIBUTIONS

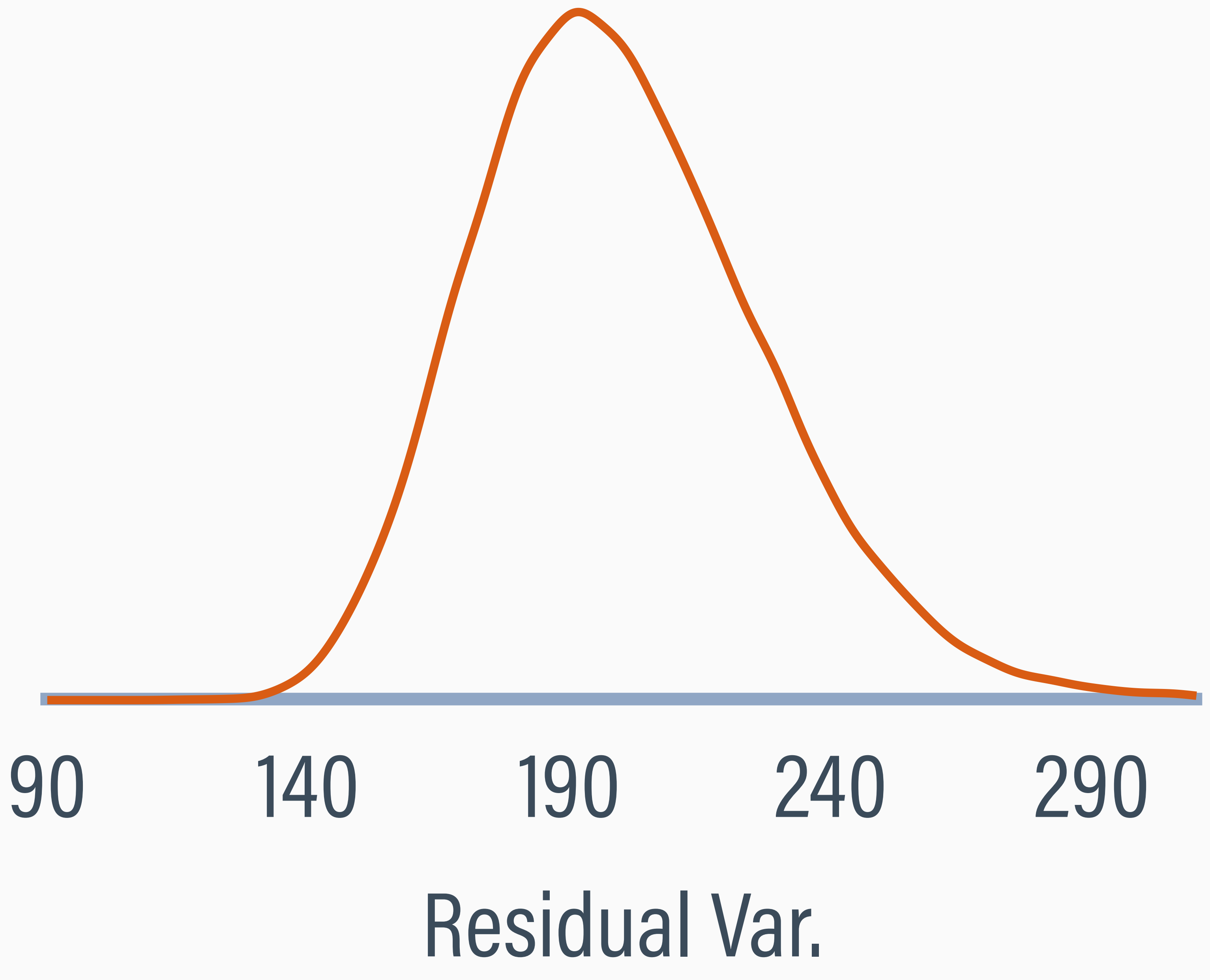
Median = 112.48
Std. Dev. = 6.81
95% CI = (99.09, 125.87)



Median = −0.47
Std. Dev. = 0.13
95% CI = (−0.73, −0.21)



Median = 197.63
Std. Dev. = 27.62
95% CI = (153.26, 260.29)



ESTIMATOR COMPARISON

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

	MCMC			FIML		
Parameter	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 ML estimates and frequentist inference (Levy & McNeish, 2021)
- MCMC is simply a flexible way to estimate frequentist quantities in cases where FIML solutions are unavailable (e.g., missing data)

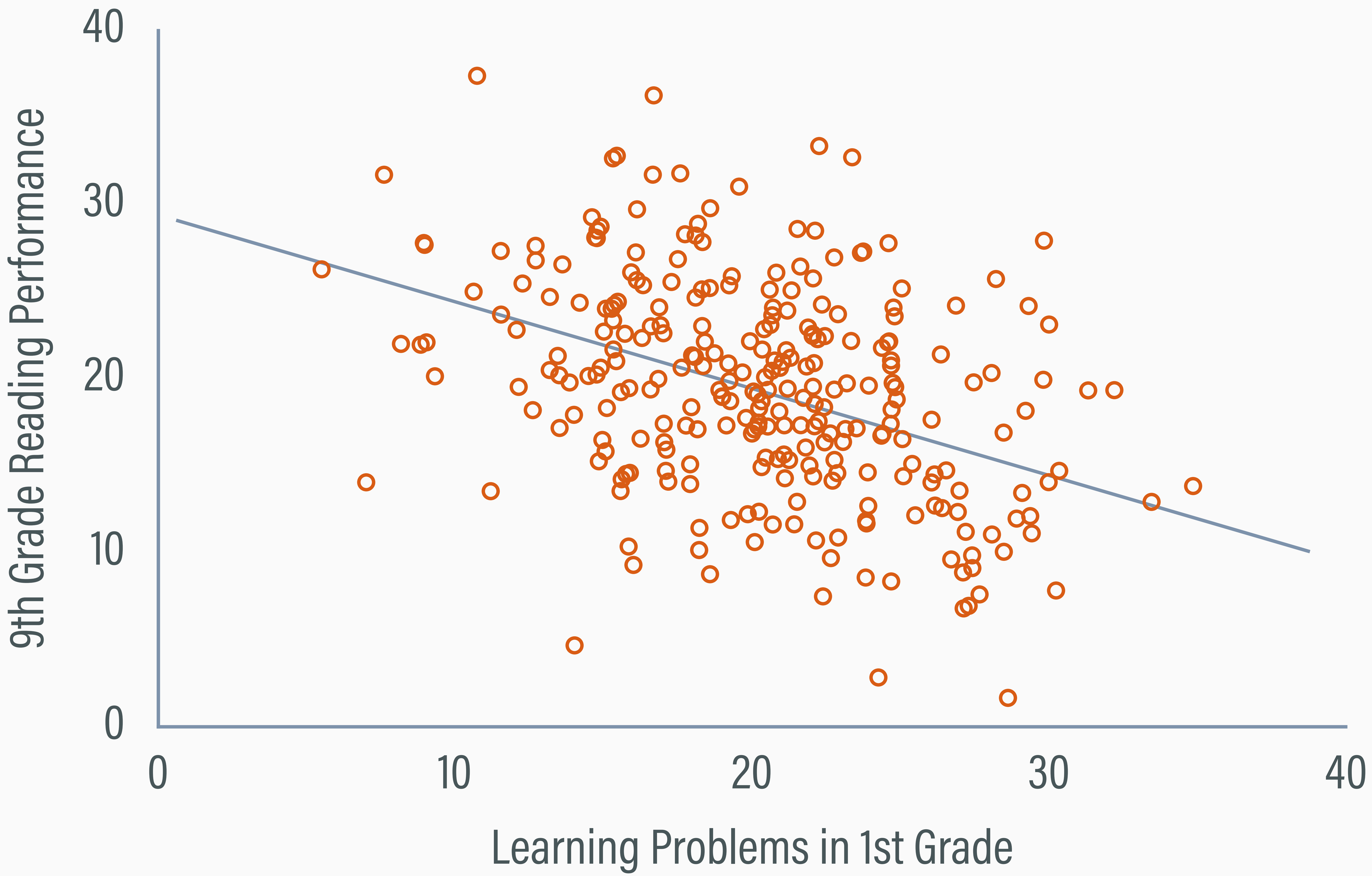
Bayesian Inference

Computational frequentism

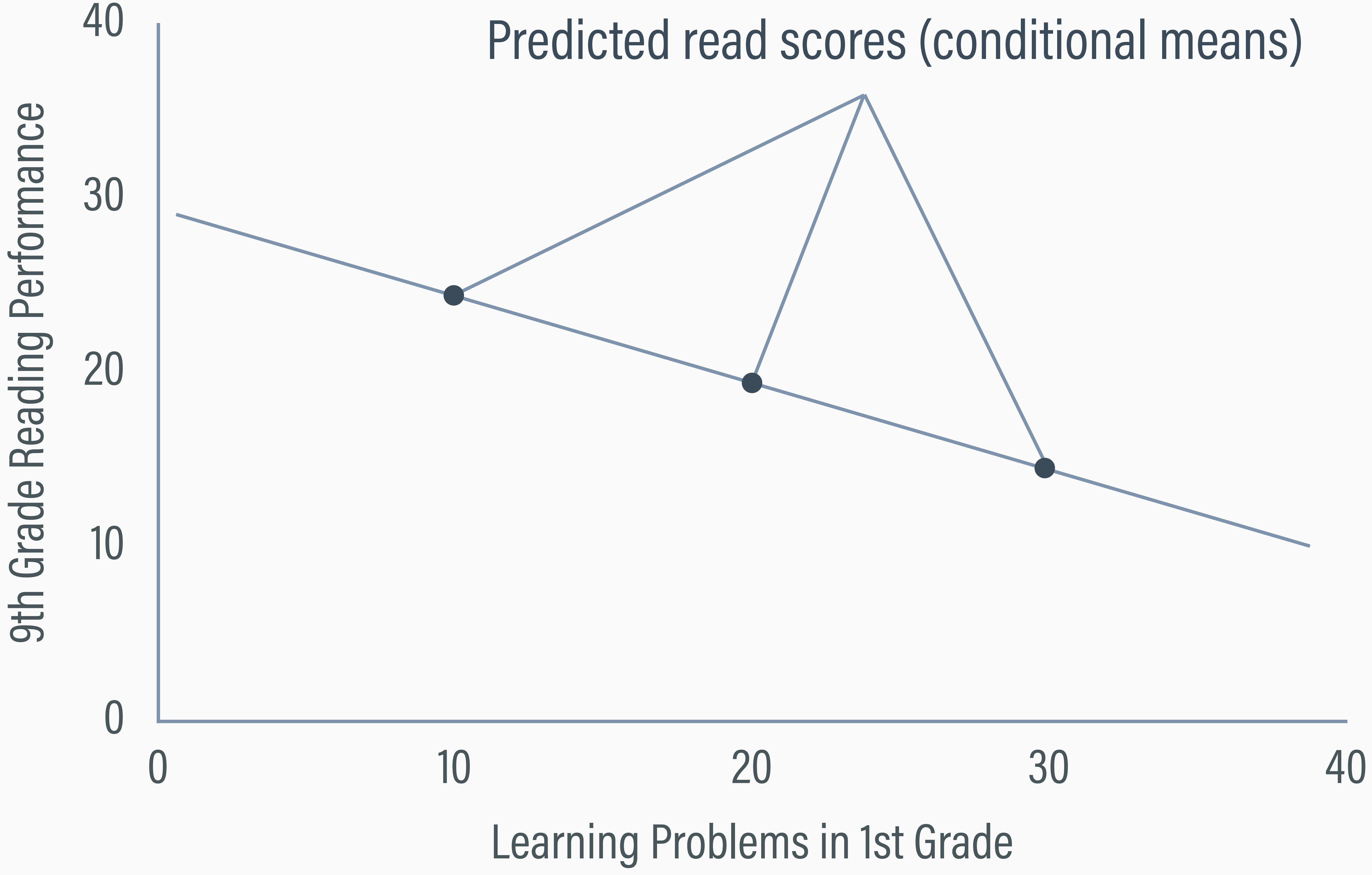
MISSING DATA IMPUTATION STEP

- Missing scores are imputed by drawing replacement scores at random from a distribution of plausible values
- The model parameters combine to define the center and spread of the missing data imputations
- Each iteration yields unique model parameters and unique imputations based on those parameters

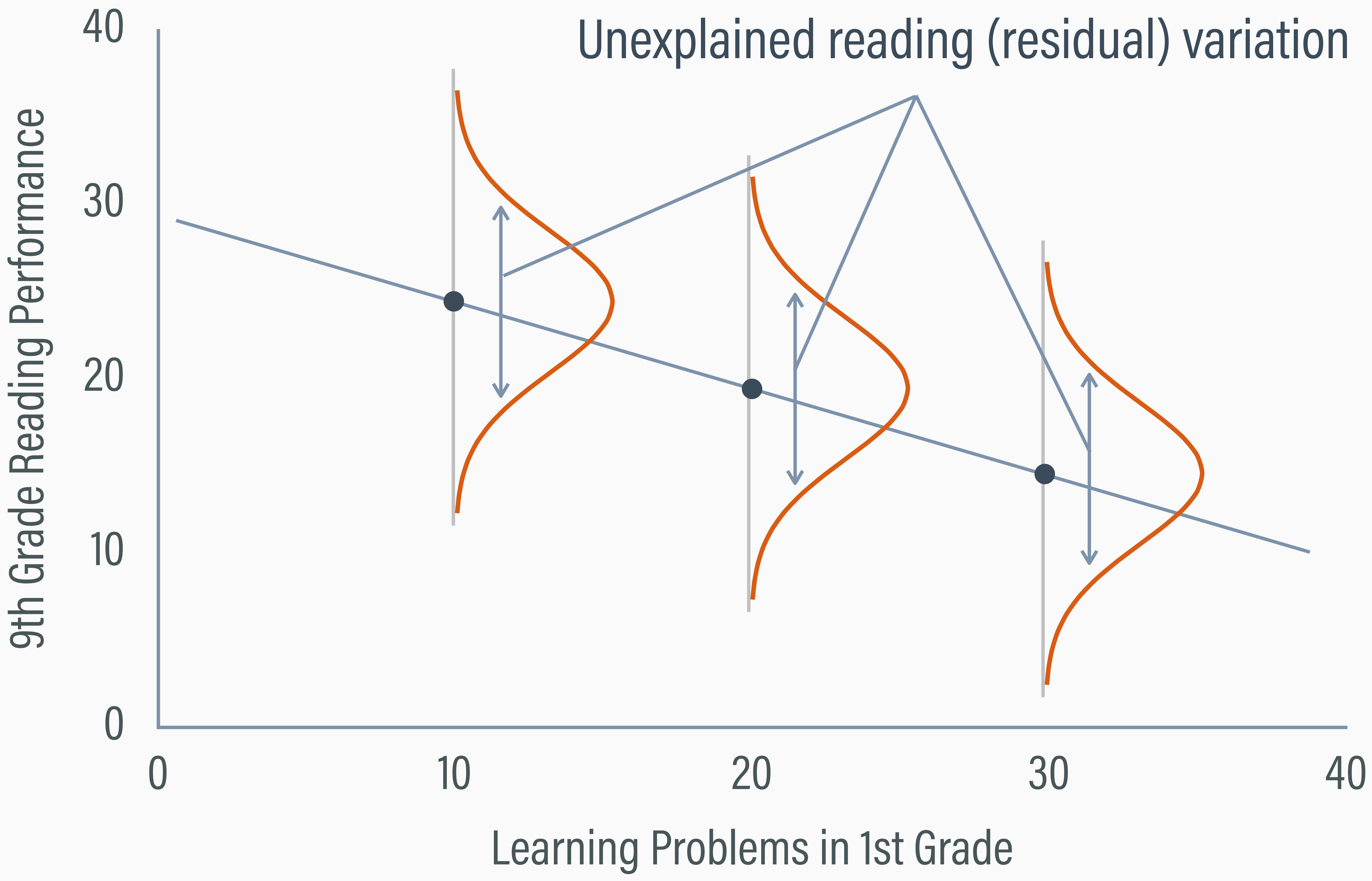
REGRESSION FROM ONE FILLED-IN DATA SET



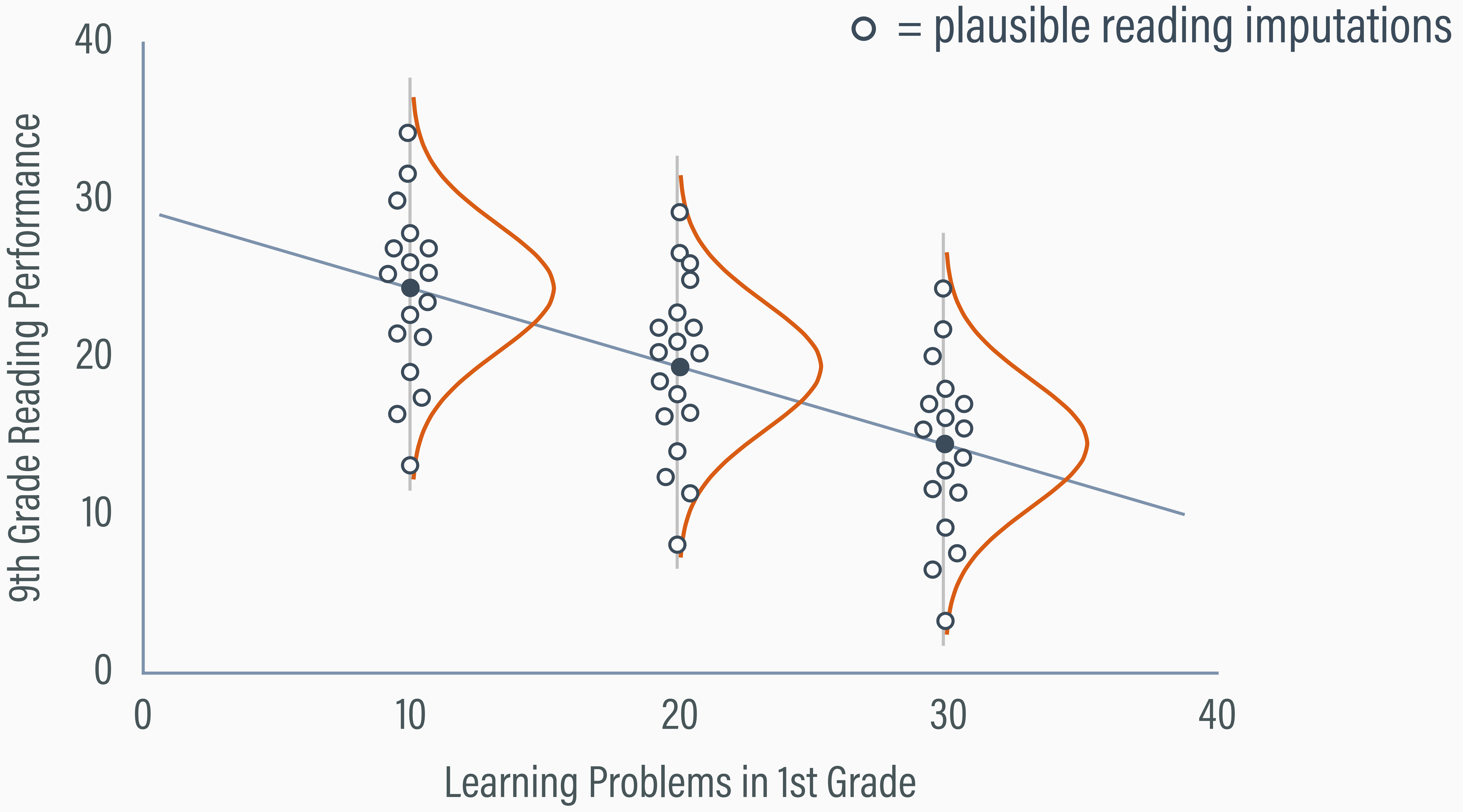
PREDICTED VALUES



RESIDUAL VARIATION

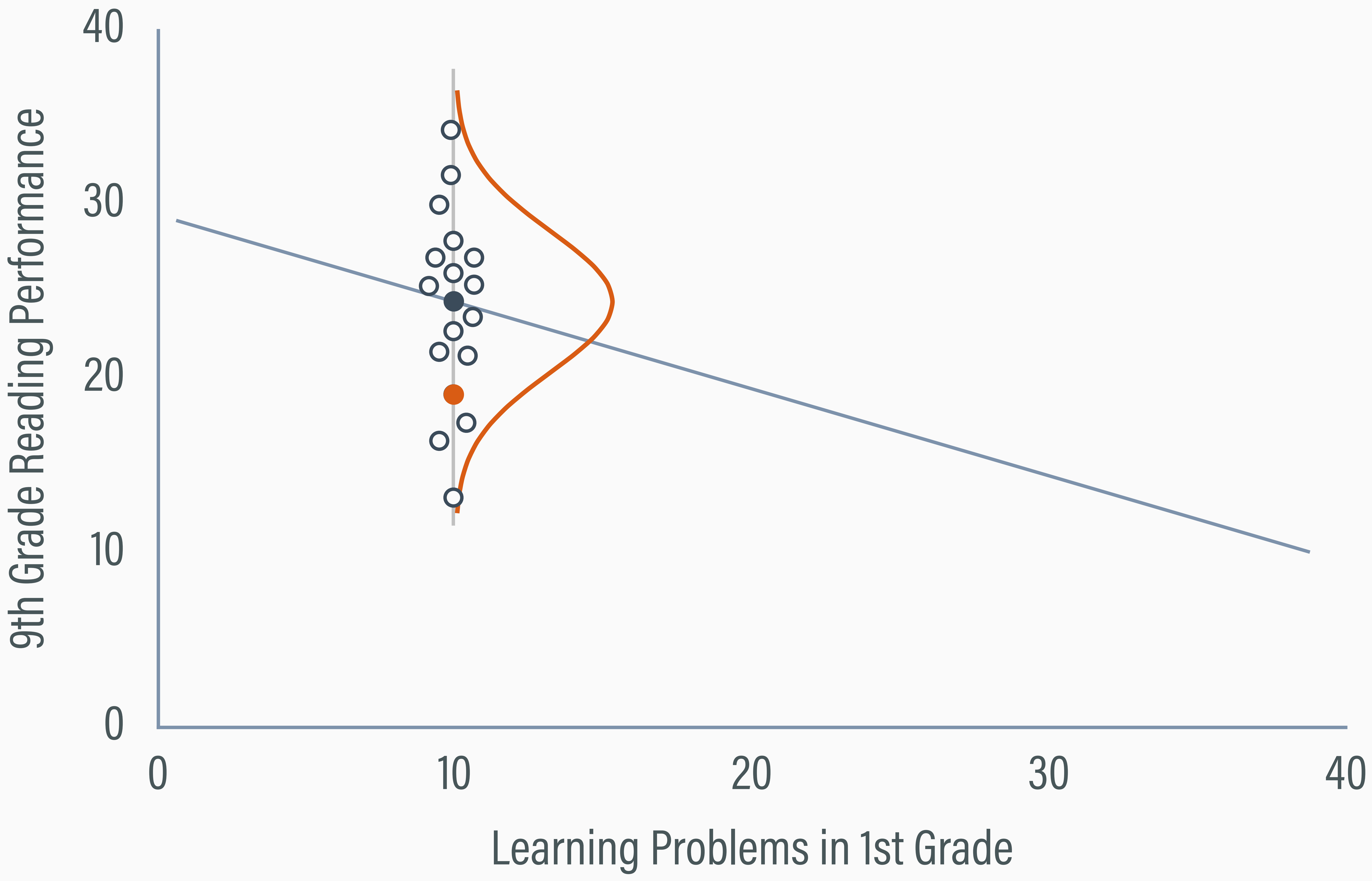


DISTRIBUTIONS OF IMPUTATIONS



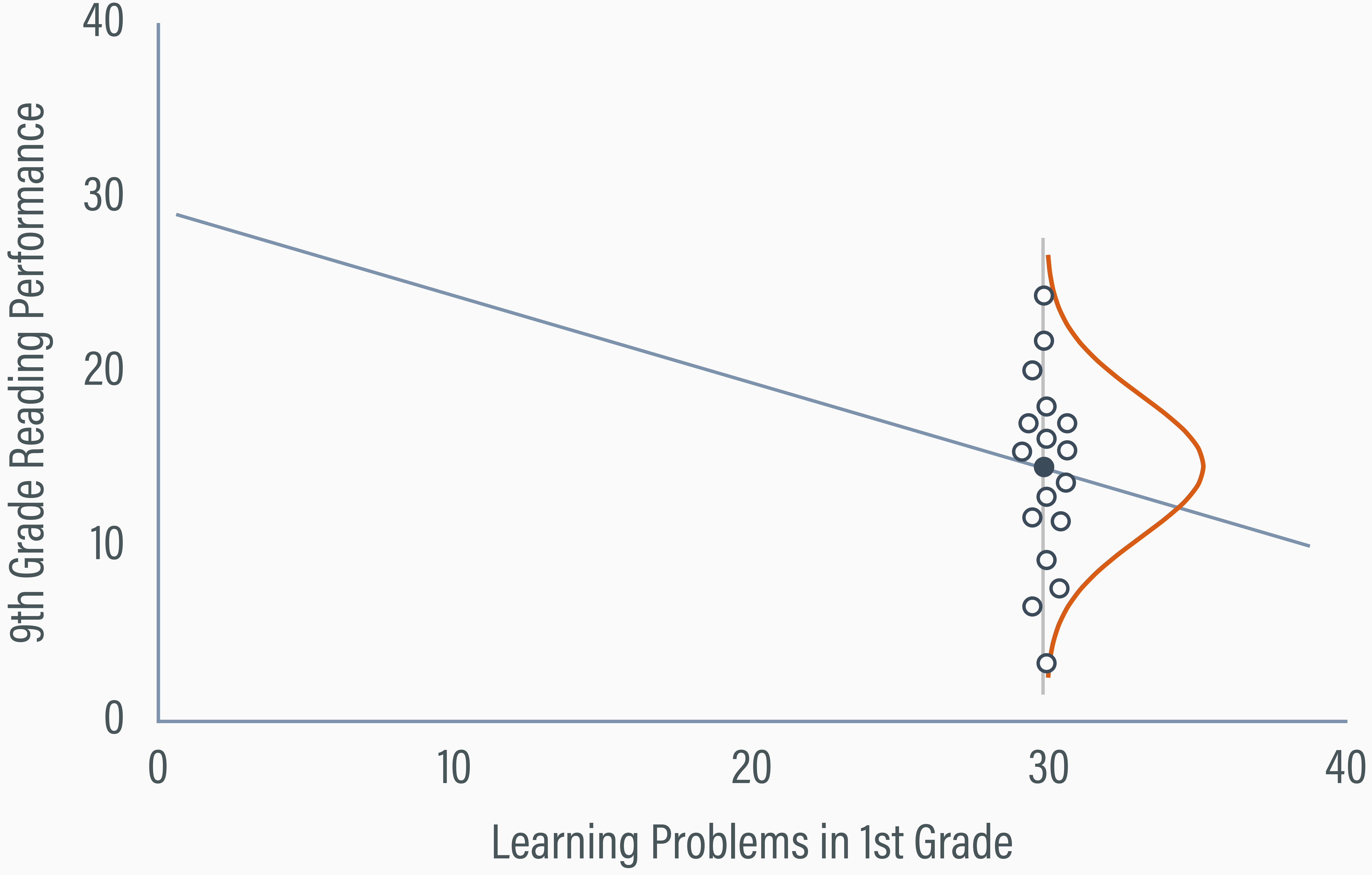
IMPUTATION FOR LOW LEARNING PROBLEMS

DRAW AN IMPUTATION AT RANDOM

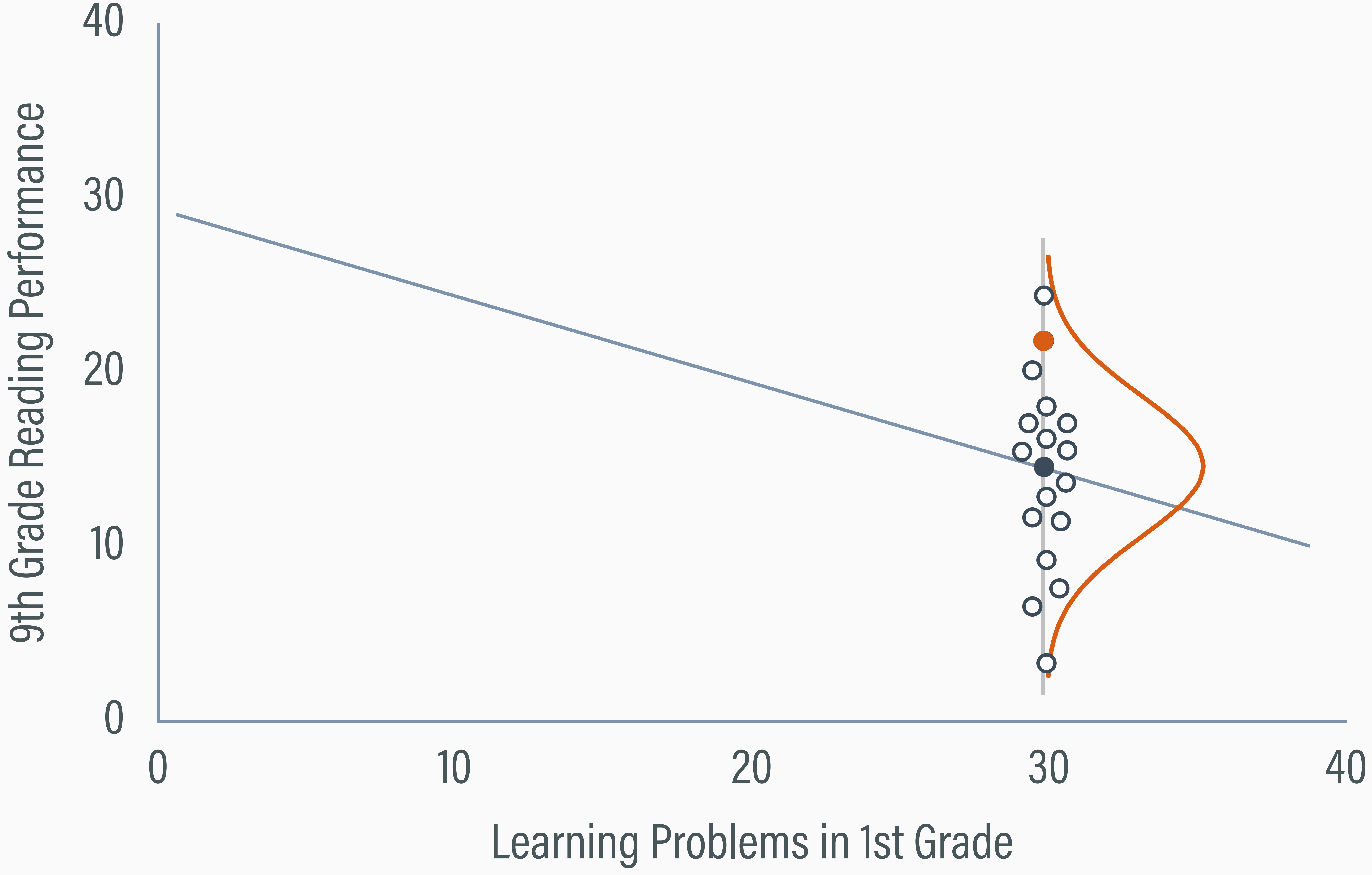


IMPUTATION = PREDICTION + NOISE

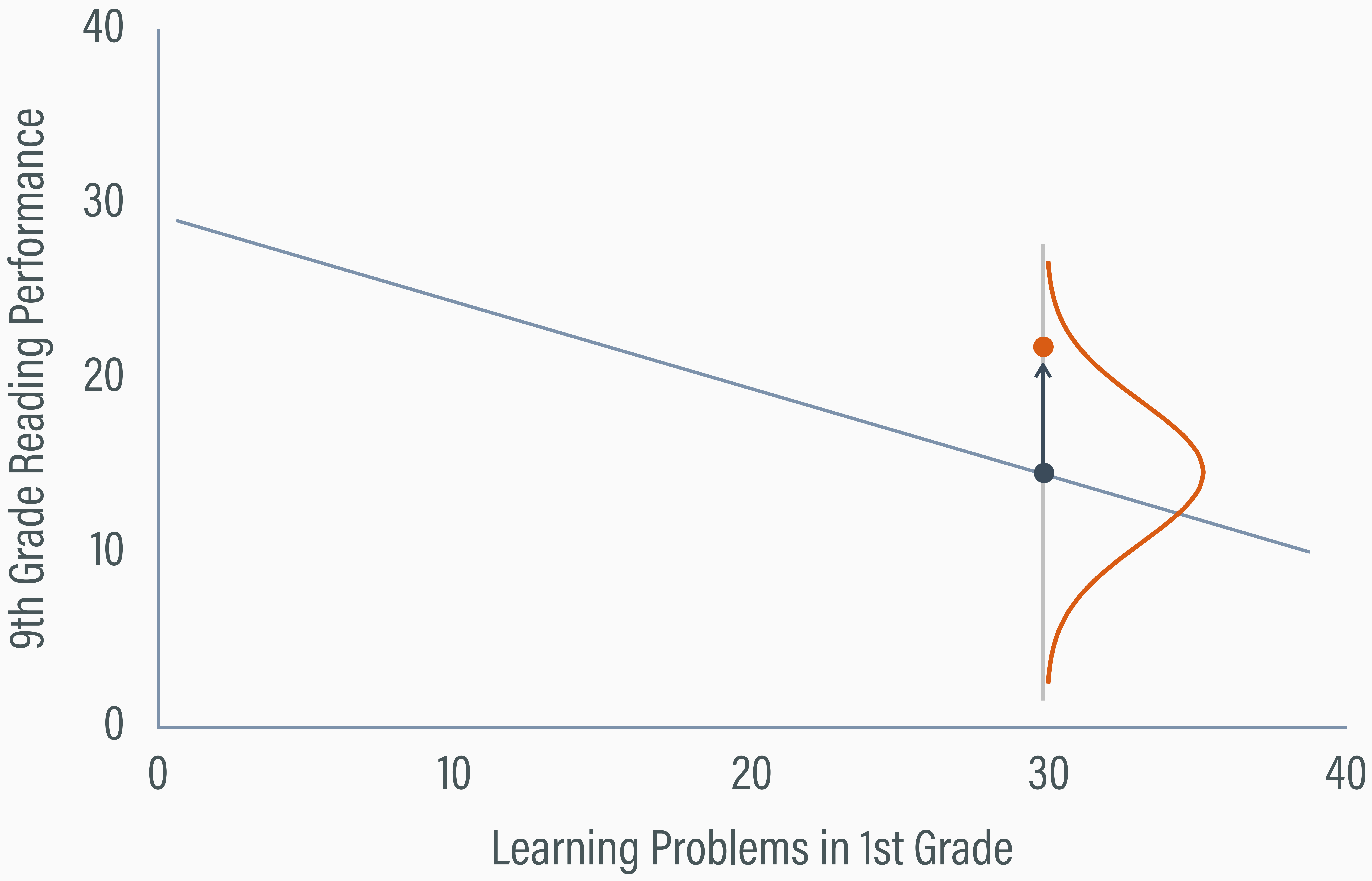
IMPUTATION FOR HIGH LEARNING PROBLEMS



DRAW AN IMPUTATION AT RANDOM



DRAW AN IMPUTATION AT RANDOM

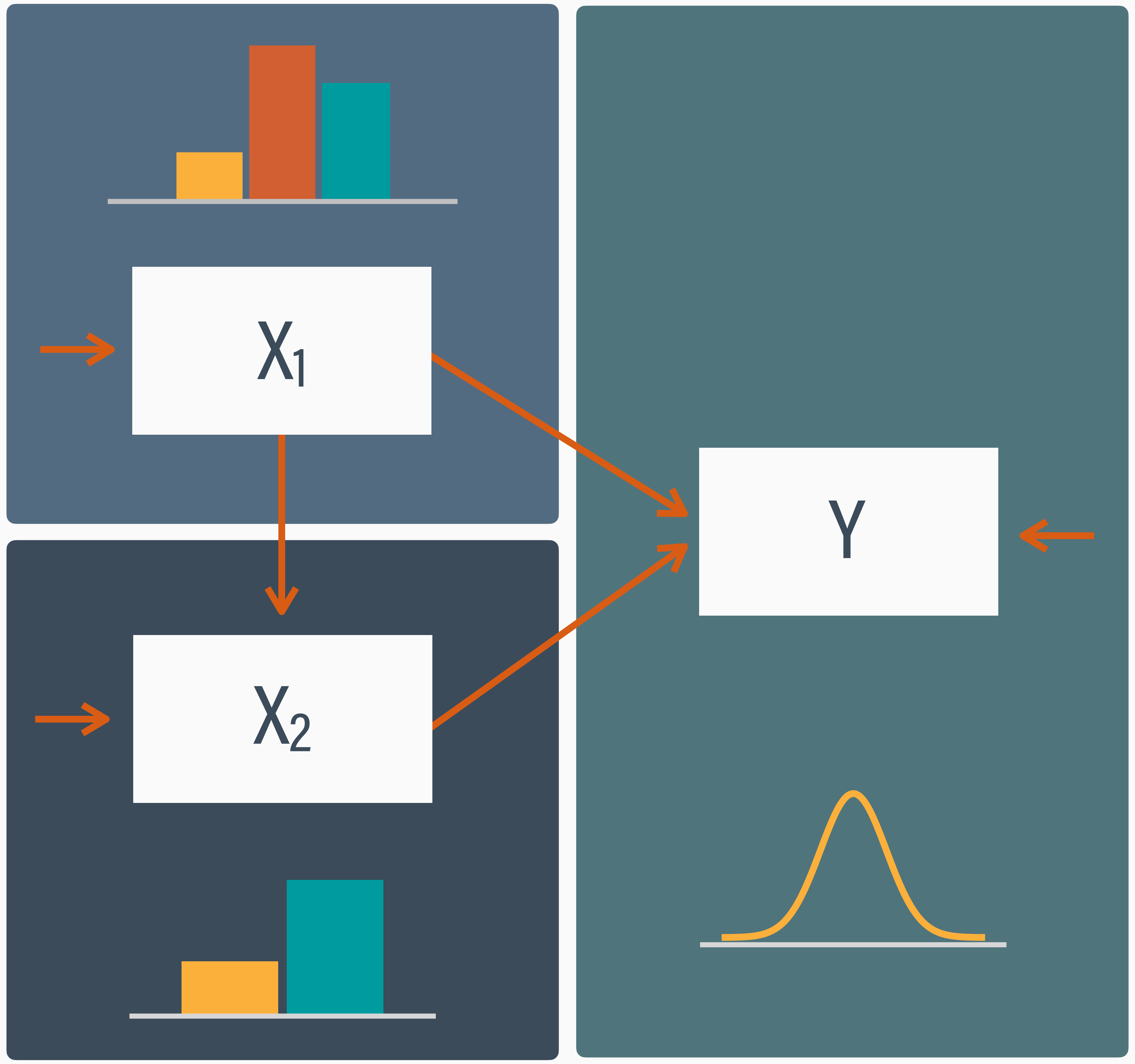


INCOMPLETE PREDICTORS

- Incomplete predictors require their own model and distributional assumptions
- Multivariate normal methods (e.g., FIML) can mis-specify the data distributions in a way that introduces bias
- Factored regression uses a modular specification where a sequence of models replaces a general multivariate model

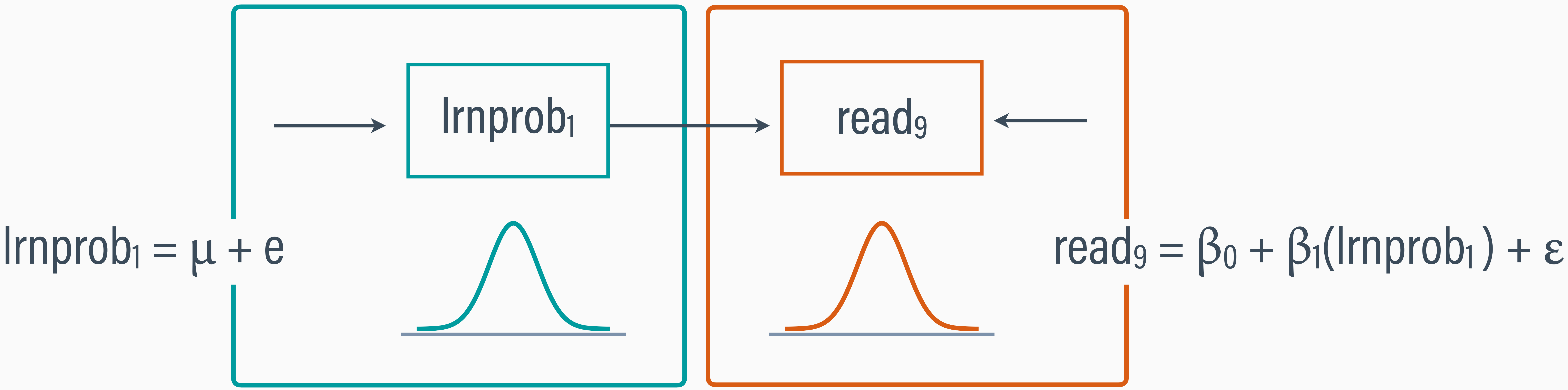
FACTORED REGRESSION SPECIFICATIONS

- Factored regression specifications invoke a unique distribution for each variable
- The analysis consists of a collection of univariate regression models
- Each model can include terms that are at odds with multivariate normality



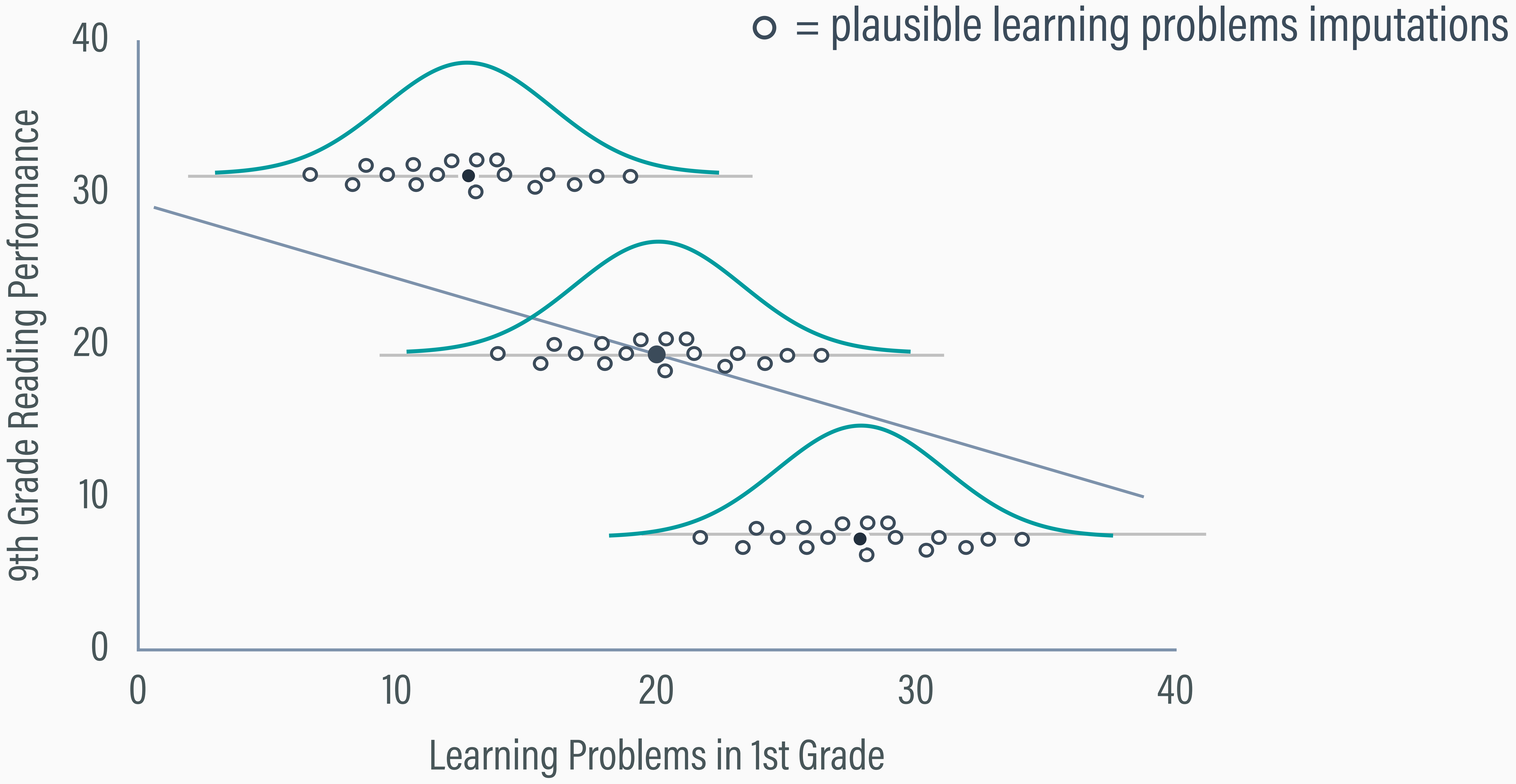
INCOMPLETE PREDICTORS

- Learning problems is the regressor in the focal model and an outcome in its own empty model



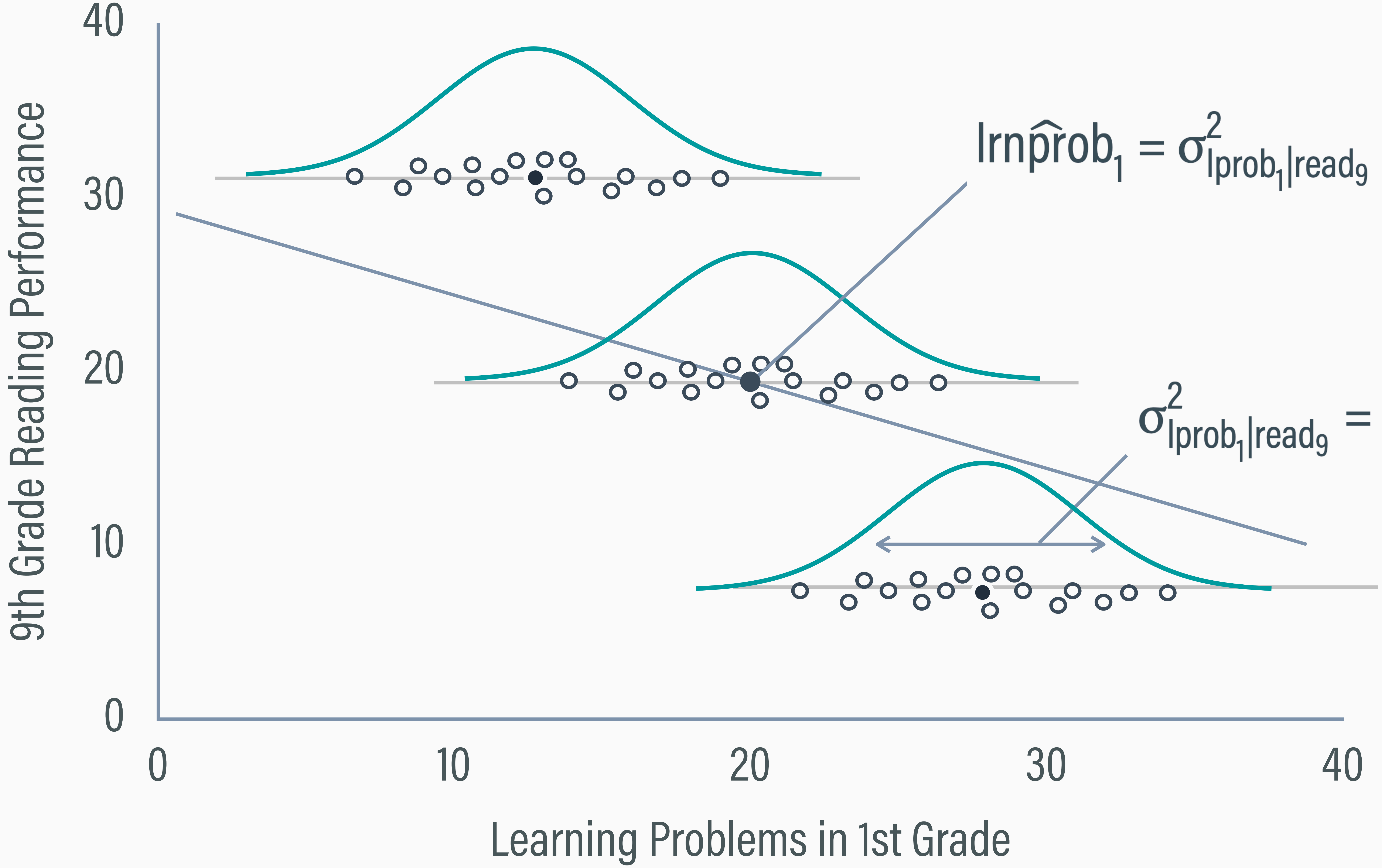
- Both sets of parameter estimates inform the distribution of predictor imputations

DISTRIBUTIONS OF IMPUTATIONS



PREDICTED VALUES AND VARIATION

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

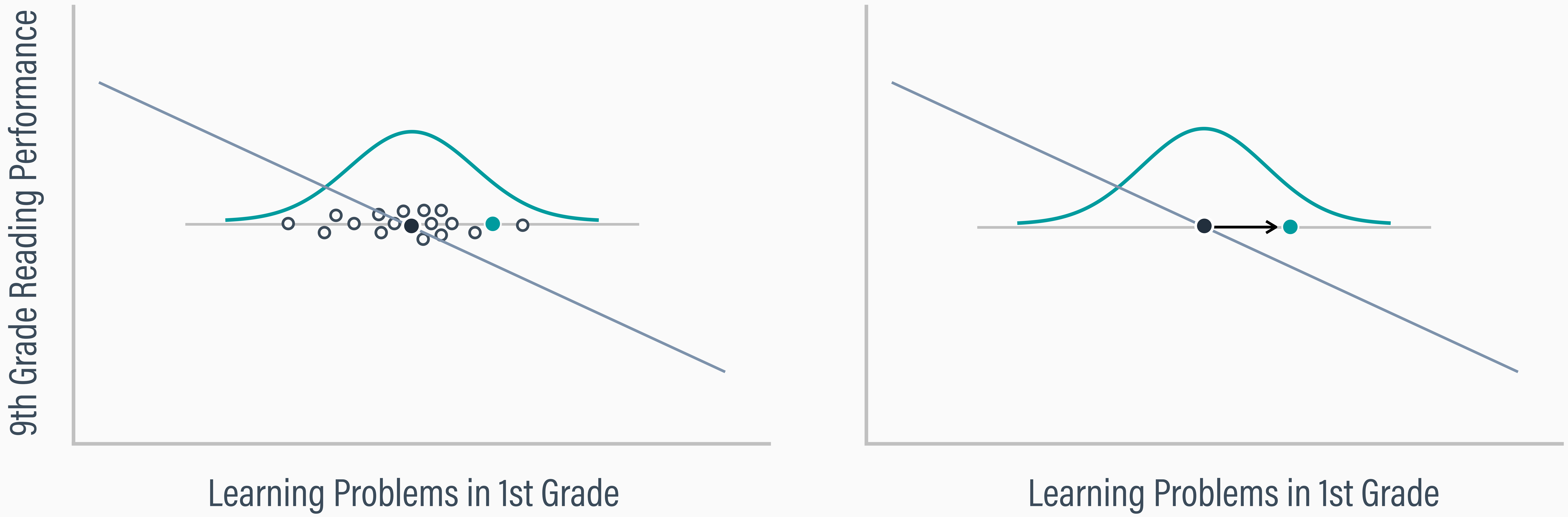


$$\mathrm{lrn}\widehat{\mathrm{p}}\mathrm{rob}_1 = \sigma^2_{\mathrm{lprob}_1|\mathrm{read}_9} \times \left(\frac{\mu}{\sigma^2_e} + \frac{\beta_1\left(\mathrm{read}_9 - \beta_0\right)}{\sigma^2_\varepsilon}\right)$$

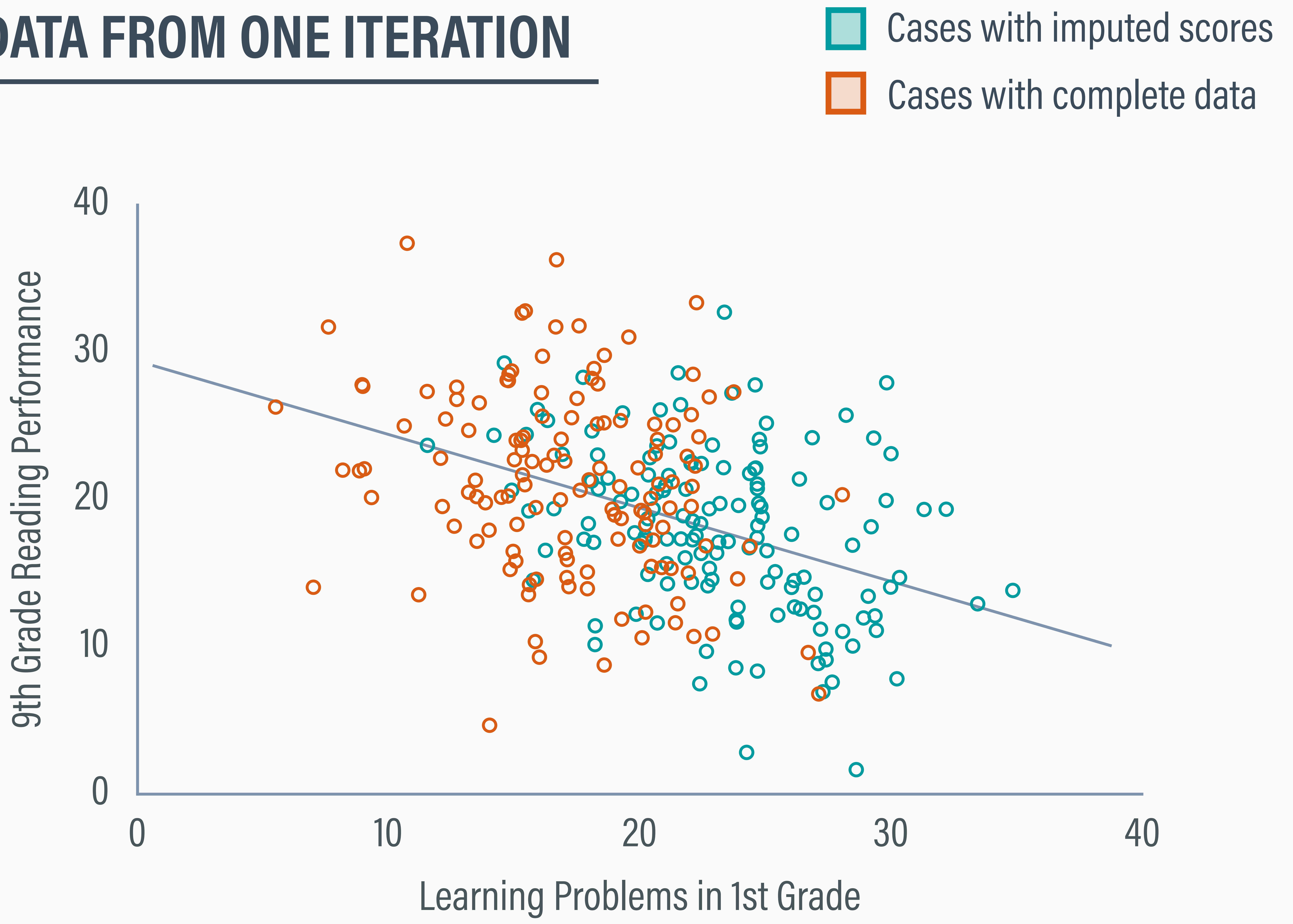
$$\sigma^2_{\mathrm{lprob}_1|\mathrm{read}_9} = \left(\frac{1}{\sigma^2_e} + \frac{\beta_1^2}{\sigma^2_\varepsilon}\right)^{-1}$$

IMPUTATION EXAMPLE

Imputation = predicted value + random normal noise



FILLED-IN DATA FROM ONE ITERATION



OUTLINE

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. MCMC Estimation and Missing Data Imputation
4. Fitting Regression Models in Blimp
5. Incomplete Categorical Variables
6. Interaction Effects
7. Reporting Results From a Missing Data Analysis

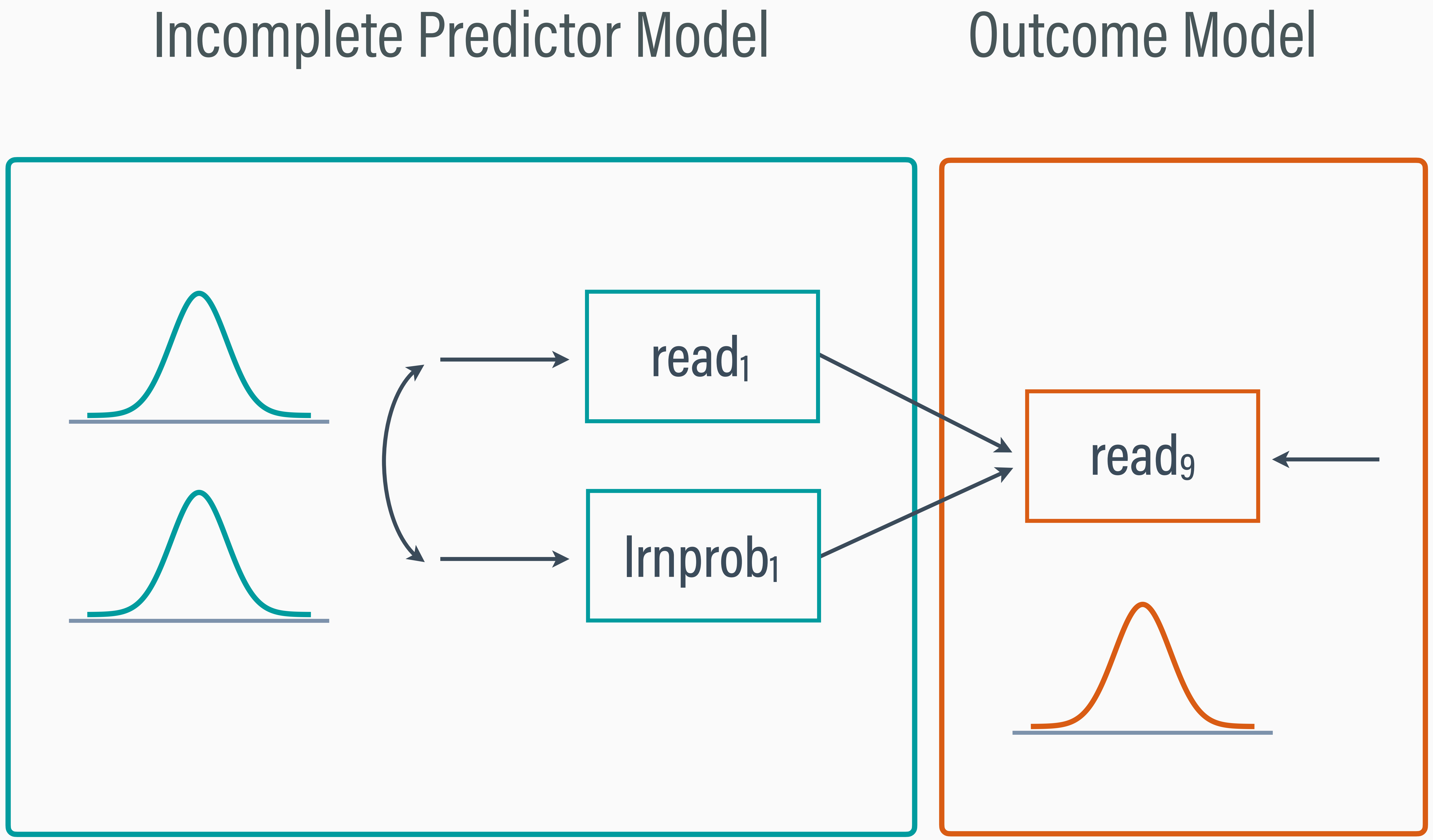
ANALYSIS 1 REGRESSION MODEL

- Academic variables from 1st grade predicting 9th grade reading achievement in middle school

$$read_9 = \beta_0 + \beta_1(read_1) + \beta_2(lrnprob_1) + \varepsilon$$

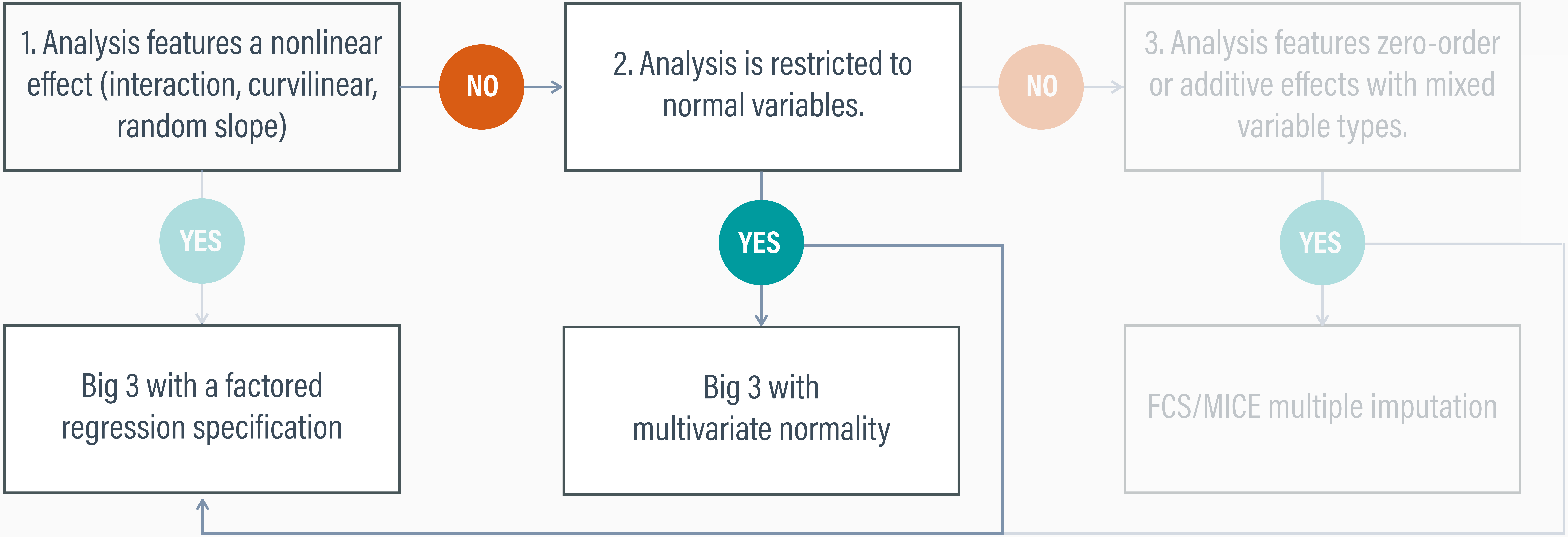
Variable	Definition	Missing %	Scale
atrisk	Emotional/behavioral risk code	2.2	0 = Low, 1 = Medium/high
lrnprob1	1st grade learning problems	2.2	Numeric (31 to 88)
read1	1st grade broad reading composite	6.5	Numeric (39 to 153)
read9	9th grade broad reading composite	17.4	Numeric (41 to 123)

TWO-PART FACTORED SPECIFICATION



MISSING DATA DECISION TREE

This is an analysis where we have multiple equivalent options for missing data handling!



BLIMP STUDIO SCRIPT 1

```
DATA: reading.dat;
VARIABLES: id male hispanic riskgrp atrisk behsymp1 lrnprob1 read1 read2 read3 read9
   read9grp stanread7 math1 math2 math3 math9 math9grp stanmath7;
MISSING: 999;
MODEL:  read9 ~ read1 lrnprob1;
BURN: 10000;
ITER: 10000;
SEED: 90291;
```

DATA AND VARIABLES

```
DATA: reading.dat;                    # data in same directory as the script
VARIABLES: id male hispanic riskgrp atrisk behsymp1 lrnprob1 read1 read2 read3 read9
  read9grp stanread7 math1 math2 math3 math9 math9grp stanmath7;  # name data columns
MISSING: 999;                         # missing value code
MODEL:  read9 ~ read1 lrnprob1;
BURN: 10000;
ITER: 10000;
SEED: 90291;
```

MODEL DETAILS

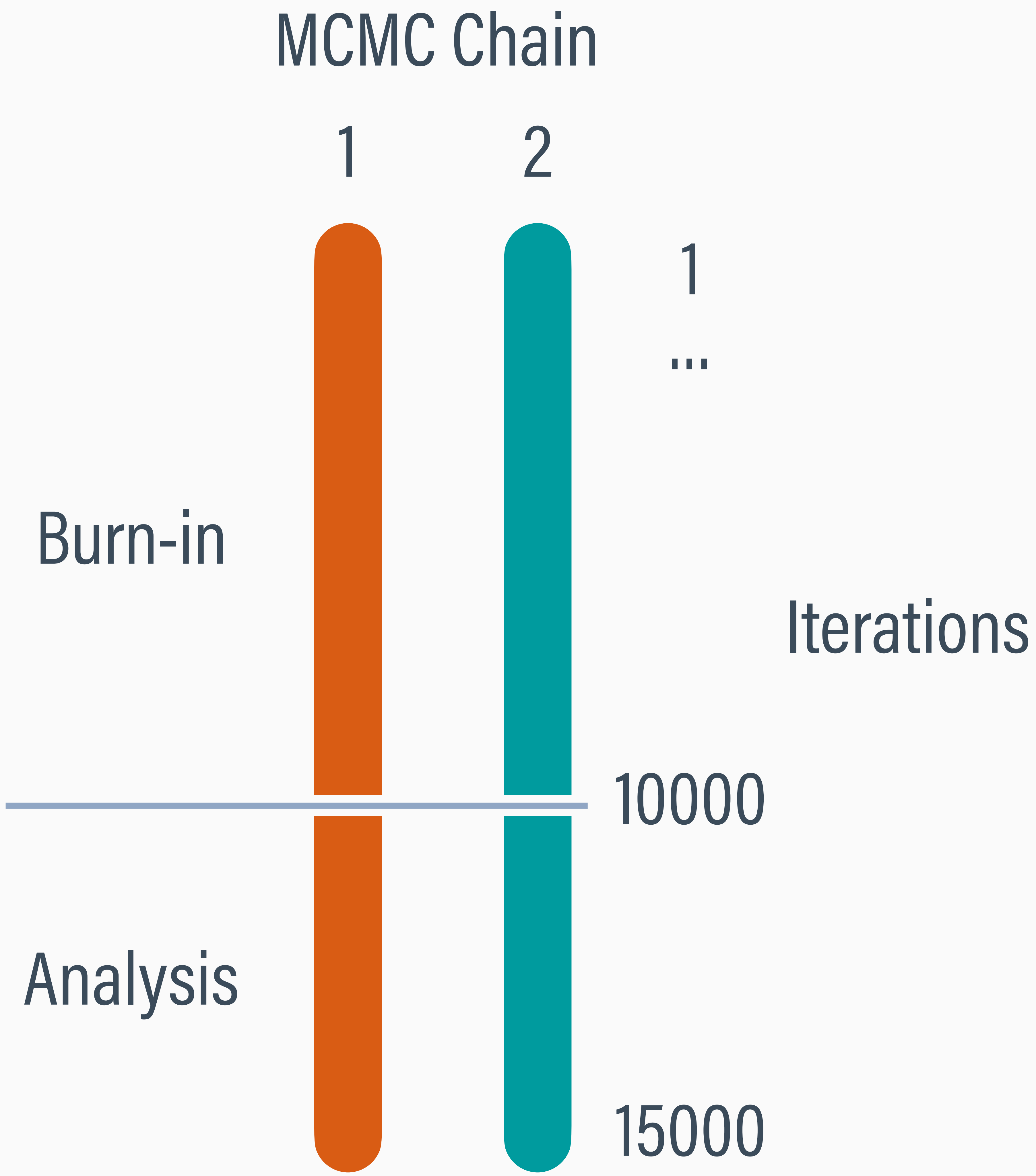
```
DATA: reading.dat;
VARIABLES: id male hispanic riskgrp atrisk behsymp1 lrnprob1 read1 read2 read3 read9
  read9grp stanread7 math1 math2 math3 math9 math9grp stanmath7;
MISSING: 999;
MODEL:  read9 ~ read1 lrnprob1;          # regression model
BURN: 10000;
ITER: 10000;
SEED: 90291;
```

COMPUTATIONAL DETAILS

```
DATA: reading.dat;
VARIABLES: id male hispanic riskgrp atrisk behsymp1 lrnprob1 read1 read2 read3 read9
  read9grp stanread7 math1 math2 math3 math9 math9grp stanmath7;
MISSING: 999;
MODEL:  read9 ~ read1 lrnprob1;
BURN: 10000;                     # number of warm-up iterations
ITER: 10000;                     # number of parameter values for the analysis
SEED: 90291;                     # integer seed for Monte Carlo simulation
```

COMPUTATIONAL DETAILS

BURN: 10000; # warm-up iterations (per chain)
ITER: 10000; # estimates to summarize (both chains)
SEED: 90291; # random number seed



RBLIMP SCRIPT (MODEL 1)

```
# fit model
model1 <- rblimp(
   data = reading,                          # R data frame
   model = 'read9 ~ read1 lrnprob1',        # regression model
   seed = 90291,                            # integer seed for Monte Carlo simulation
   burn = 10000,                            # number of warm-up iterations
   iter = 10000)                            # number of parameter values for the analysis

# summarize results
output(model1)                              # print output
posterior_plot(model1, ‘read9')             # plot parameter distributions
```

MISSING DATA INFORMATION

- Per-variable missingness rates appear in the rows, and missing data patterns appear in the columns (M = missing)

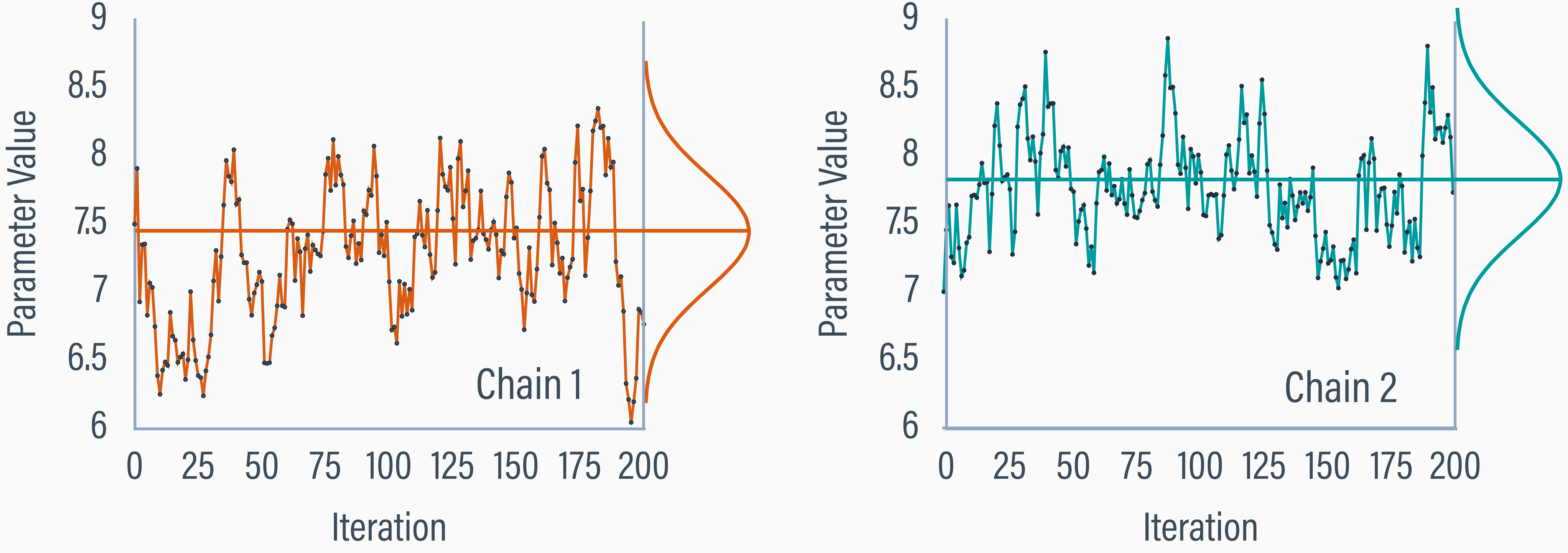
```
DATA INFORMATION:

  Sample Size:                 138
  Missing Data Info:
                        miss %       1     2     3     4     5
                                  --------------------------------
                  read9 = 17.4       -     M     -     -     M
               lrnprob1 =  2.2       -     -     -     M     -
                  read1 =  6.5       -     -     M     -     M
                                  --------------------------------
                                  %  74.6  16.7   5.8   2.2   0.7
```

MCMC CONVERGENCE

- MCMC parameter estimates continually vary across iterations
- MCMC converges when parameter estimates oscillate around a stable mean, and variation doesn’t change with more iterations
- The potential scale reduction factor (PSRF) compares the similarity of parameters generated from two MCMC processes

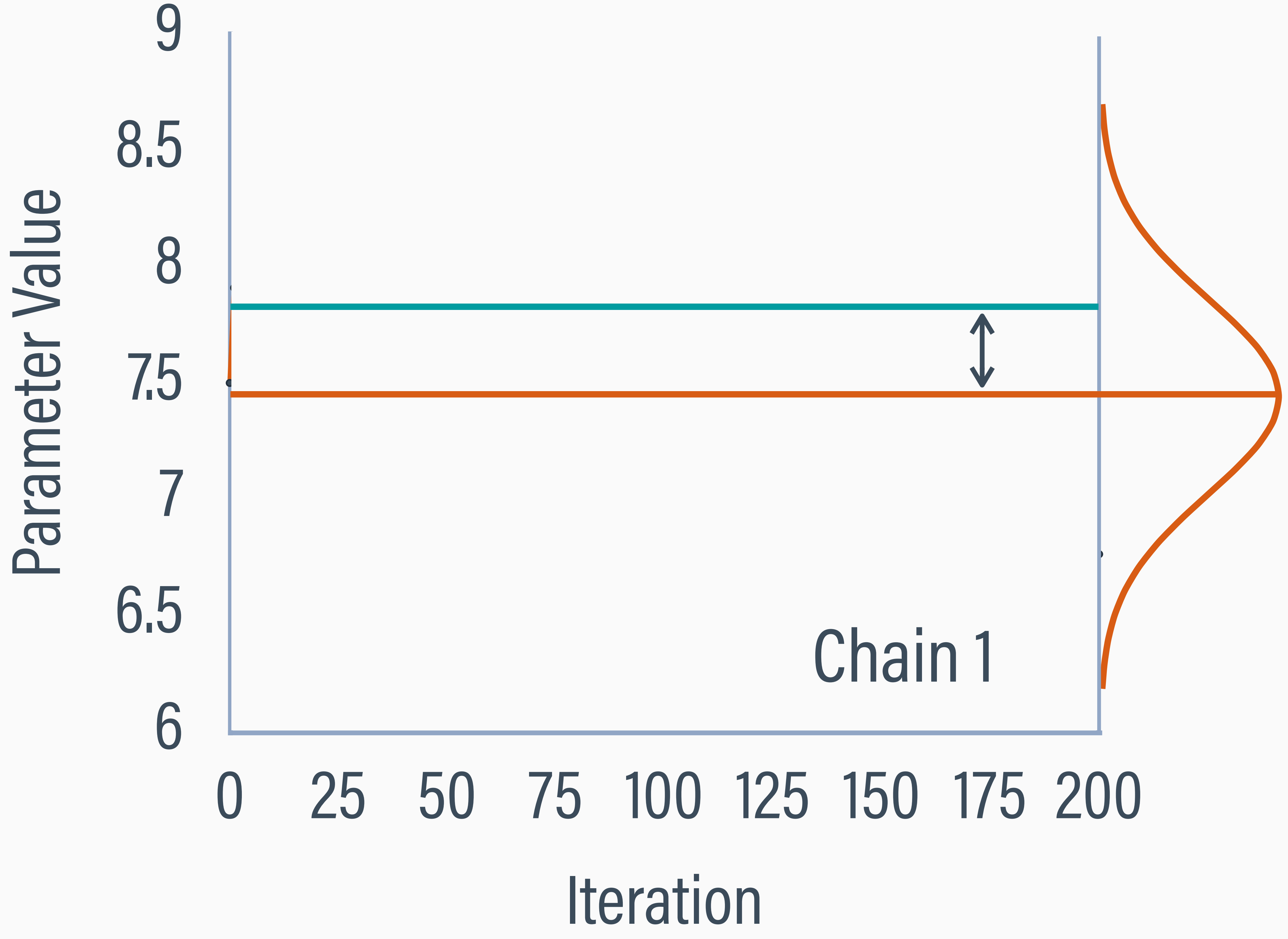
POTENTIAL SCALE REDUCTION FACTOR

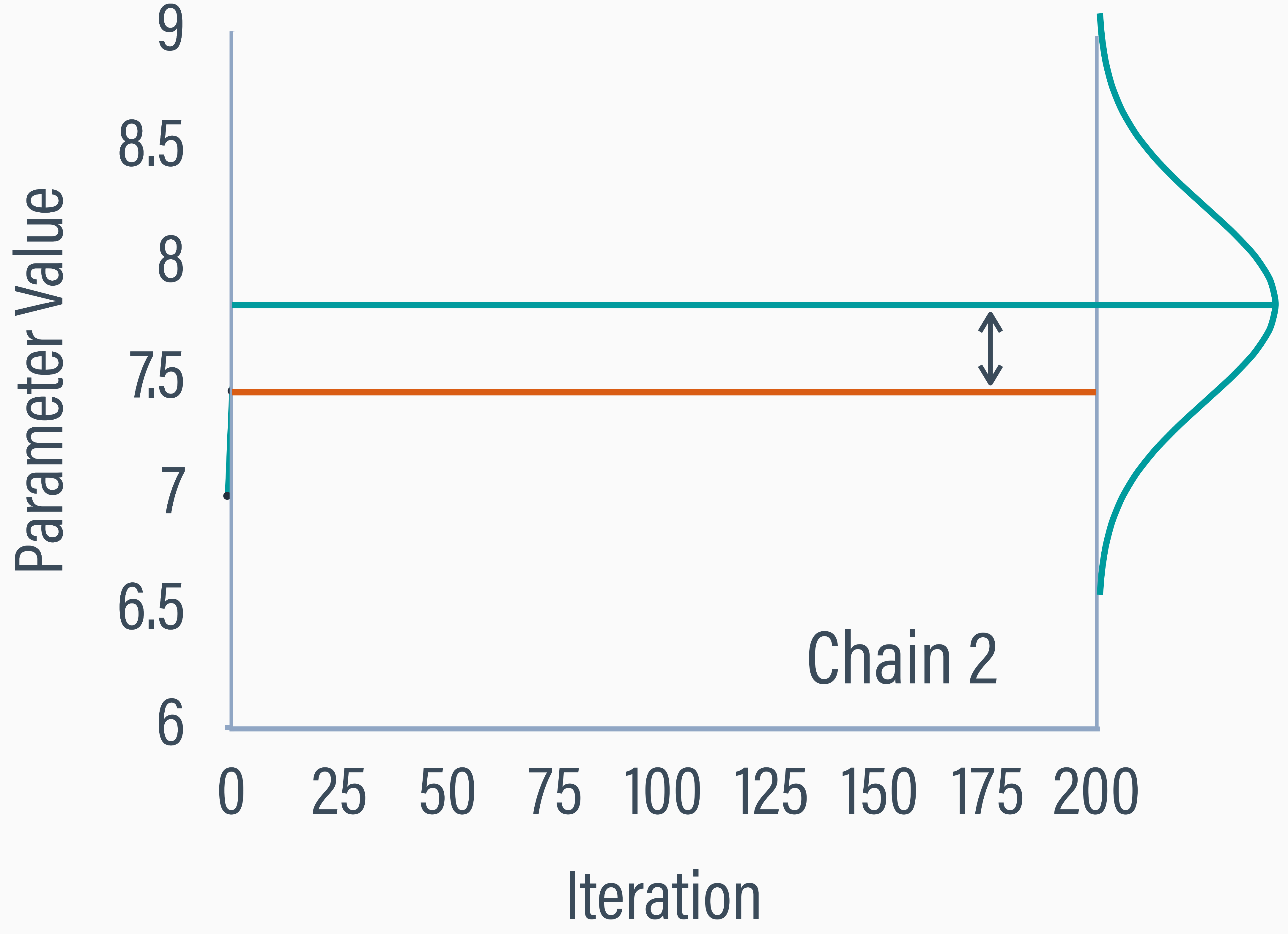
$$\text{PSRF} = \sqrt{\frac{\text{mean difference between chains + within-chain variation}}{\text{within-chain variation}}}$$



BETWEEN-CHAIN MEAN DIFFERENCE

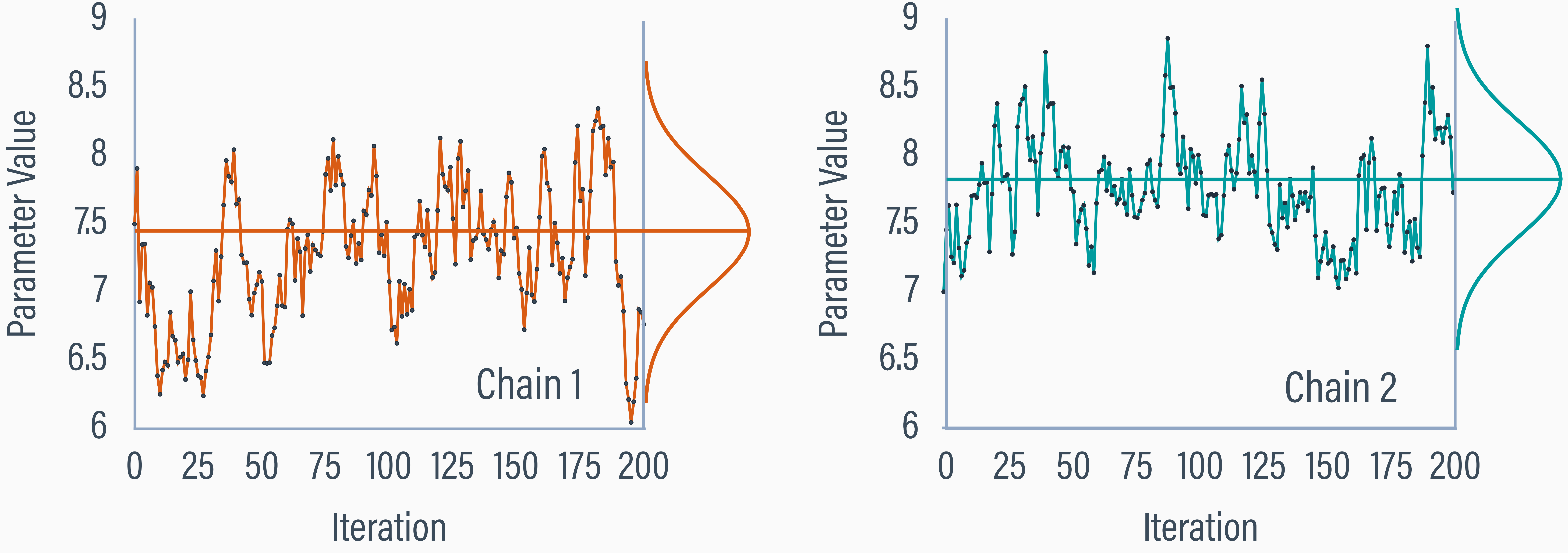
$$\text{PSRF} = \sqrt{\frac{\text{mean difference between chains} + \text{within-chain variation}}{\text{within-chain variation}}}$$





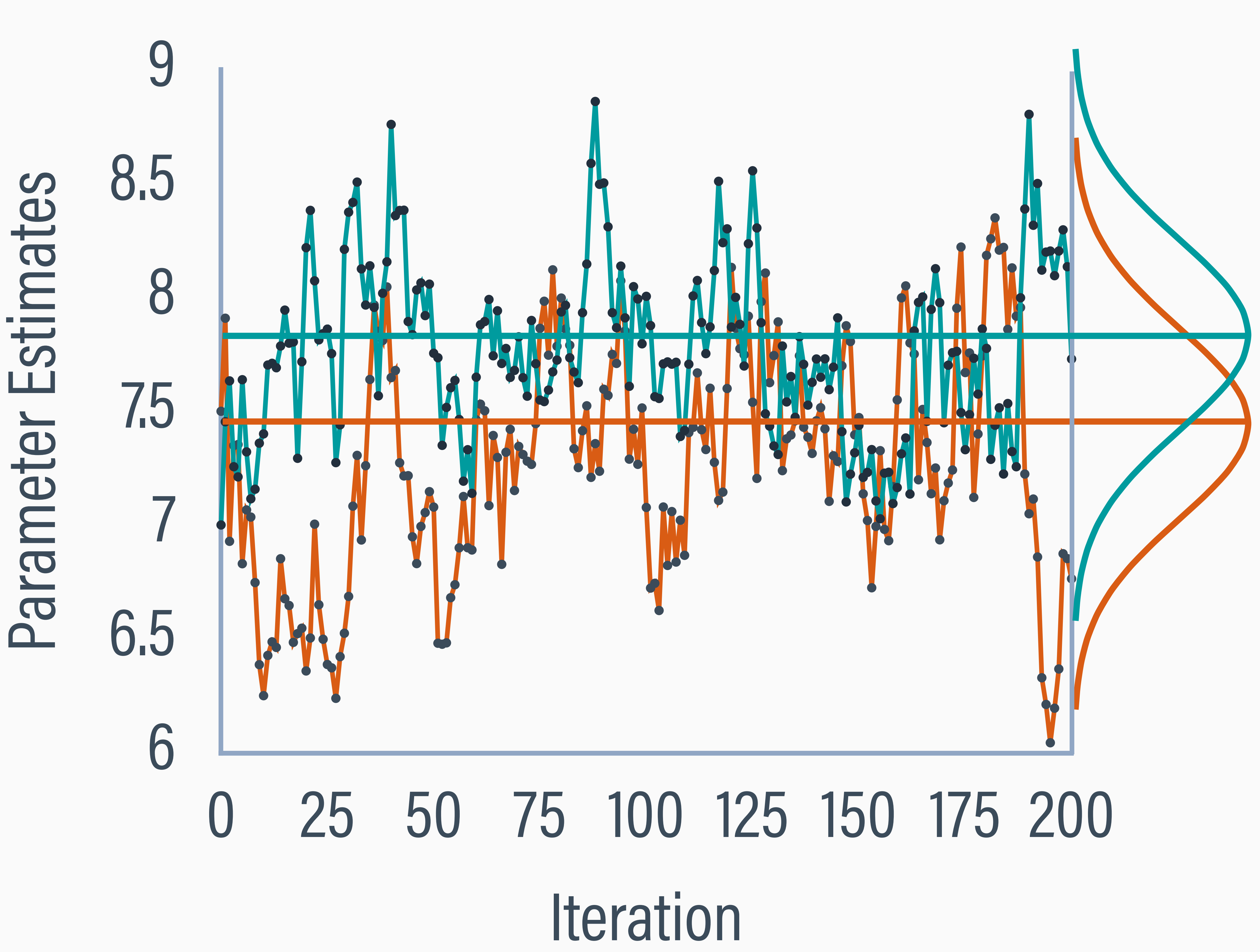
WITHIN-CHAIN VARIATION

$$\text{PSRF} = \sqrt{\frac{\text{mean difference between chains} + \text{within-chain variation}}{\text{within-chain variation}}}$$

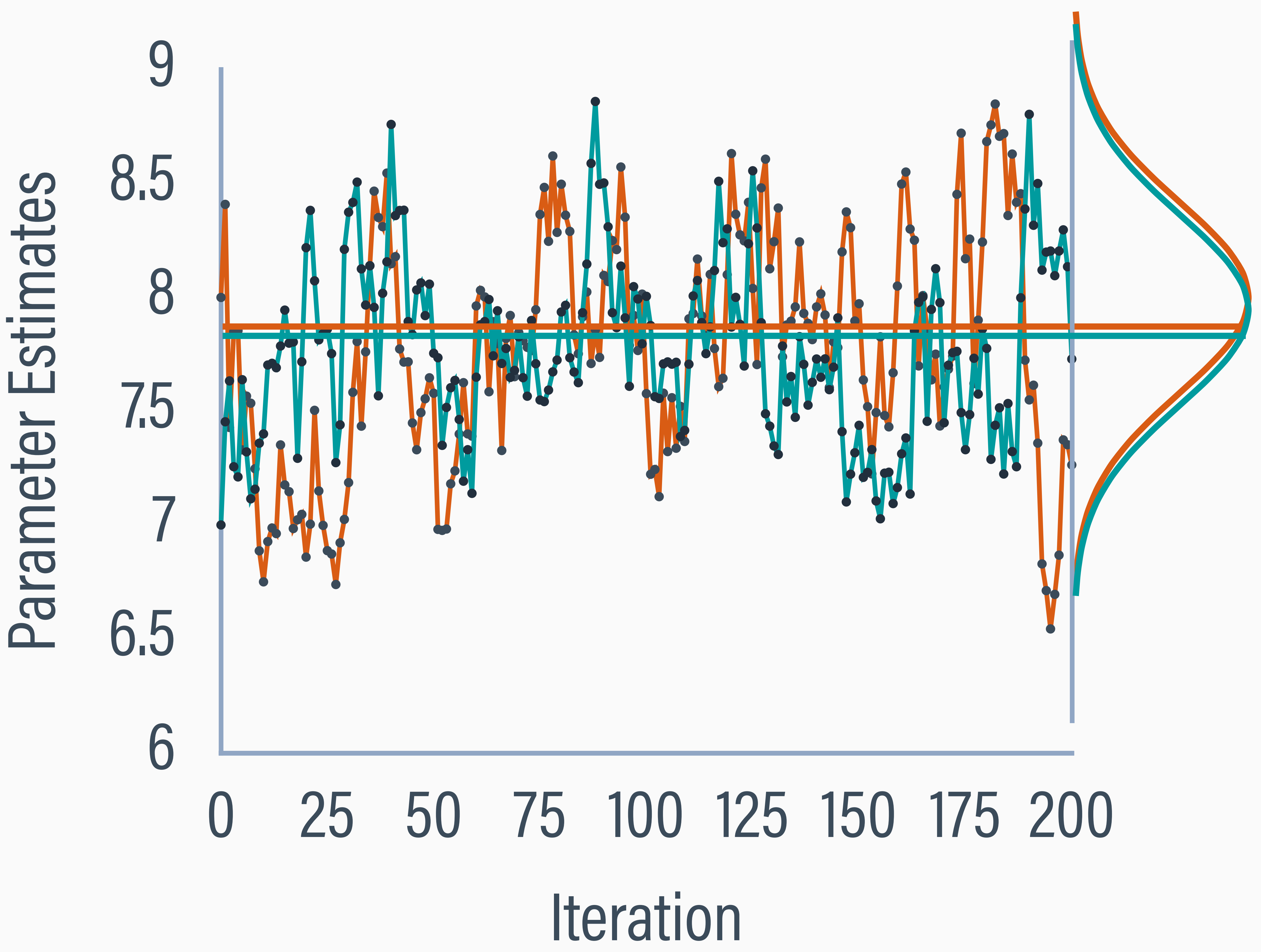


CONVERGENCE

MCMC has not converged because between-chain mean difference is large (PSR > 1.05)



MCMC has converged because between-chain mean difference is very small (PSR < 1.05)



PSR DIAGNOSTIC OUTPUT

The number of burn-in iterations is sufficient because the highest PSRF across all parameters is < 1.05 at the end of the burn-in period

```
BURN-IN POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

  NOTE: Split chain PSR is being used. This splits each chain's
        iterations to create twice as many chains.

  Comparing iterations across 2 chains     Highest PSR   Parameter #
                          251 to 500             1.029             6
                          501 to 1000            1.012             1
                          751 to 1500            1.006             8
                         1001 to 2000            1.003             1
                          ...     ...              ...           ...
                         4001 to 8000            1.001             6
                         4251 to 8500            1.001             4
                         4501 to 9000            1.001            14
                         4751 to 9500            1.000            14
                         5001 to 10000        »  1.001             6
```

SUMMARY TABLE: VARIANCES AND COVARIANCES

OUTCOME MODEL ESTIMATES:

Summaries based on 10000 iterations using 2 chains.
NOTE: Estimate column based on posterior median.

Outcome Variable: **read9**

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	94.449	13.550	72.536	125.090	---	---	5358.708
Coefficients:							
Intercept	65.176	6.748	51.876	78.518	93.364	0.000	3601.650
read1	0.505	0.048	0.409	0.597	112.098	0.000	5608.777
lrnprob1	-0.404	0.098	-0.599	-0.209	16.811	0.000	2401.917
Standardized Coefficients:							
read1	0.680	0.045	0.577	0.753	226.950	0.000	4685.157
lrnprob1	-0.297	0.069	-0.426	-0.156	18.342	0.000	2399.433
Proportion Variance Explained							
by Coefficients	0.562	0.055	0.438	0.654	---	---	5115.914
by Residual Variation	0.438	0.055	0.346	0.562	---	---	5115.914

SUMMARY TABLE: REGRESSION COEFFICIENTS

OUTCOME MODEL ESTIMATES:

Summaries based on 10000 iterations using 2 chains.
NOTE: Estimate column based on posterior median.

Outcome Variable: **read9**

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	94.449	13.550	72.536	125.090	---	---	5358.708
Coefficients:							
Intercept	65.176	6.748	51.876	78.518	93.364	0.000	3601.650
read1	0.505	0.048	0.409	0.597	112.098	0.000	5608.777
lrnprob1	-0.404	0.098	-0.599	-0.209	16.811	0.000	2401.917
Standardized Coefficients:							
read1	0.680	0.045	0.577	0.753	226.950	0.000	4685.157
lrnprob1	-0.297	0.069	-0.426	-0.156	18.342	0.000	2399.433
Proportion Variance Explained							
by Coefficients	0.562	0.055	0.438	0.654	---	---	5115.914
by Residual Variation	0.438	0.055	0.346	0.562	---	---	5115.914

SUMMARY TABLE: STANDARDIZED COEFFICIENTS

OUTCOME MODEL ESTIMATES:

Summaries based on 10000 iterations using 2 chains.
NOTE: Estimate column based on posterior median.

Outcome Variable: read9

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	94.449	13.550	72.536	125.090	---	---	5358.708
Coefficients:							
Intercept	65.176	6.748	51.876	78.518	93.364	0.000	3601.650
read1	0.505	0.048	0.409	0.597	112.098	0.000	5608.777
lrnprob1	-0.404	0.098	-0.599	-0.209	16.811	0.000	2401.917
Standardized Coefficients:							
read1	0.680	0.045	0.577	0.753	226.950	0.000	4685.157
lrnprob1	-0.297	0.069	-0.426	-0.156	18.342	0.000	2399.433
Proportion Variance Explained							
by Coefficients	0.562	0.055	0.438	0.654	---	---	5115.914
by Residual Variation	0.438	0.055	0.346	0.562	---	---	5115.914

SUMMARY TABLE: R^2 EFFECT SIZES

OUTCOME MODEL ESTIMATES:

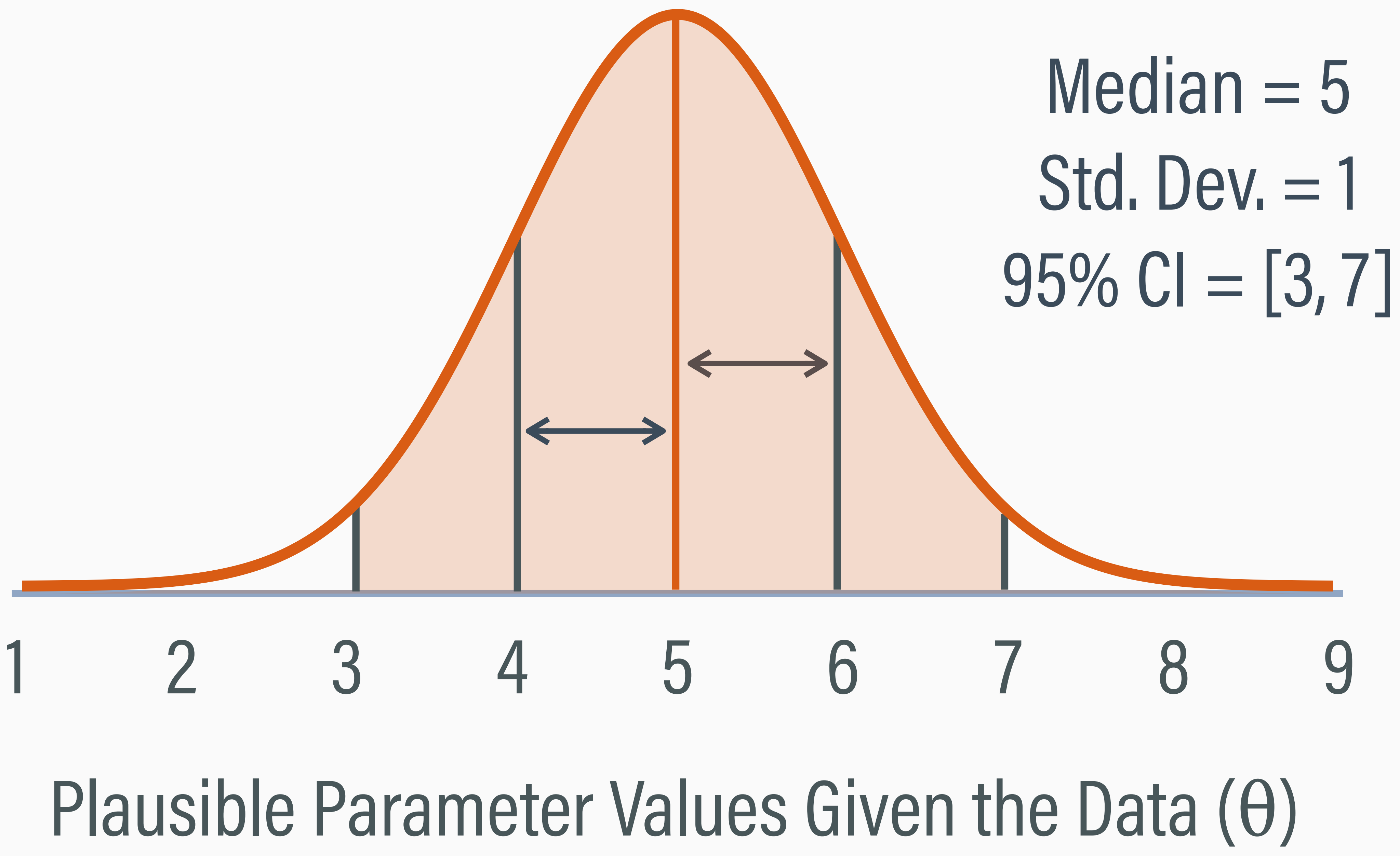
Summaries based on 10000 iterations using 2 chains.
NOTE: Estimate column based on posterior median.

Outcome Variable: **read9**

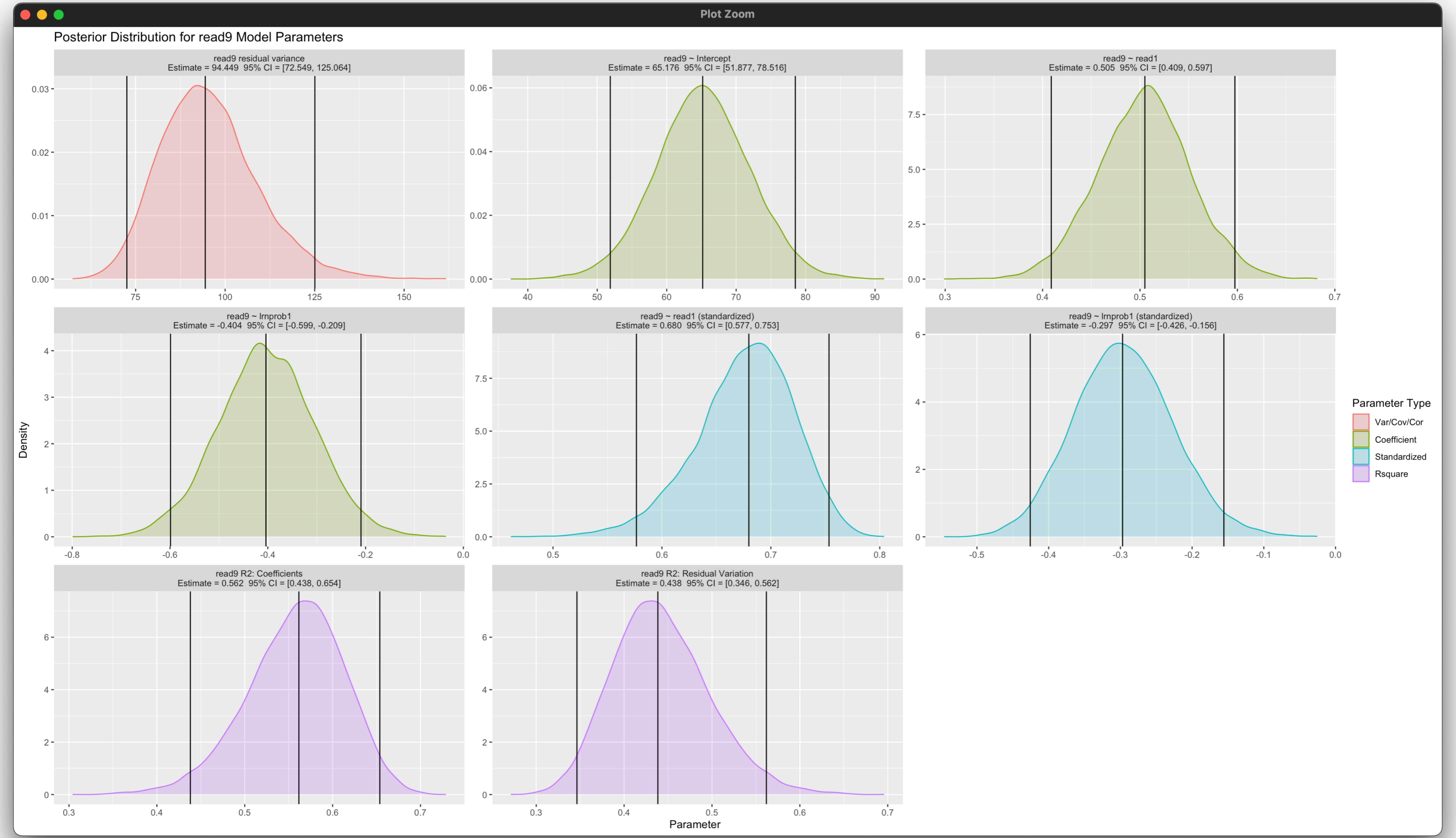
Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	94.449	13.550	72.536	125.090	---	---	5358.708
Coefficients:							
Intercept	65.176	6.748	51.876	78.518	93.364	0.000	3601.650
read1	0.505	0.048	0.409	0.597	112.098	0.000	5608.777
lrnprob1	-0.404	0.098	-0.599	-0.209	16.811	0.000	2401.917
Standardized Coefficients:							
read1	0.680	0.045	0.577	0.753	226.950	0.000	4685.157
lrnprob1	-0.297	0.069	-0.426	-0.156	18.342	0.000	2399.433
Proportion Variance Explained							
by Coefficients	0.562	0.055	0.438	0.654	---	---	5115.914
by Residual Variation	0.438	0.055	0.346	0.562	---	---	5115.914

DISTRIBUTION SUMMARIES

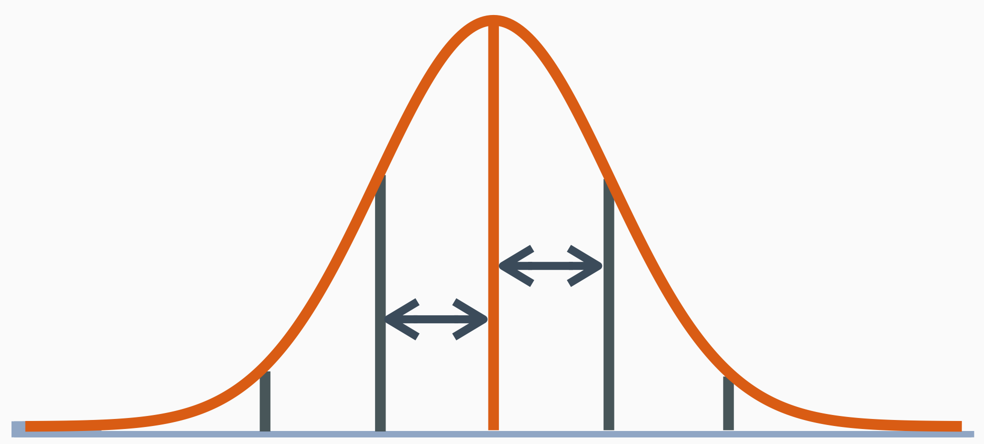
- The median (or mean) quantifies the most likely parameter value
- The standard deviation quantifies spread of the parameter's distribution
- 95% intervals define plausible parameter values that could have produced the data



DISTRIBUTION PLOTS (RBLIMP ONLY)



ESTIMATES AND “BAYESIAN STANDARD ERRORS”



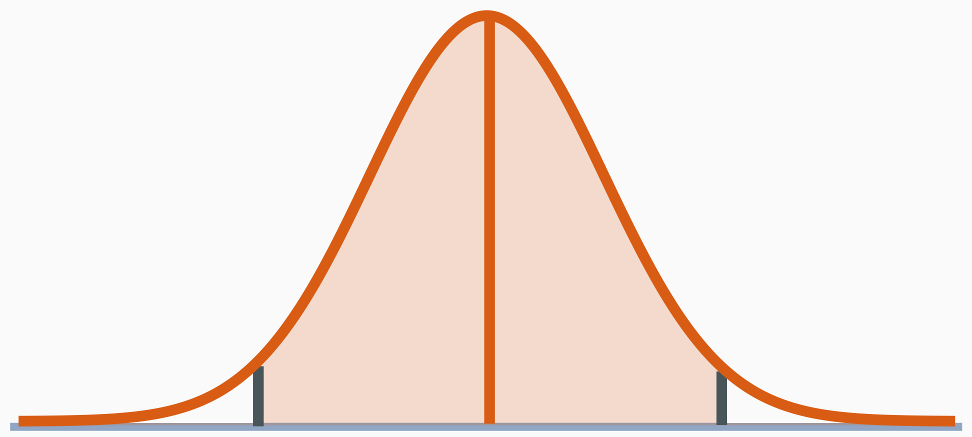
OUTCOME MODEL ESTIMATES:

Summaries based on 10000 iterations using 2 chains.
NOTE: Estimate column based on posterior median.

Outcome Variable: **read9**

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	94.449	13.550	72.536	125.090	---	---	5358.708
Coefficients:							
Intercept	65.176	6.748	51.876	78.518	93.364	0.000	3601.650
read1	0.505	0.048	0.409	0.597	112.098	0.000	5608.777
lrnprob1	-0.404	0.098	-0.599	-0.209	16.811	0.000	2401.917
Standardized Coefficients:							
read1	0.680	0.045	0.577	0.753	226.950	0.000	4685.157
lrnprob1	-0.297	0.069	-0.426	-0.156	18.342	0.000	2399.433
Proportion Variance Explained							
by Coefficients	0.562	0.055	0.438	0.654	---	---	5115.914
by Residual Variation	0.438	0.055	0.346	0.562	---	---	5115.914

95% CREDIBLE INTERVALS



OUTCOME MODEL ESTIMATES:

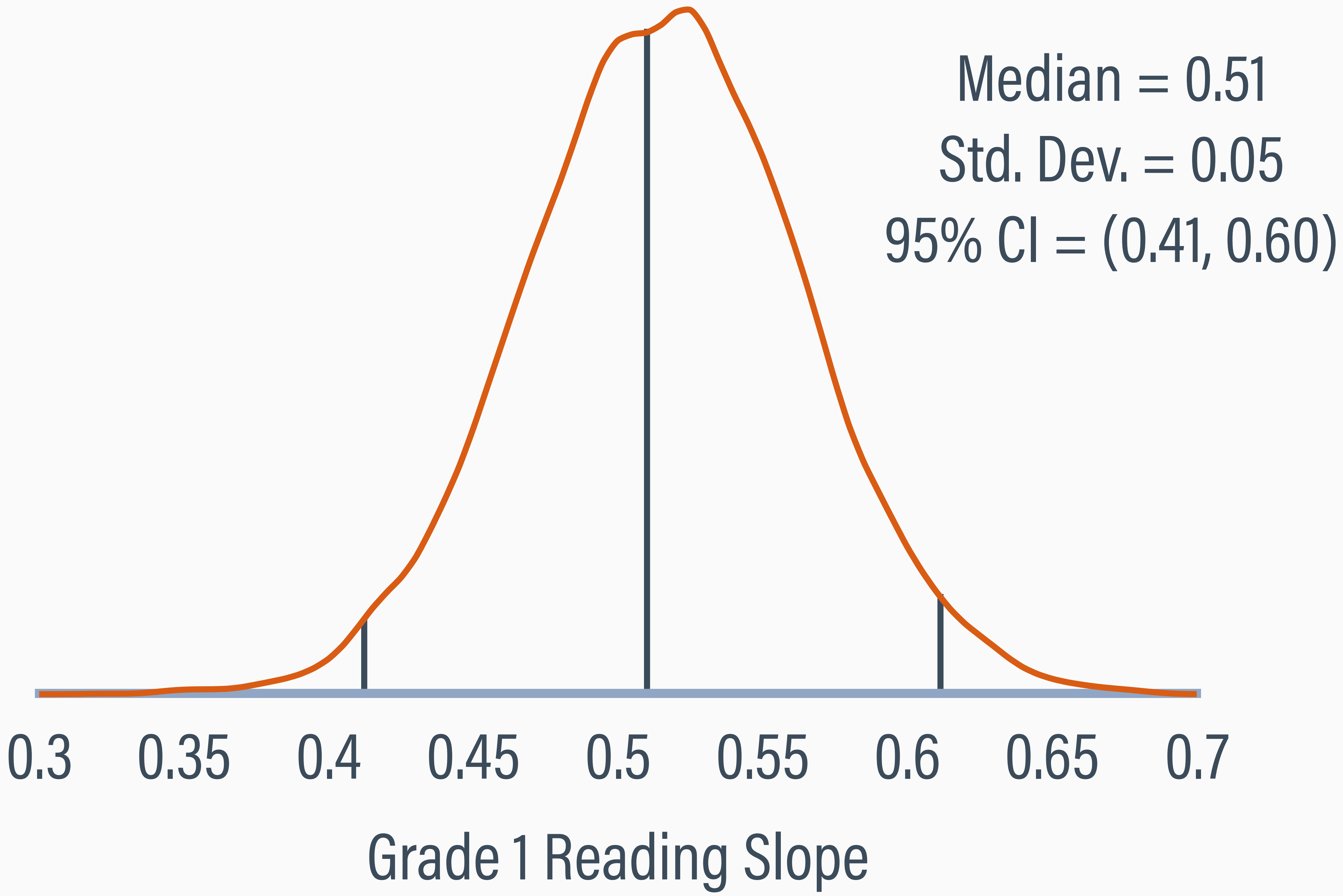
Summaries based on 10000 iterations using 2 chains.
NOTE: Estimate column based on posterior median.

Outcome Variable: **read9**

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	94.449	13.550	72.536	125.090	---	---	5358.708
Coefficients:							
Intercept	65.176	6.748	51.876	78.518	93.364	0.000	3601.650
read1	0.505	0.048	0.409	0.597	112.098	0.000	5608.777
lrnprob1	-0.404	0.098	-0.599	-0.209	16.811	0.000	2401.917
Standardized Coefficients:							
read1	0.680	0.045	0.577	0.753	226.950	0.000	4685.157
lrnprob1	-0.297	0.069	-0.426	-0.156	18.342	0.000	2399.433
Proportion Variance Explained							
by Coefficients	0.562	0.055	0.438	0.654	---	---	5115.914
by Residual Variation	0.438	0.055	0.346	0.562	---	---	5115.914

INTERPRETATIONS

- For two students with same first grade learning problems rating, scoring one point higher on the first grade reading test predicts a 0.51 increase in grade 9 reading
- The parameter's standard deviation is 0.05
- The range from 0.41 to 0.60 captures 95% of the plausible parameter values that could have produced these data

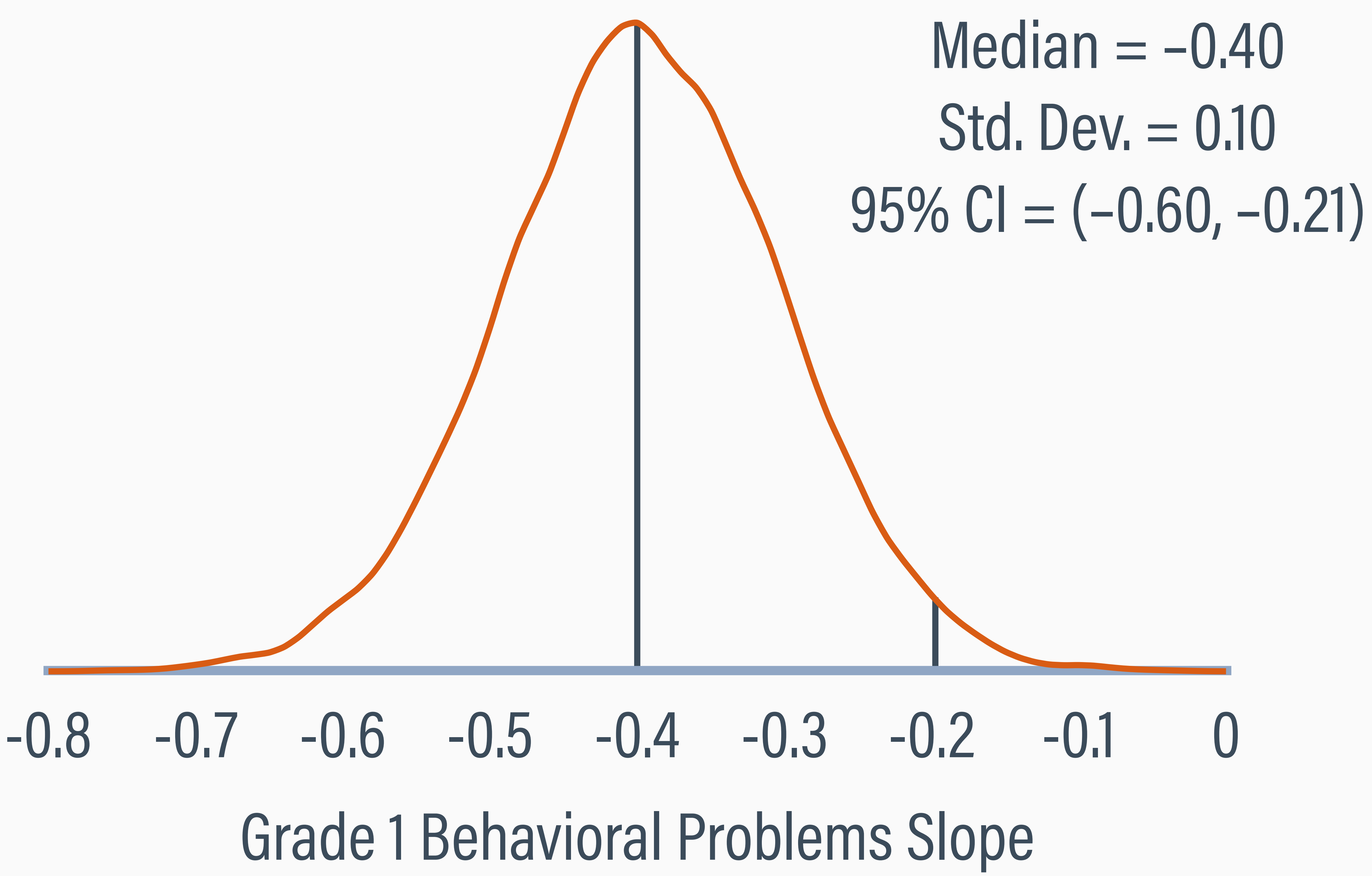


INTERPRETATIONS, CONTINUED

- For two students with same first grade reading, scoring one point higher on the first grade learning problems measure predicts a –0.44 decrease in grade 9 reading

- The parameter's standard deviation is 0.10

- The range from –0.60 to –0.21 captures 95% of the plausible parameter values that could have produced these data

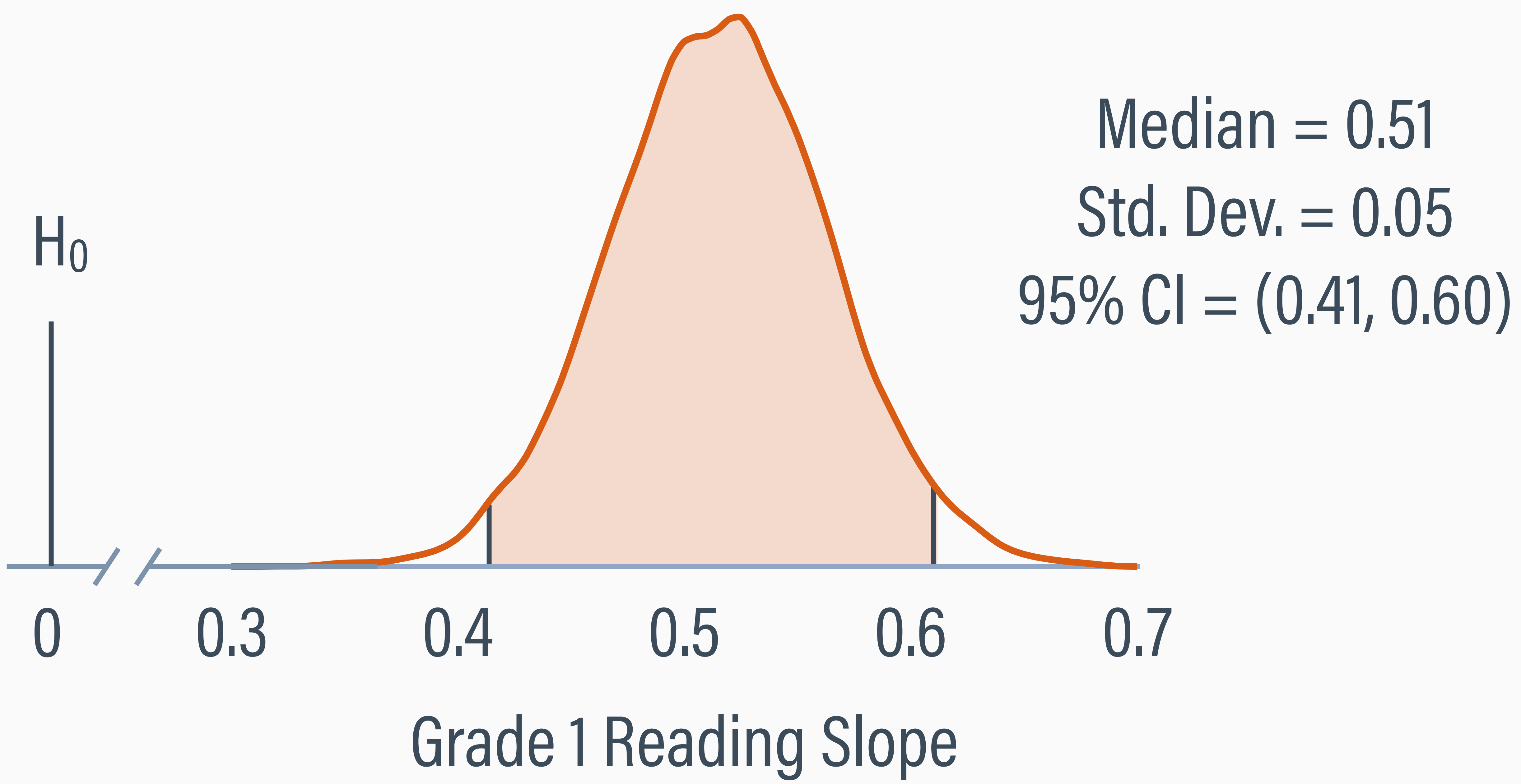


COMPUTATIONAL FREQUENTISM

- Many common missing data problems are challenging or impossible with maximum likelihood
- A computational frequentist perspective views MCMC parameter summaries are replacements for unobtainable maximum likelihood estimates
- Use MCMC-generated quantities for frequentist inference

SIGNIFICANCE TESTING VIA INTERVALS

- If the 95% credible interval does not include zero, so the we refute the null hypothesis ($p < .05$)
- A population slope equal to zero is unlikely to have produced these data



FREQUENTIST WALD TEST

- The Wald chi-square is an alternate test statistic that equals the square of the z-statistic (or t-test)

$$\chi^2_{ML} = \frac{\left(\hat{\theta} - \theta_0\right)^2}{SE^2} = \frac{(\text{estimate} - \text{null})^2}{(\text{standard error})^2} = z^2$$

- Multivariate versions of the Wald test can evaluate multiple parameters simultaneously

MCMC WALD TEST

- The Bayesian or MCMC Wald test statistic (Asparouhov & Muthén, 2021) replaces the point estimate and standard error with the posterior mean and standard deviation

$$\chi^2_{MCMC} = \frac{(\theta - \theta_0)^2}{SD^2} = \frac{(\text{posterior mean} - \text{null})^2}{(\text{posterior standard deviation})^2} = z^2$$

- MCMC-generated test statistic and p-value for frequentist inference (computational frequentism)

WALD CHI-SQUARE TESTS

OUTCOME MODEL ESTIMATES:

Summaries based on 10000 iterations using 2 chains.
NOTE: Estimate column based on posterior median.

Outcome Variable: read9

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	94.449	13.550	72.536	125.090	---	---	5358.708
Coefficients:							
Intercept	65.176	6.748	51.876	78.518	93.364	0.000	3601.650
read1	0.505	0.048	0.409	0.597	112.098	0.000	5608.777
lrnprob1	-0.404	0.098	-0.599	-0.209	16.811	0.000	2401.917
Standardized Coefficients:							
read1	0.680	0.045	0.577	0.753	226.950	0.000	4685.157
lrnprob1	-0.297	0.069	-0.426	-0.156	18.342	0.000	2399.433
Proportion Variance Explained							
by Coefficients	0.562	0.055	0.438	0.654	---	---	5115.914
by Residual Variation	0.438	0.055	0.346	0.562	---	---	5115.914

ESTIMATOR COMPARISON

The two estimators are effectively numerically equivalent!!!

	MCMC						FIML					
Parameter	Median	SD	2.5%	97.5%	Chi-Sq.	p	Est.	SE	2.5%	97.5%	z	p
Intercept	65.18	6.75	51.88	78.52	93.36	< .001	65.13	6.50	52.39	77.87	10.02	< .001
1st Grade Reading	0.51	0.05	0.41	0.60	112.10	< .001	0.51	0.05	0.42	0.59	11.07	< .001
Learning Problems	−0.40	0.10	−0.60	−0.21	16.81	< .001	−0.40	0.10	−0.59	−0.22	−4.24	< .001
Residual variance	94.45	13.55	72.54	125.09	—	—	89.35	12.33	65.19	113.51	—	—
R^2	.56	.06	.44	.65	—	—	0.57	0.06	—	—	—	—

QUALITY CONTROL CHECK: EFFECTIVE SAMPLE SIZE

- The effective sample size (N_Eff) diagnostic quantifies the number of independent MCMC estimates contributing to the parameter summaries after removing autocorrelation

- An acceptable value (N_Eff > 100) implies that the number of iterations after the burn-in period is sufficient, whereas low values suggest to increase the number of iterations

- Low values often indicate that the data lack support for certain model parameters (e.g., due to overfitting)

DIAGNOSTIC OUTPUT

All N_Eff values > 100, the number of iterations for the summary is sufficient!

OUTCOME MODEL ESTIMATES:

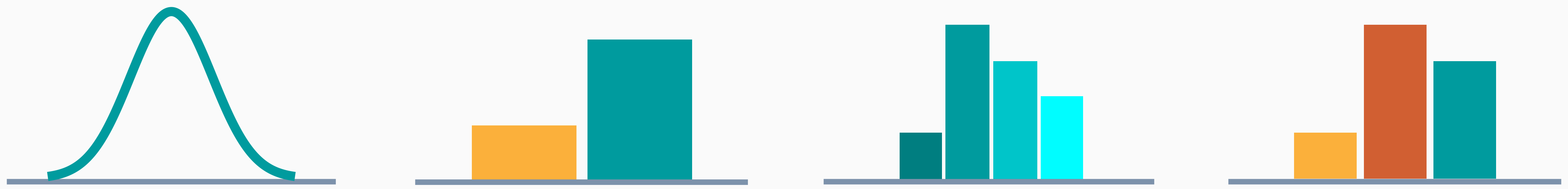
```
  Summaries based on 10000 iterations using 2 chains.
  NOTE: Estimate column based on posterior median.

Outcome Variable:  read9
```

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	94.449	13.550	72.536	125.090	---	---	5358.708
Coefficients:							
Intercept	65.176	6.748	51.876	78.518	93.364	0.000	3601.650
read1	0.505	0.048	0.409	0.597	112.098	0.000	5608.777
lrnprob1	-0.404	0.098	-0.599	-0.209	16.811	0.000	2401.917
Standardized Coefficients:							
read1	0.680	0.045	0.577	0.753	226.950	0.000	4685.157
lrnprob1	-0.297	0.069	-0.426	-0.156	18.342	0.000	2399.433
Proportion Variance Explained							
by Coefficients	0.562	0.055	0.438	0.654	---	---	5115.914
by Residual Variation	0.438	0.055	0.346	0.562	---	---	5115.914

OUTLINE

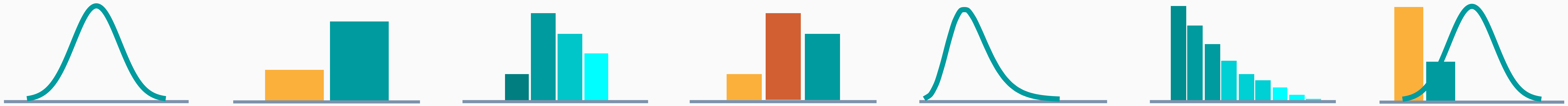
BLIMP VARIABLE TYPES

Exogenous Predictors



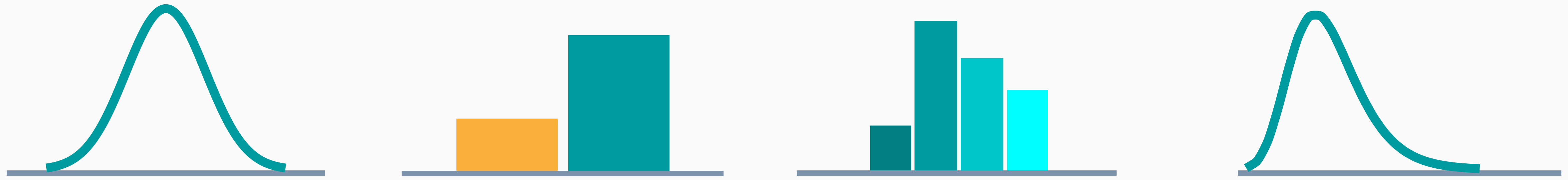
Normal (Manifest) | Binary | Ordinal | Nominal

Univariate Outcomes



Normal (Manifest or Latent) | Binary | Ordinal | Nominal | Skewed (Manifest or Latent) | Count | Two-Part (Floor Effects)

Multivariate Outcomes



Normal (Manifest or Latent) | Binary | Ordinal | Skewed (Manifest or Latent)

LATENT RESPONSE FORMULATION

READING DATA

- In first grade, students are classified according to their risk of developing emotional or behavioral problems
- 65% are classified as being at medium or high risk

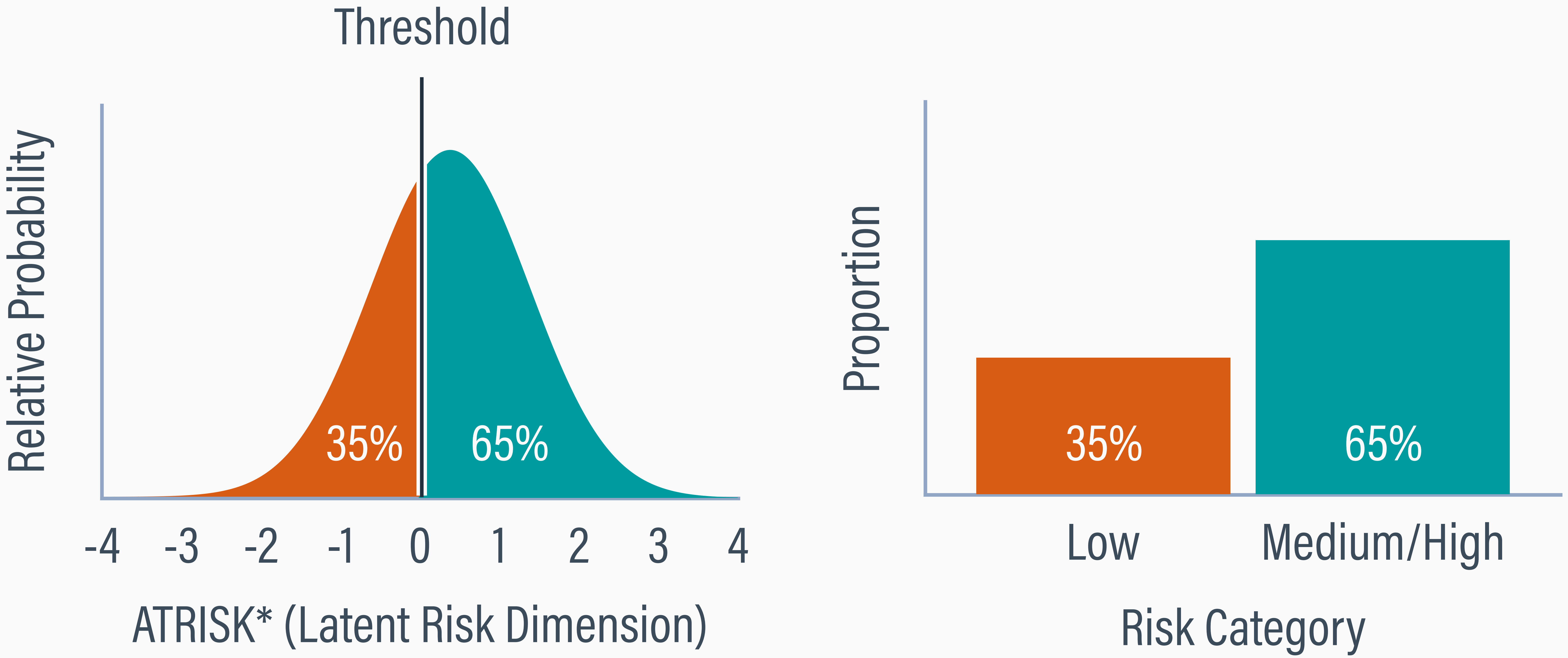
Variable	Definition	Missing %	Scale
atrisk	Emotional/behavioral risk code	2.2	0 = Low, 1 = Medium/high
lrnprob1	1st grade learning problems	2.2	Numeric (31 to 88)
read1	1st grade broad reading composite	6.5	Numeric (39 to 153)
read9	9th grade broad reading composite	17.4	Numeric (41 to 123)

INCOMPLETE BINARY VARIABLES

- Probit regression envisions binary and ordinal variables arising from an underlying normal latent response variable
- Applied to the at risk indicator, the latent variable represents an unobserved, continuous propensity for emotional disorders
- A threshold carves the latent distribution into segments

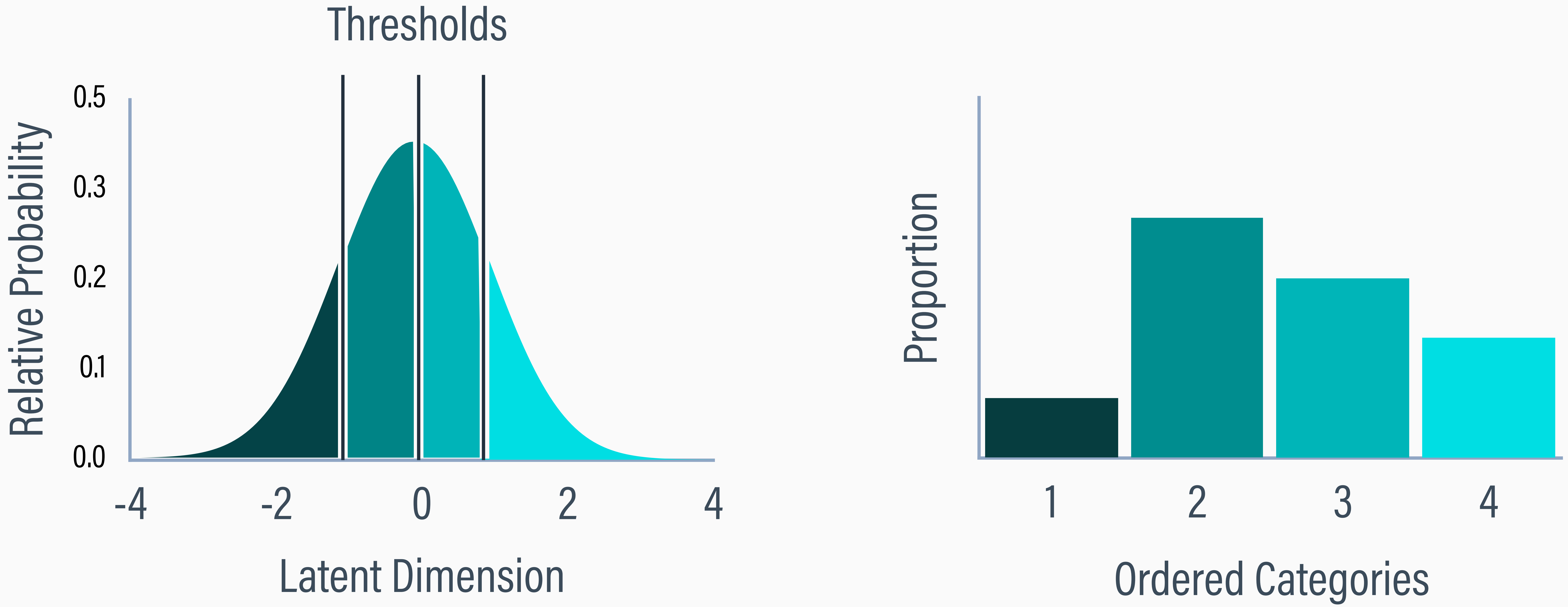
LATENT AND DISCRETE DISTRIBUTIONS

- The threshold parameter divides the latent distribution into segments, with areas under the curve matching the bar plot



ORDINAL VARIABLES

- Multiple threshold parameters divide the latent distribution into segments, with areas under the curve matching the bar plot

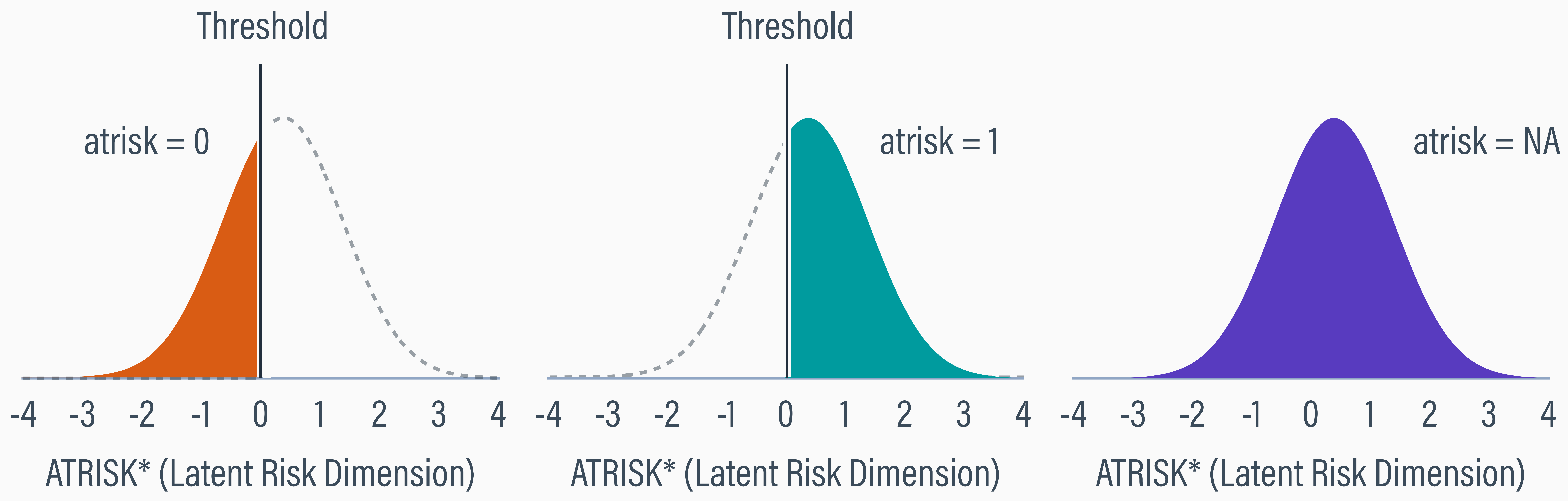


IMPUTING LATENT RESPONSE SCORES

- Latent response scores are missing data to be imputed
- MCMC uses computer simulation to “sample” latent response scores distributions, just like any other incomplete variable
- Blimp uses the latent response scores to link the categorical predictor to other continuous predictors, but the binary dummy code is the regressor in the focal model

LATENT AND DISCRETE DISTRIBUTIONS

- Latent imputations must fall below or above threshold if the binary variable is observed, and they are unconstrained if missing



ANALYSIS 2 REGRESSION MODEL

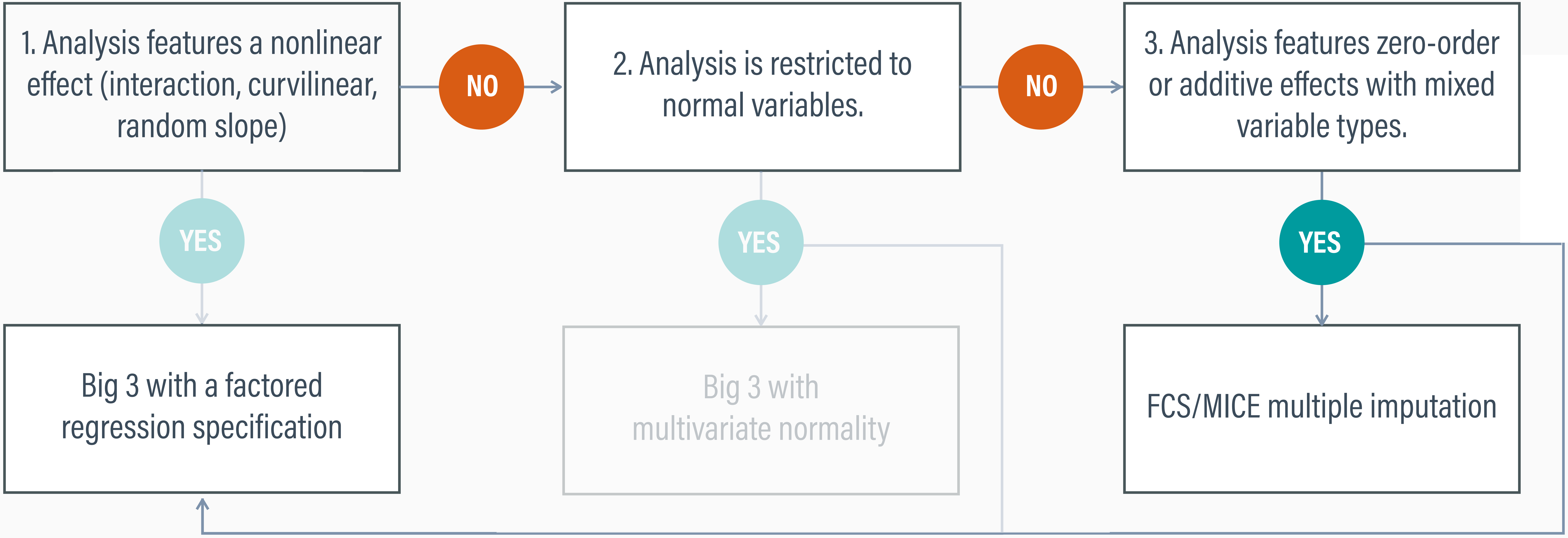
- Academic variables from 1st grade predicting 9th grade reading achievement in middle school

$$read_9 = \beta_0 + \beta_1(read_1) + \beta_2(lrnprob_1) + \beta_3(atrisk) + \varepsilon$$

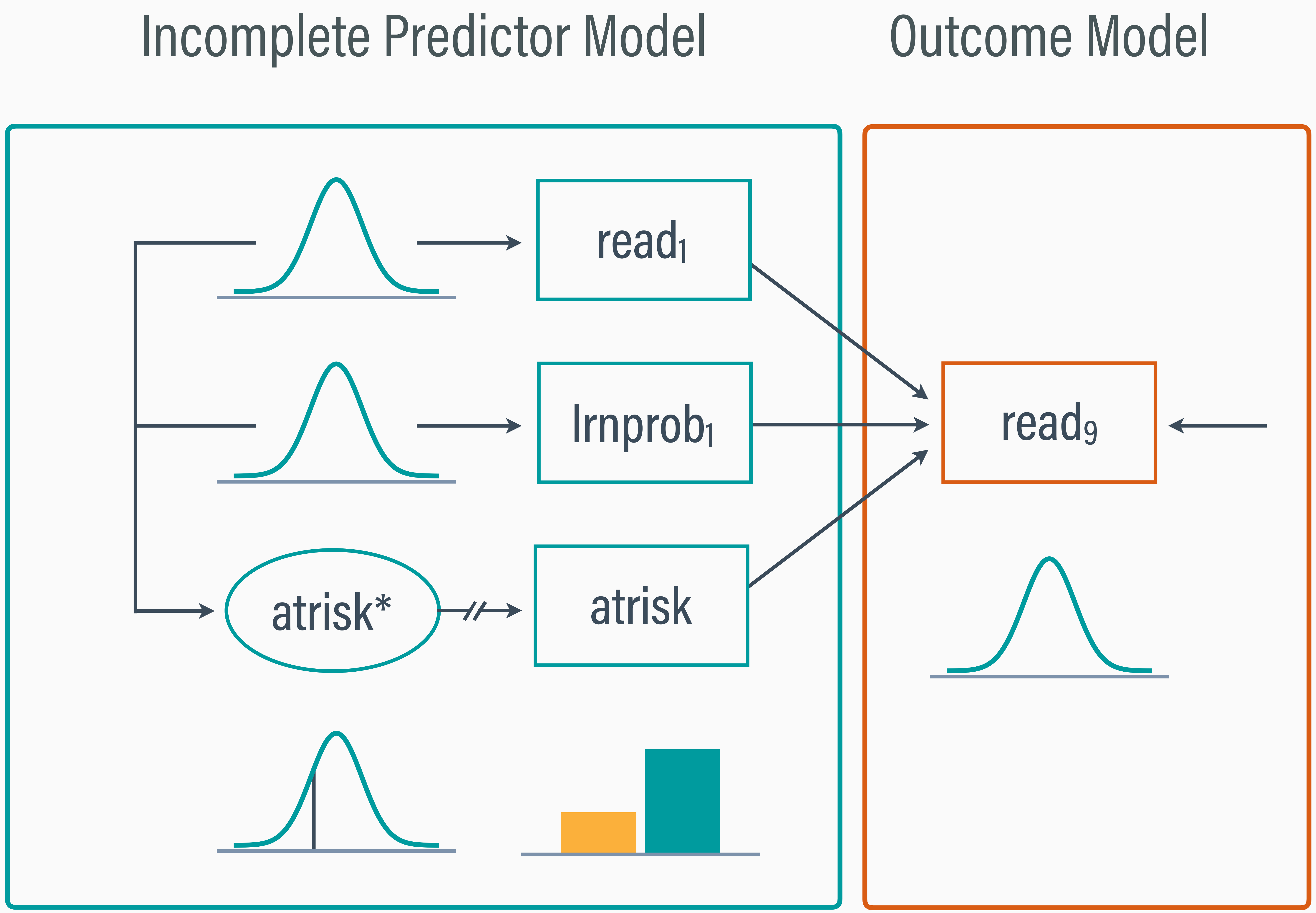
Variable	Definition	Missing %	Scale
atrisk	Emotional/behavioral risk code	2.2	0 = Low, 1 = Medium/high
lrnprob1	1st grade learning problems	2.2	Numeric (31 to 88)
read1	1st grade broad reading composite	6.5	Numeric (39 to 153)
read9	9th grade broad reading composite	17.4	Numeric (41 to 123)

MISSING DATA DECISION TREE

Mixtures of numeric and categorical variables reduce the number of viable options.



TWO-PART FACTORED SPECIFICATION



BLIMP STUDIO SCRIPT 2

```
DATA: reading.dat;
VARIABLES: id male hispanic riskgrp atrisk behsymp1 lrnprob1 read1 read2 read3 read9
  read9grp stanread7 math1 math2 math3 math9 math9grp stanmath7;
MISSING: 999;
NOMINAL:  atrisk;        # automatic dummy coding with lowest score as the reference
MODEL:  read9 ~ read1 lrnprob1 atrisk;
BURN: 10000;
ITER: 10000;
SEED: 90291;
```

RBLIMP SCRIPT (MODEL 2)

```
# fit model
model2 <- rblimp(
   data = reading,
   nominal = 'atrisk',        # automatic dummy coding with lowest score as the reference
   model = 'read9 ~ read1 lrnprob1 atrisk',
   seed = 90291,
   burn = 10000,
   iter = 10000)

# summarize results
output(model2)
posterior_plot(model2, 'read9')
```

PSR DIAGNOSTIC OUTPUT

The number of burn-in iterations is sufficient because the highest PSRF across all parameters is < 1.05 at the end of the burn-in period

```
BURN-IN POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

  NOTE: Split chain PSR is being used. This splits each chain's
        iterations to create twice as many chains.

  Comparing iterations across 2 chains     Highest PSR   Parameter #
                           251 to 500            1.034           21
                           501 to 1000           1.010           11
                           751 to 1500           1.010            9
                          1001 to 2000           1.006           13
                           ...     ...             ...          ...
                          4001 to 8000           1.001           18
                          4251 to 8500           1.001           21
                          4501 to 9000           1.001            8
                          4751 to 9500           1.001            8
                          5001 to 10000       »  1.001            8
```

MISSING DATA INFORMATION

- Per-variable missingness rates appear in the rows, and missing data patterns appear in the columns (M = missing)

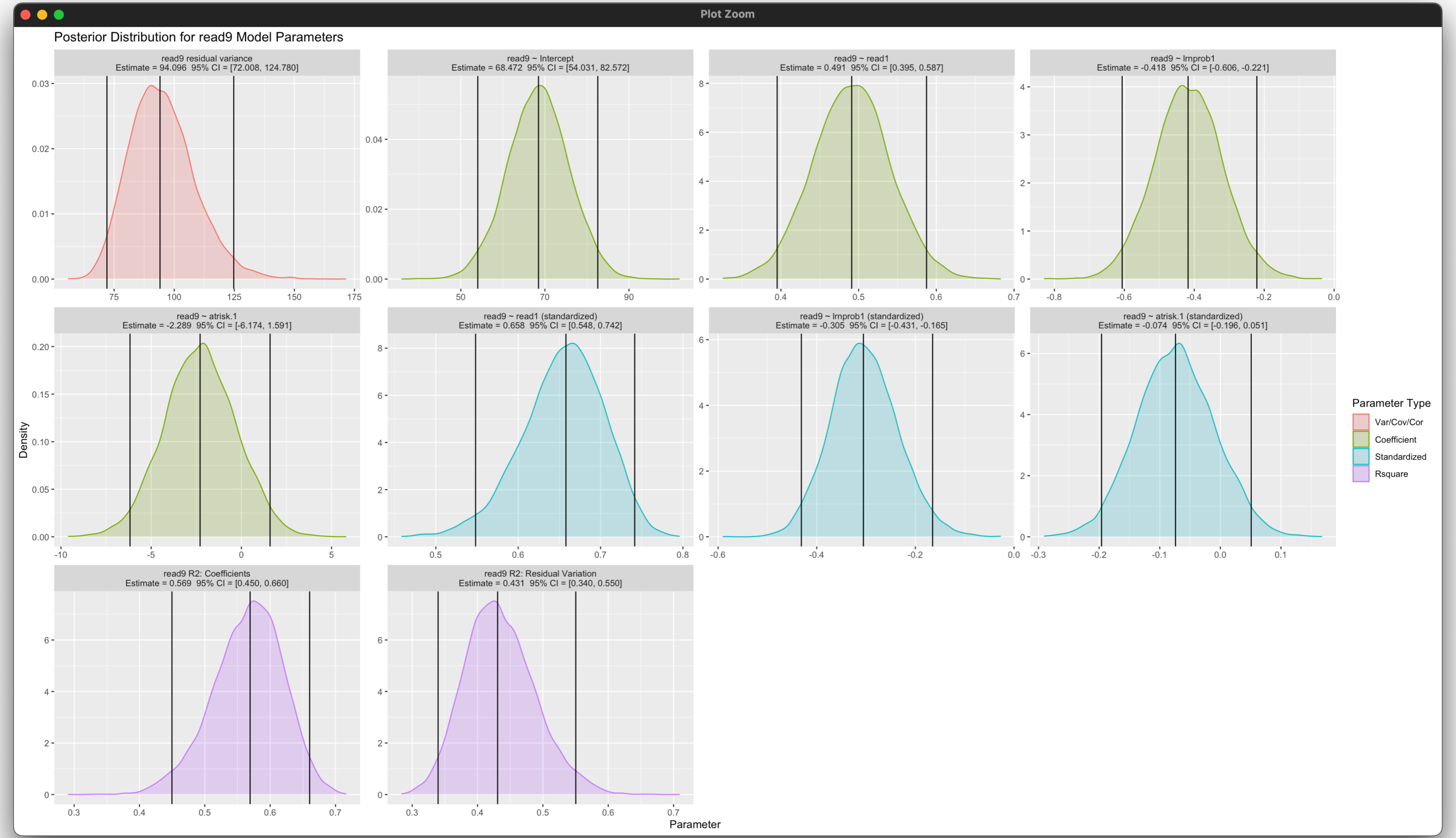
```
DATA INFORMATION:

  Sample Size:                 138
  Nominal Dummy Codes:

                 atrisk = atrisk.1
  Missing Data Info:
                         miss %      1     2     3     4     5     6
                               ---------------------------------------
                  read9 = 17.4       -     M     -     -     -     M
                 atrisk =  2.2       -     -     -     M     -     -
               lrnprob1 =  2.2       -     -     -     -     M     -
                  read1 =  6.5       -     -     M     -     -     M
                               ---------------------------------------
                               %  72.5  16.7   5.8   2.2   2.2   0.7
```

REGRESSION SUMMARY TABLE

OUTCOME MODEL ESTIMATES:

```
  Summaries based on 10000 iterations using 2 chains.
  NOTE: Estimate column based on posterior median.

Outcome Variable:  read9

Parameters                         Estimate     StdDev       2.5%      97.5%      ChiSq     PValue      N_Eff
                                ------------------------------------------------------------------------------
Variances:
  Residual Var.                      94.096     13.521     71.998    124.781        ---        ---   5093.390

Coefficients:
  Intercept                          68.472      7.253     54.030     82.578     89.064      0.000   3563.037
  read1                               0.491      0.049      0.395      0.587     99.355      0.000   6011.826
  lrnprob1                           -0.418      0.098     -0.606     -0.221     18.237      0.000   2740.547
  atrisk.1                           -2.289      1.990     -6.180      1.591      1.320      0.251   5821.121

...

Proportion Variance Explained
  by Coefficients                     0.569      0.054      0.450      0.660        ---        ---   5125.545
  by Residual Variation               0.431      0.054      0.340      0.550        ---        ---   5125.545

                                ------------------------------------------------------------------------------
```

DISTRIBUTION PLOTS (RBLIMP ONLY)



INTERPRETATIONS

- Blimp automatically dummy codes nominal variables, treating the lowest numeric code (low risk) as the reference
- For two students with same first grade reading and learning problems scores, being classified as at risk is associated with a –2.29 point decrease in grade 9 reading (parameter SD = 1.99)
- Slopes for numeric predictors are similar to the first analysis

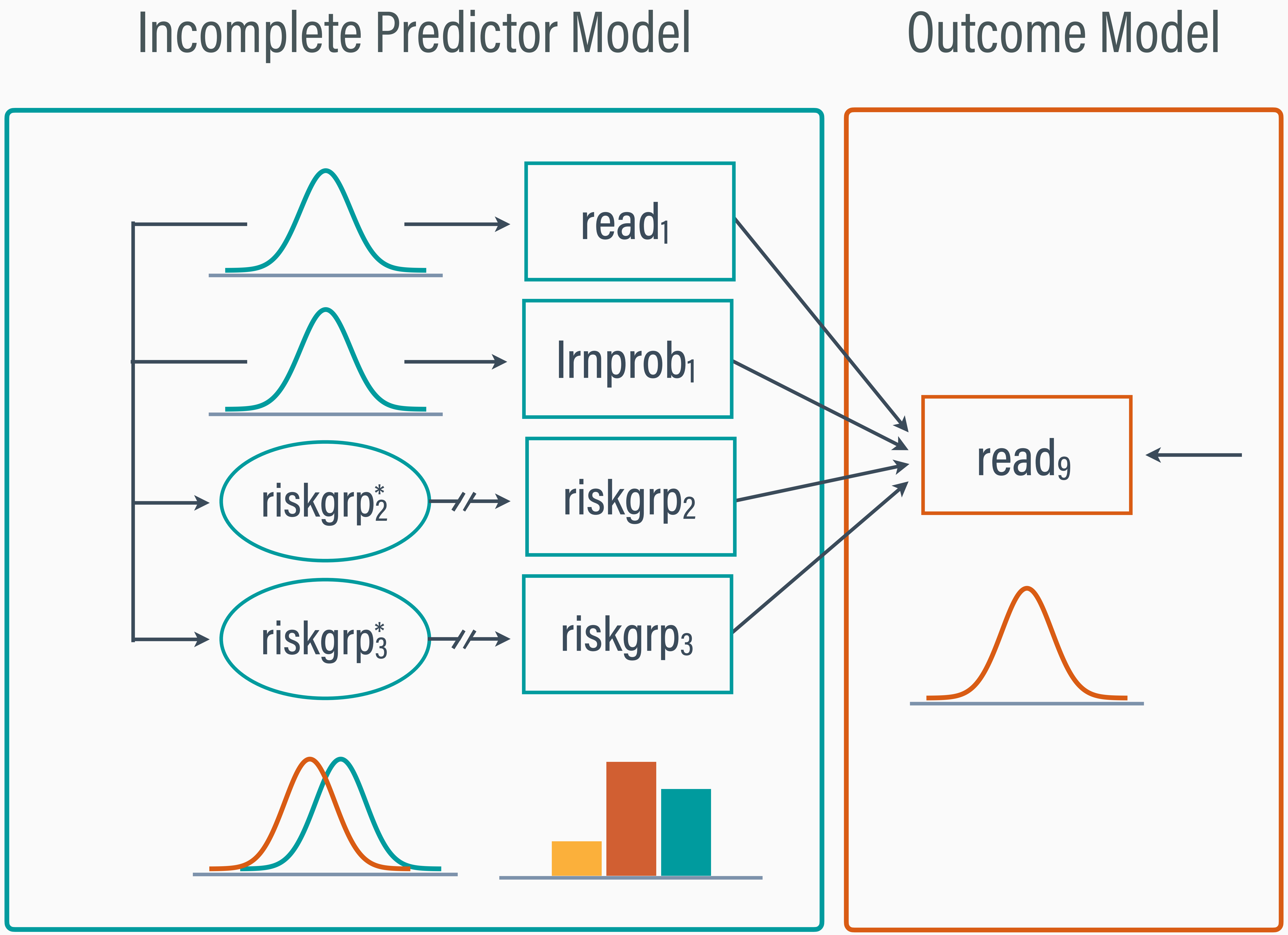
ANALYSIS 3 REGRESSION MODEL

- Risk status is represented by two dummy codes with the low-risk group (riskgrp = 1) serving as the reference group

$$read_9 = \beta_0 + \beta_1(read_1) + \beta_2(lrnprob_1) + \beta_3(riskgrp_2) + \beta_4(riskgrp_3) + \varepsilon$$

Variable	Definition	Missing %	Scale
riskgrp	Emotional/behavioral risk code	2.2	1 = Low, 2 = Medium, 3 = 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)

TWO-PART FACTORED SPECIFICATION



BLIMP STUDIO SCRIPT 3

```
DATA: reading.dat;
VARIABLES: id male hispanic riskgrp atrisk behsymp1 lrnprob1 read1 read2 read3 read9
  read9grp stanread7 math1 math2 math3 math9 math9grp stanmath7;
MISSING: 999;
NOMINAL:  riskgrp;        # automatic dummy coding with lowest score as the reference
MODEL:  read9 ~ read1 lrnprob1 riskgrp;
BURN: 10000;
ITER: 10000;
SEED: 90291;
```

RBLIMP SCRIPT (MODEL 3)

```
# fit model
model3 <- rblimp(
   data = reading,
   nominal = 'riskgrp',       # automatic dummy coding with lowest score as the reference
   model = 'read9 ~ read1 lrnprob1 riskgrp',
   seed = 90291,
   burn = 10000,
   iter = 10000)

# summarize results
output(model3)
posterior_plot(model3, 'read9')
```

PSR DIAGNOSTIC OUTPUT

The number of burn-in iterations is sufficient because the highest PSRF across all parameters is < 1.05 at the end of the burn-in period

```
BURN-IN POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

  NOTE: Split chain PSR is being used. This splits each chain's
        iterations to create twice as many chains.

  Comparing iterations across 2 chains     Highest PSR   Parameter #
                          251 to 500             1.096            18
                          501 to 1000            1.037            18
                          751 to 1500            1.012            14
                         1001 to 2000            1.018            14
                          ...    ...              ...            ...
                         4001 to 8000            1.003            25
                         4251 to 8500            1.003            18
                         4501 to 9000            1.002            18
                         4751 to 9500            1.002            18
                         5001 to 10000        »  1.002            14
```

MISSING DATA INFORMATION

- Per-variable missingness rates appear in the rows, and missing data patterns appear in the columns (M = missing)

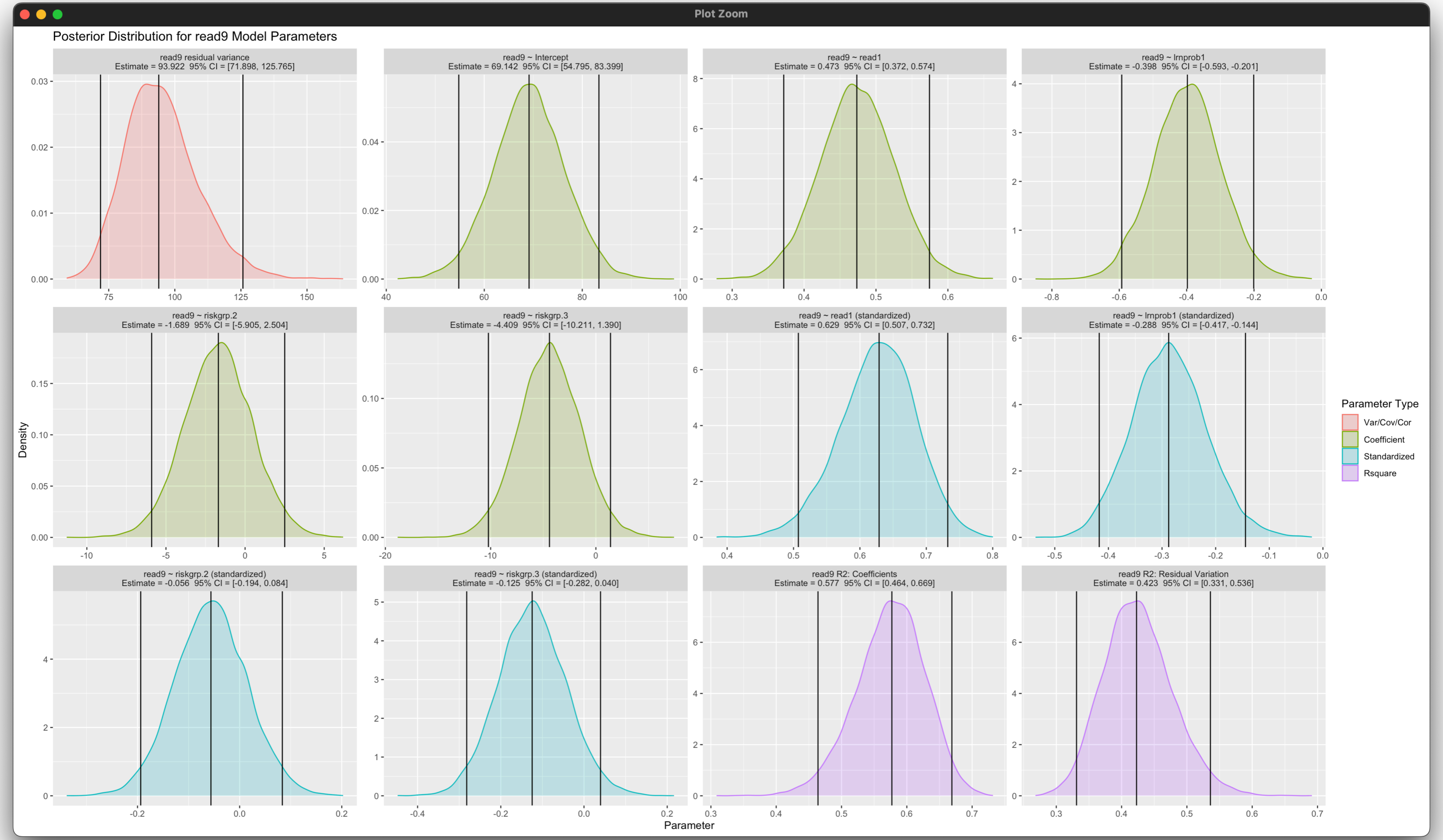
```
DATA INFORMATION:

  Sample Size:                138
  Nominal Dummy Codes:

                riskgrp = riskgrp.2 riskgrp.3
  Missing Data Info:
                        miss %      1     2     3     4     5     6
                                  --------------------------------------
                  read9 = 17.4      -     M     -     -     -     M
                riskgrp =  2.2      -     -     -     M     -     -
               lrnprob1 =  2.2      -     -     -     -     M     -
                  read1 =  6.5      -     -     M     -     -     M
                                  --------------------------------------
                                  %  72.5  16.7   5.8   2.2   2.2   0.7
```

REGRESSION SUMMARY TABLE

OUTCOME MODEL ESTIMATES:

```
  Summaries based on 10000 iterations using 2 chains.
  NOTE: Estimate column based on posterior median.

Outcome Variable:   read9

Parameters                      Estimate     StdDev       2.5%      97.5%      ChiSq     PValue      N_Eff
                                ----------------------------------------------------------------------------
Variances:
  Residual Var.                   93.922     13.762     71.897    125.783        ---        ---   5004.950

Coefficients:
  Intercept                       69.142      7.228     54.792     83.413     91.460      0.000   3685.210
  read1                            0.473      0.052      0.372      0.574     84.275      0.000   6143.602
  lrnprob1                        -0.398      0.100     -0.593     -0.201     15.889      0.000   3028.647
  riskgrp.2                       -1.689      2.132     -5.906      2.505      0.633      0.426   5716.054
  riskgrp.3                       -4.409      2.936    -10.212      1.391      2.263      0.133   4779.567

...

Proportion Variance Explained
  by Coefficients                  0.577      0.053      0.464      0.669        ---        ---   5100.899
  by Residual Variation            0.423      0.053      0.331      0.536        ---        ---   5100.899

                                ----------------------------------------------------------------------------
```

DISTRIBUTION PLOTS (RBLIMP ONLY)



INTERPRETATIONS

- Blimp automatically dummy codes nominal variables, treating the lowest numeric code (low risk) as the reference
- For two students with same first grade reading and learning problems scores, being classified as moderate versus low risk is associated with a –1.69 point decrease in grade 9 reading
- For two students with same first grade reading and learning problems scores, being classified as high versus low risk is associated with a –4.41 point decrease in grade 9 reading

OUTLINE

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. MCMC Estimation and Missing Data Imputation
4. Fitting Regression Models in Blimp
5. Incomplete Categorical Variables
6. Interaction Effects
7. Reporting Results From a Missing Data Analysis

MODERATED REGRESSION

- Moderation occurs when a focal predictor's influence on an outcome depends on a third variable called a moderator
- Moderated regression answers the question, for whom does an effect apply?
- Does the diagnostic utility of first-grade reading performance on ninth-grade reading achievement depend on whether a student is experiencing learning problems in first grade?

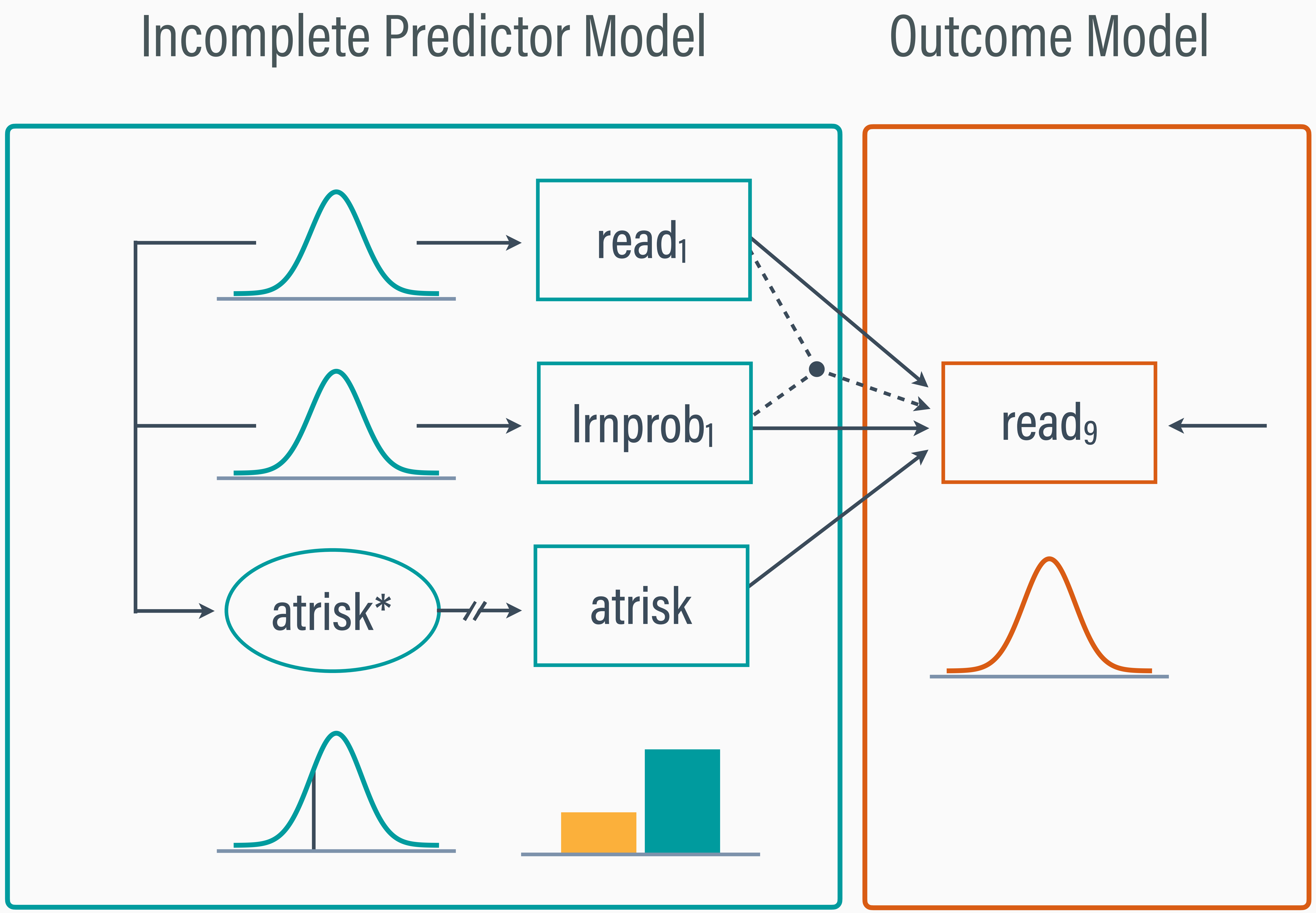
ANALYSIS 4 REGRESSION MODEL

- An interaction is formed by multiplying two predictors, either of which (or both) could be incomplete

$$read_9 = \beta_0 + \beta_1(read_1) + \beta_2(lrnprob_1) + \beta_3(read_1)(lrnprob_1) + \beta_4(atrisk) + \varepsilon$$

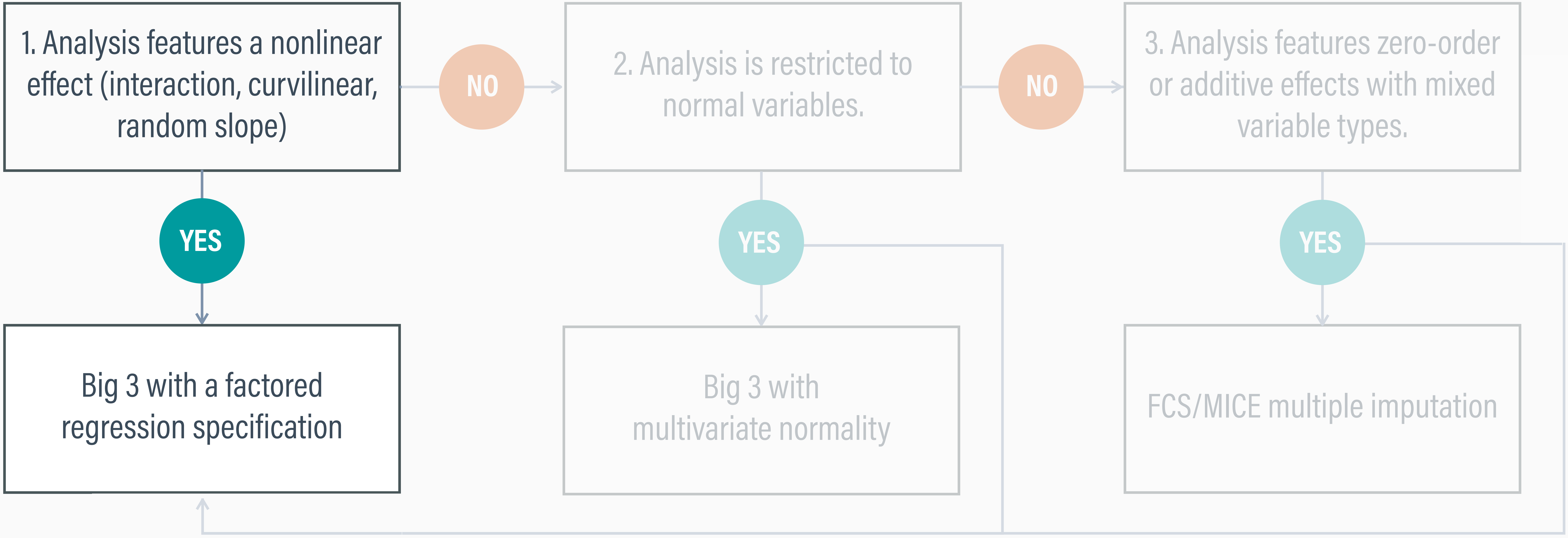
Variable	Definition	Missing %	Scale
riskgrp	Emotional/behavioral risk code	2.2	1 = Low, 2 = Medium, 3 = High
lrnprob1	1st grade learning problems	2.2	Numeric (31 to 88)
read1	1st grade broad reading composite	6.5	Numeric (39 to 153)
read9	9th grade broad reading composite	17.4	Numeric (41 to 123)

TWO-PART FACTORED SPECIFICATION



MISSING DATA DECISION TREE

Interaction and nonlinear effects require factored specifications with specialized software.

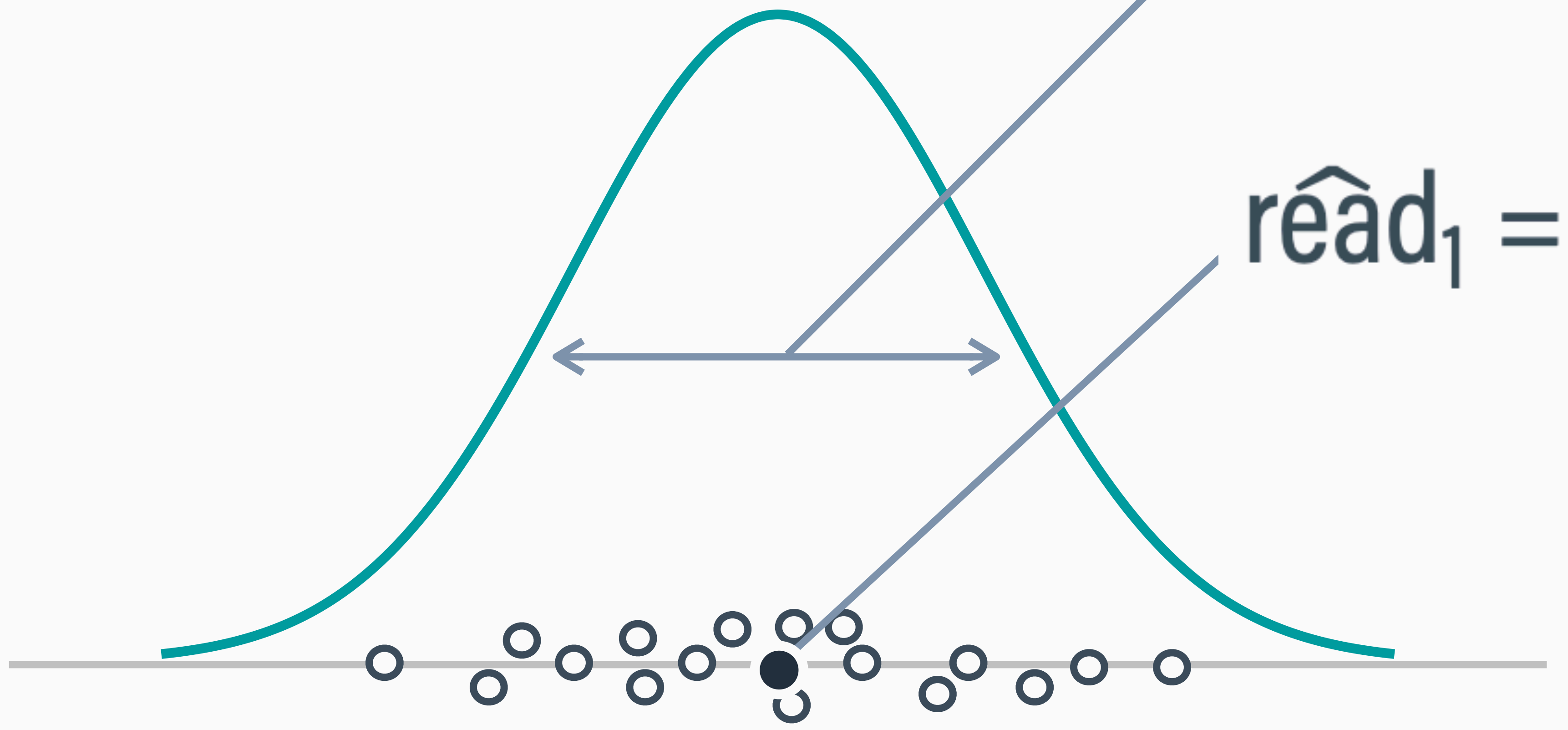


INCOMPLETE PRODUCT TERMS

- Products are deterministic functions of lower-order predictors in the focal model rather than unique variables themselves
- The two-part imputation for incomplete predictors remains the same as before
- When the interaction is non-zero, the two-part specification produces non-normal (heteroscedastic) predictor imputations that accommodate the nonlinear term in the focal model

PREDICTED VALUES AND VARIATION

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



$$\sigma^2_{read_1|\ldots} = \left(\frac{1}{\sigma^2_e} + \frac{\left(\beta_1 + \beta_3 lrnprob_1\right)^2}{\sigma^2_\varepsilon}\right)^{-1}$$

$$\widehat{read}_1 = \sigma^2_{read_1|\ldots} \times \left(\frac{\gamma_0 + \gamma_1 lrnprob_1 + \gamma_2 atrisk^*}{\sigma^2_e} + \frac{\left(\beta_1 + \beta_3 lrnprob_1\right)\left(read_9 - \beta_0 - \beta_3 lrnprob_1 - \beta_4 atrisk\right)}{\sigma^2_\varepsilon}\right)$$

BLIMP STUDIO SCRIPT 4

```
DATA: reading.dat;
VARIABLES: id male hispanic riskgrp atrisk behsymp1 lrnprob1 read1 read2 read3 read9
   read9grp stanread7 math1 math2 math3 math9 math9grp stanmath7;
MISSING: 999;
NOMINAL:  atrisk;
CENTER: read1 lrnprob1;
MODEL:  read9 ~ read1 lrnprob1 read1*lrnprob1 atrisk;  # product in focal model
SIMPLE: read1 | lrnprob1;  # conditional effects of read1 at different levels of lrnprob1
BURN: 10000;
ITER: 10000;
SEED: 90291;
```

RBLIMP SCRIPT (MODEL 4)

```
# fit model
model4 <- rblimp(
    data = reading,
    nominal = 'atrisk',
    center = 'read1 lrnprob1',
    model = 'read9 ~ read1 lrnprob1 read1*lrnprob1 atrisk',     # product in focal model
    simple = 'read1 | lrnprob1',                                 # conditional effects of read1 at different levels of lrnprob1
    seed = 90291,
    burn = 10000,
    iter = 10000)

# summarize results
output(model4)
posterior_plot(model4, 'read9')
simple_plot(read9 ~ read1 | lrnprob1, model4)   # plot simple slopes at different values of moderator
jn_plot(read9 ~ read1 | lrnprob1, model4)   # Johnson-Neyman regions of significance
```

PSR DIAGNOSTIC OUTPUT

The number of burn-in iterations is sufficient because the highest PSRF across all parameters is < 1.05 at the end of the burn-in period

BURN-IN POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

```
NOTE: Split chain PSR is being used. This splits each chain's
      iterations to create twice as many chains.

Comparing iterations across 2 chains     Highest PSR   Parameter #
                         251 to 500            1.032            15
                         501 to 1000           1.023            21
                         751 to 1500           1.054            21
                        1001 to 2000           1.055            21
                         ...     ...            ...            ...
                        4001 to 8000           1.002            21
                        4251 to 8500           1.005            21
                        4501 to 9000           1.006            21
                        4751 to 9500           1.003            21
                        5001 to 10000     »    1.007            21
```

MISSING DATA INFORMATION

- Per-variable missingness rates appear in the rows, and missing data patterns appear in the columns (M = missing)

```
DATA INFORMATION:

  Sample Size:                  138
  Nominal Dummy Codes:

                  atrisk = atrisk.1
  Missing Data Info:
                         miss %       1     2     3     4     5     6
                                  --------------------------------------
                   read9 = 17.4       -     M     -     -     -     M
                  atrisk =  2.2       -     -     -     M     -     -
                lrnprob1 =  2.2       -     -     -     -     M     -
                   read1 =  6.5       -     -     M     -     -     M
                                  --------------------------------------
                                  %  72.5  16.7   5.8   2.2   2.2   0.7
```

REGRESSION SUMMARY TABLE

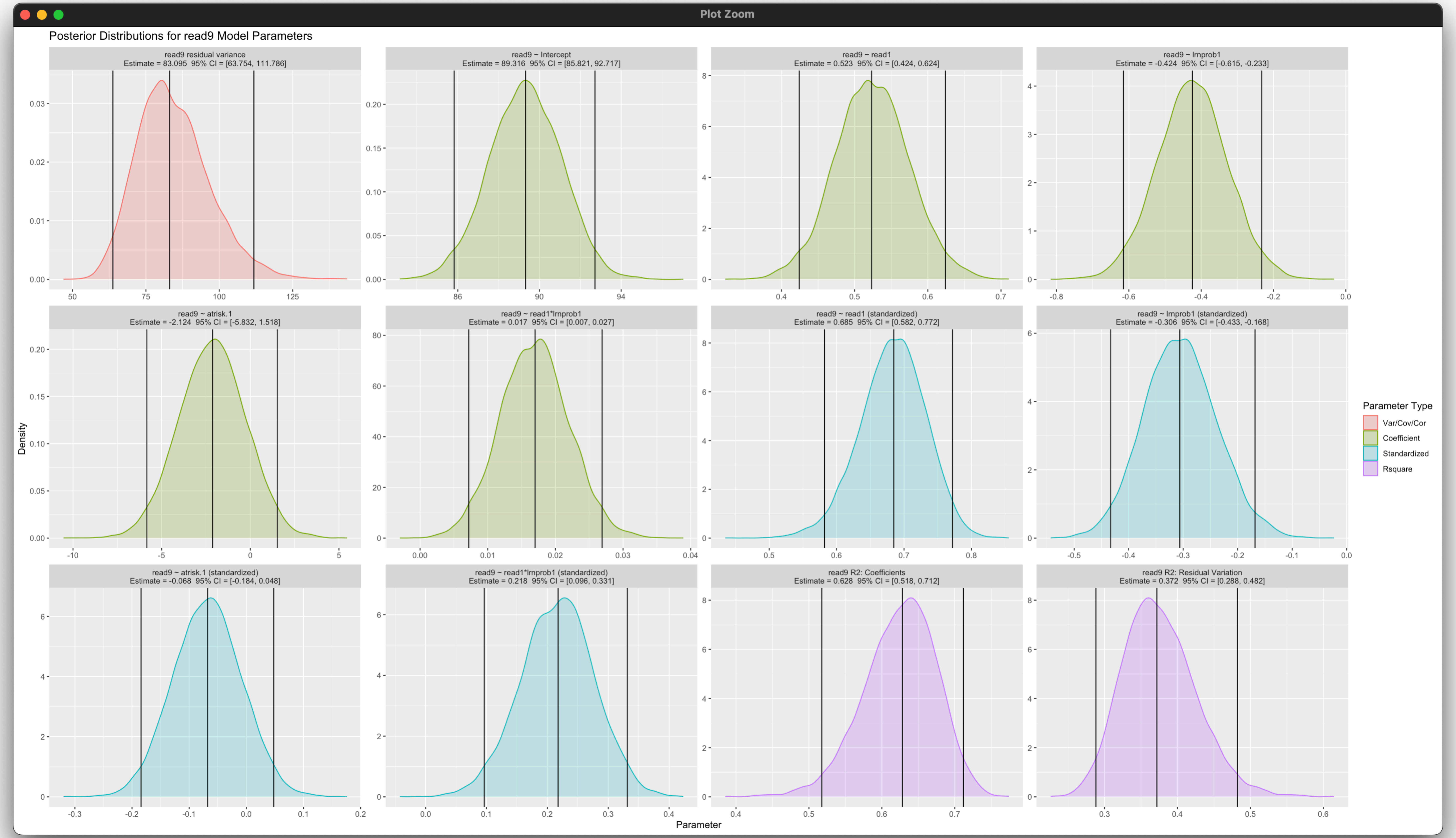
OUTCOME MODEL ESTIMATES:

Summaries based on 10000 iterations using 2 chains.
NOTE: Estimate column based on posterior median.

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
...							
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

DISTRIBUTION PLOTS (RBLIMP ONLY)



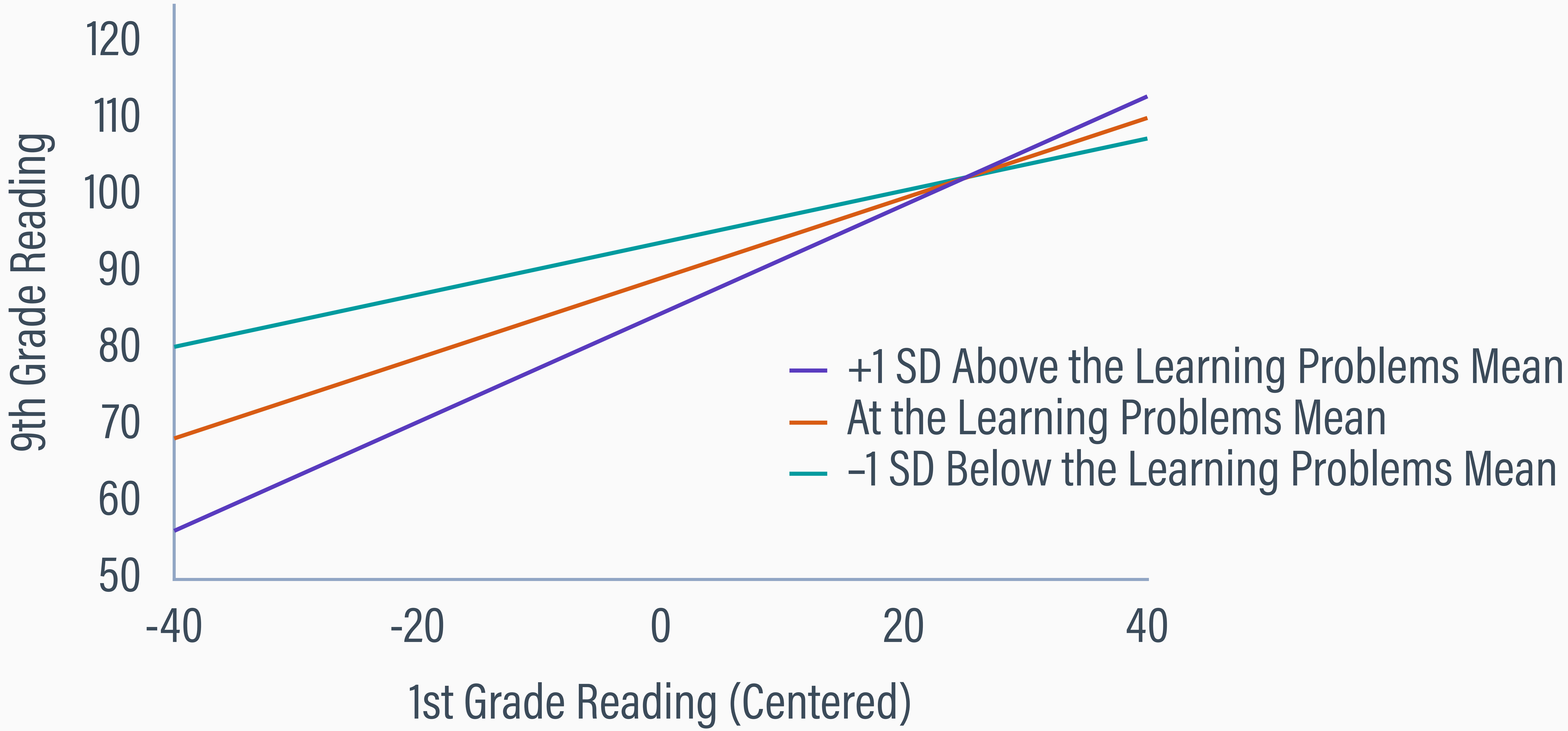
INTERPRETATIONS

- Lower-order terms are conditional effects that depend on a meaningful zero, which was achieved by centering
- For two students at the mean of the learning problems distribution, scoring one point higher on the first grade reading test is associated with a 0.52 increase in ninth-grade reading
- For two students at the mean of the first grade reading distribution, being rated one point higher on the first grade learning problems measure is associated with a –0.42 decrease in ninth-grade reading

INTERPRETATIONS, CONTINUED

- The interaction coefficient, = .02, is the amount by which the focal slope changes for a one-unit difference on the moderator
- For two students at the mean of the learning problems distribution, scoring one point higher on the first grade reading test is associated with a 0.52 increase in ninth-grade reading
- For two students one point above the learning problems mean, scoring one point higher on the first grade reading test is associated with a 0.52 + 0.02 increase in ninth-grade reading

CONDITIONAL EFFECTS (SIMPLE SLOPES)

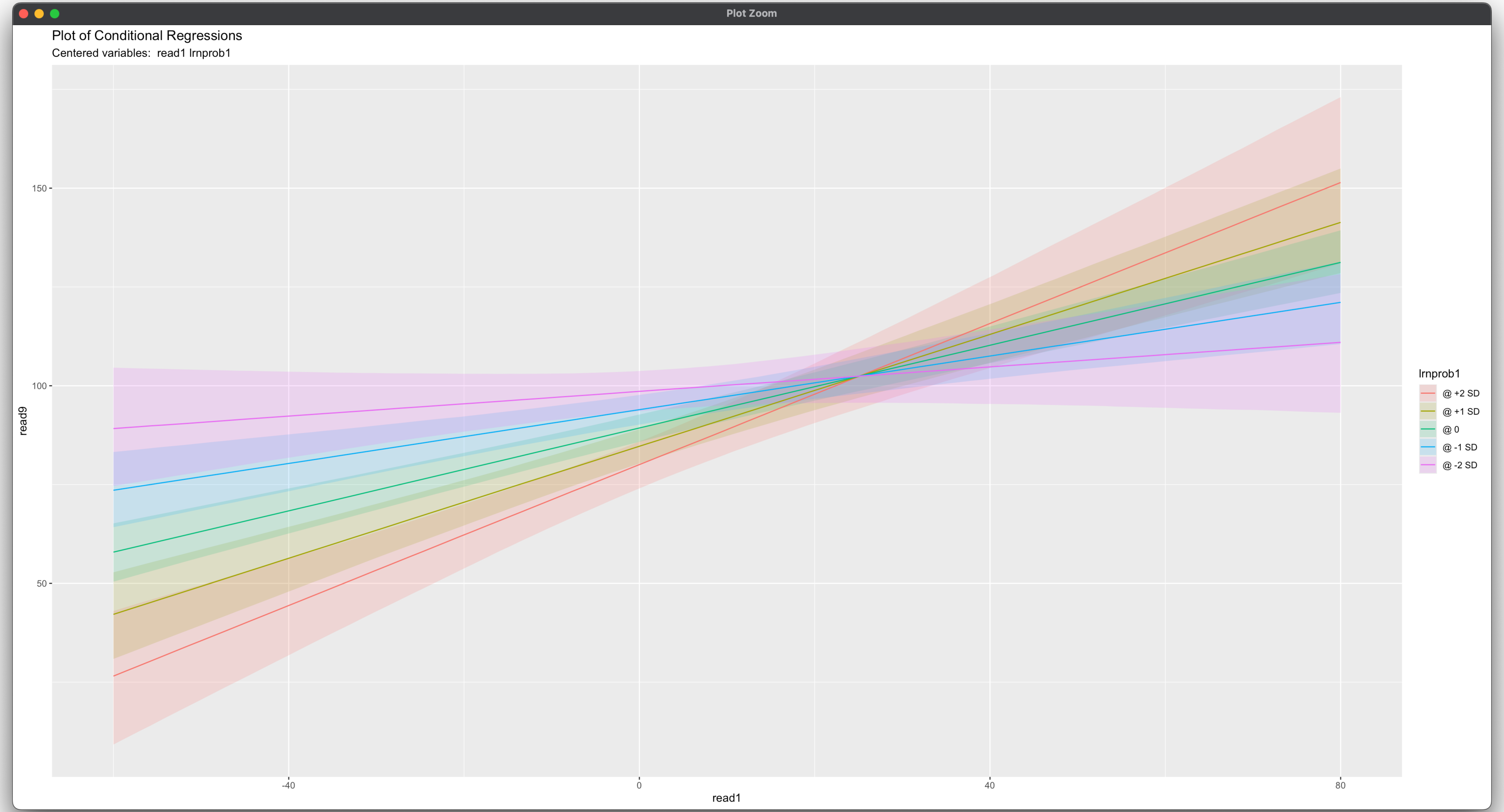


CONDITIONAL EFFECTS SUMMARY TABLE

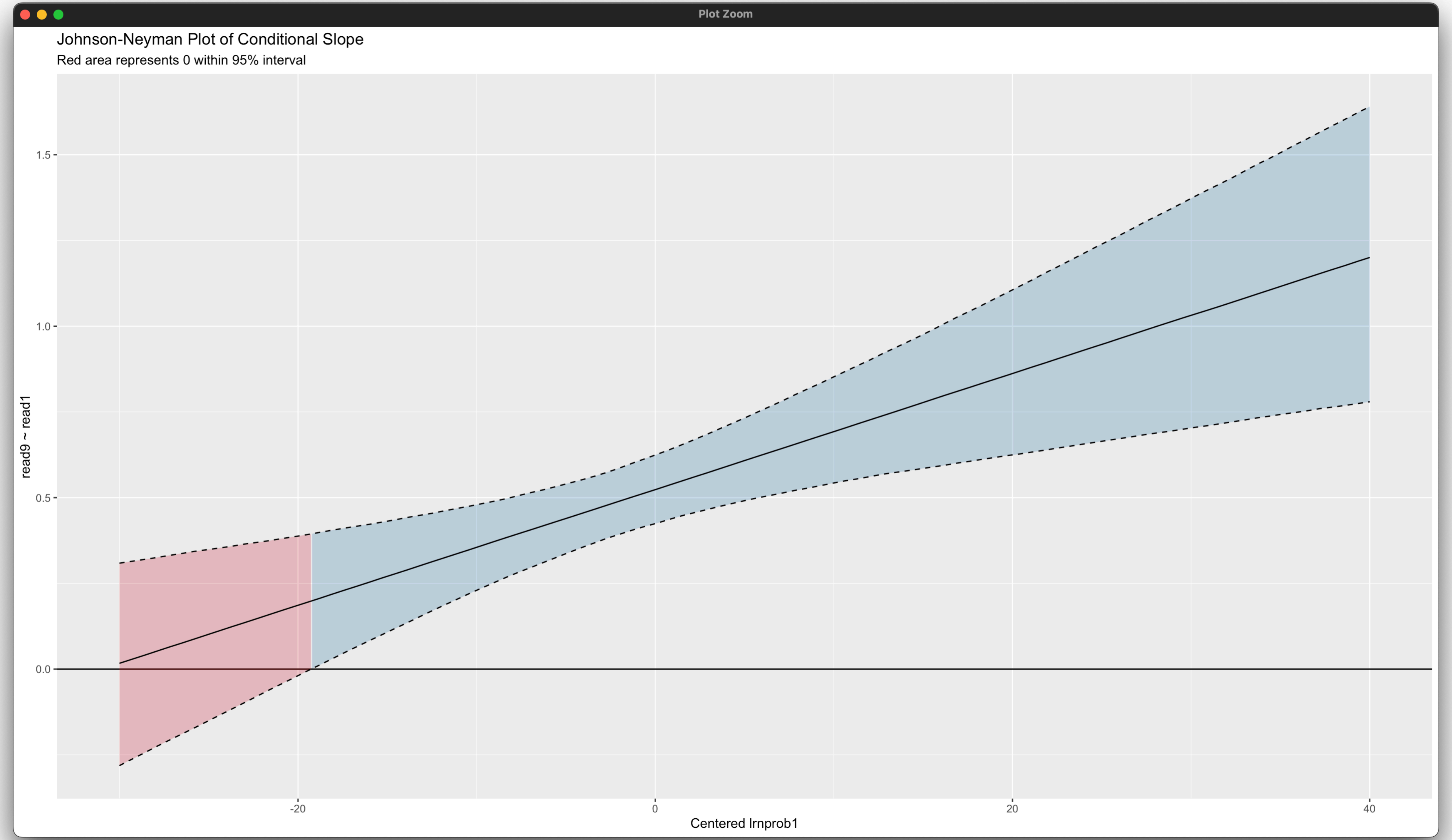
Conditional Effects	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
read1 | lrnprob1 @ +2 SD							
Intercept	80.052	3.019	74.043	85.852	702.721	0.000	1602.353
Slope	0.893	0.132	0.639	1.158	45.808	0.000	3153.284
read1 | lrnprob1 @ +1 SD							
Intercept	84.692	2.211	80.319	89.026	1467.138	0.000	1532.751
Slope	0.709	0.083	0.550	0.876	73.692	0.000	3170.278
read1 | lrnprob1 @ 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 | lrnprob1 @ -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
read1 | lrnprob1 @ -2 SD							
Intercept	98.582	2.582	93.557	103.730	1458.470	0.000	3236.081
Slope	0.156	0.114	-0.073	0.373	1.832	0.176	3688.862

NOTE: Intercepts are computed by setting all predictors not involved in the conditional effect to zero.

CONDITIONAL EFFECT PLOTS (RBLIMP ONLY)



JOHNSON-NEYMAN PLOTS (RBLIMP ONLY)



OUTLINE

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. MCMC Estimation and Missing Data Imputation
4. Fitting Regression Models in Blimp
5. Incomplete Categorical Variables
6. Interaction Effects
7. Reporting Results From a Missing Data Analysis

REPORTING CHECKLIST

1. Missing Data Rates
2. Assumed Missing Data Process
3. Missing Data Handling Method
4. Software Tools Used
5. Algorithmic and Model-Checking Details
6. Interpretation Details

MISSING DATA RATES AND ASSUMPTIONS

Missing data rates on predictor variables measured in first grade ranged between 2.2% (risk status and learning problems) and 6.5% (reading performance). Approximately 17.9% of ninth-grade reading scores were missing. Our analyses are based on the conditionally missing at random assumption where missingness systematically vary as a function of observed scores but not the unseen scores.

MISSING DATA APPROACH AND SOFTWARE

We fit our models using Bayesian MCMC estimation in the Blimp 3 software (Keller & Enders, 2021). Given the same assumptions and data, MCMC and likelihood-based missing data handling procedures are numerically equivalent (Enders, 2022). However, MCMC estimation is preferable because classic FIML estimator is known to introduce bias when applied to models interactive effects (the so-called “just another variable” approach; Lüdtke et al., 2020; Zhang & Wang, 2017). MCMC is similar to FIML in the sense that it directly estimates the model of interest, and it is similar to multiple imputation in the sense that it averages over thousands of realizations of the missing values

ALGORITHMIC DETAILS

We used Blimp's default diffuse (non-informative) prior distributions, as described in Section 1.7 of the user guide. We used two MCMC chains with random starting values to generate model summaries consisting of 10,000 estimates following an initial burn-in period of 10,000 cycles. The potential scale reduction factor convergence diagnostics (Gelman & Rubin, 1992) indicated that MCMC converged in fewer than 1,000 iterations, so a 10,000-cycle warm-up period was sufficiently conservative. We verified the total number of iterations for the analysis was sufficient by examining the effective number of independent MCMC samples for each parameter, all of which were greater than the recommended value of 100 (Gelman et al., 2014, p. 287).

REPORTING TEMPLATE CONTINUED

Table 1 displays the regression summary table from the analysis. From a Bayesian perspective, the posterior medians and standard deviations are analogous to frequentist point estimates and standard errors, and the 95% credible interval limits are akin to confidence intervals. However, these quantities make no reference to repeated samples but instead convey parameter values that are consistent with the observed data. Given the same assumptions and data, Bayesian and likelihood-based missing data handling procedures are numerically equivalent (Enders, 2022). Taking a computational frequentist perspective, these MCMC-generated summaries can also be viewed as surrogates for frequentist point estimates, standard errors, and confidence intervals (Levy & McNeish, 2021)

SUMMARY TABLE

Table 1

Regression Model Summary

Parameter	Est.	SD	2.5%	97.5%	Chi-Sq.	p
Intercept	89.32	1.75	85.82	92.72	2599.23	< .001
1st Grade Reading	0.52	0.05	0.42	0.62	109.00	< .001
Learning Problems	−0.42	0.10	−0.62	−0.23	19.14	< .001
Reading by Problems	0.02	0.01	0.01	0.03	11.19	< .001
At Risk Indicator	−2.12	1.89	−5.84	1.52	1.28	.26
Residual variance	83.10	12.31	63.75	111.83	—	—
R^2	0.63	0.05	0.52	0.71	—	—

SUMMARY OF RESULTS

Collectively, the predictors explained approximately 63% of the variation in 9th grade reading scores. At the learning problems mean, first grade reading exhibited a significant positive association with 9th grade reading performance ($\beta = 0.52$, SD = 0.05, CI = [0.42,0.62]), controlling for other predictors. The MCMC Wald test (Asparouhov & Muthén, 2021) of the interaction effect was statistically significant, $\chi^2(1) = 11.19$, $p < .001$. The positive interaction coefficient indicates that the first grade reading slope increases as learning problems increase ($\beta = 0.02$, SD = 0.01, CI = [0.01,0.03]), such that first grade test scores become increasingly predictive of later achievement. Figure 1 displays the simple slopes at three levels of learning problems.

For more information go to

WWW.APPLIEDMISSINGDATA.COM