ACQUISITION INNOVATION
RESEARCH CENTER

A Comparison of Methods for Integrated Evaluation of Complex Systems

Justin Krometis
Virginia Tech National Security Institute

VIRGINIA TECH®

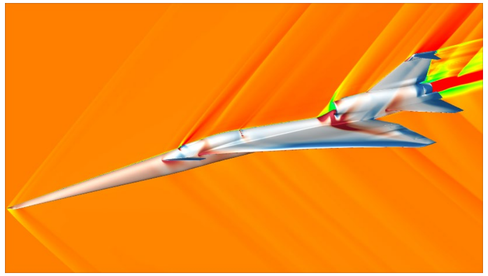
Background: Integrated Test & Evaluation

Motivation and Methods

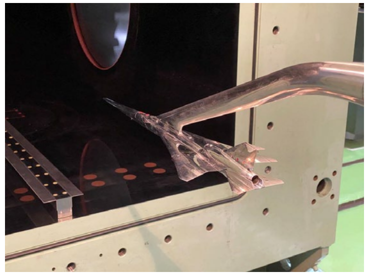
Why Integrated Testing?

DoD programs collect data throughout the acquisition life cycle, for example:

Historical Data,
SME Knowledge



Model Data



Developmental
Test Data

Operational
Test Data

Operational
Performance

Leveraging all data enables better understanding of systems *earlier*…

…allowing for fewer or more optimal tests *later*

Policy Motivation (DoDI 5000.98)

3.1(a):

"OT&E and LFT&E planning, execution, analysis, and reporting activities will use **the latest advances in science (e.g., design of experiments, statistical inference methods, or big data analytics)** to … determine, with scientific rigor, the preliminary and final operational effectiveness, suitability, survivability, and lethality (as applicable) of DoD systems."

3.1(c):

"Science and technology-based OT&E and LFT&E will **enable efficient use of data from multiple data sources** (e.g., contractor test (CT), developmental test (DT), operational test (OT), and live fire test (LFT) data or M&S results). Improved **sequential testing using Bayesian or similar inference methods** … are critical to dynamically optimize the planning, execution, analysis, and reporting of integrated T&E, OT&E, and LFT&E across the acquisition life cycle."

Frequentist Methods

Fit a statistical model to all of the data

Naïve: Assume all data is equivalent and fit to all data equally

Blocking: Try to account for differences in data sources by adding source or phase-specific factors to the model

- Example: Add a shift parameter to account for possible biases in data sources

Bayesian Inference

Relates the probability of a parameter value θ given data Y ($P(\theta|Y)$) to the probability of Y given θ and the probability of θ:

$$P(\theta|Y) \propto P(Y|\theta)\,P(\theta)$$

Relates the probability of a parameter value θ given data Y ($P(\theta|Y)$) to the probability of Y given θ and the probability of θ:

$$P(\theta|Y) \propto P(Y|\theta)\boxed{P(\theta)}$$

Prior

Relates the probability of a parameter value θ given data Y ($P(\theta|Y)$) to the probability of Y given θ and the probability of θ:

$$P(\theta|Y) \propto \boxed{P(Y|\theta)}\,\boxed{P(\theta)}$$

Likelihood (Data) **Prior**

Bayesian Inference

Relates the probability of a parameter value θ given data Y ($P(\theta|Y)$) to the probability of Y given θ and the probability of θ:

$$P(\theta|Y) \propto P(Y|\theta) P(\theta)$$

Posterior — $P(\theta|Y)$

Likelihood (Data) — $P(Y|\theta)$

Prior — $P(\theta)$

Bayesian Inference

ACQUISITION INNOVATION RESEARCH CENTER

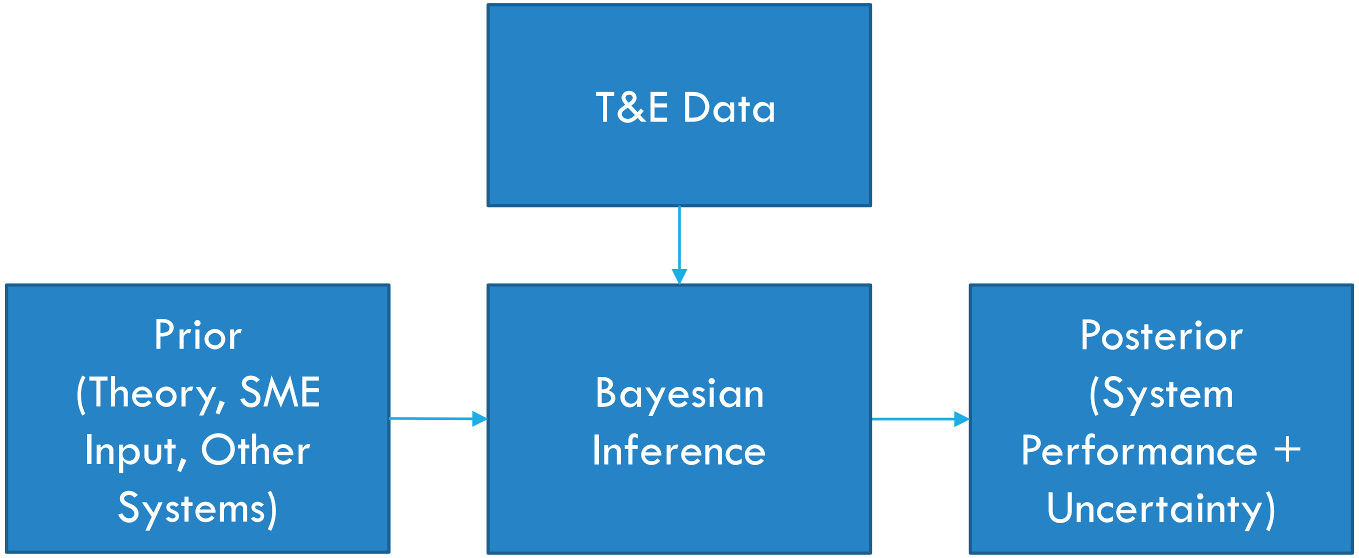
Relates the probability of a parameter value θ given data Y ($P(\theta|Y)$) to the probability of Y given θ and the probability of θ:

$$P(\theta|Y) \propto P(Y|\theta)P(\theta)$$

Posterior **Likelihood (Data)** **Prior**

Models human learning: Understanding (prior) + Experience (likelihood) = Updated understanding (posterior)

Single Test Phase



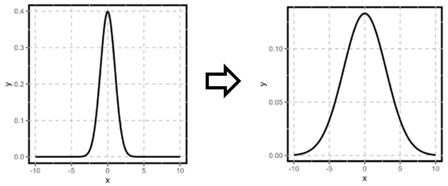
Single Test Phase

T&E Data

Prior (Theory, SME Input, Other Systems)

Bayesian Inference

Posterior (System Performance + Uncertainty)



Downweight/Adjust as Necessary

Prior (Based on Previous Test Phase)

Informative Priors

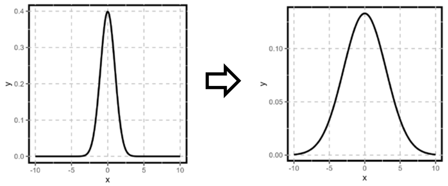
Single Test Phase

T&E Data

Prior (Theory, SME Input, Other Systems)

Bayesian Inference

Posterior (System Performance + Uncertainty)



Downweight/Adjust as Necessary

Next Test Phase

Prior (Based on Previous Test Phase)

T&

Ba

Inf

Application to T&E for DoD Programs

Considerations for Design & Evaluation of Tests

ACQUISITION INNOVATION RESEARCH CENTER

Consideration	Possible Levels (best case → worst case)
Safety	Minimal risk (not live projectile) → High risk (live fire)
Cost	$ → $$$
Resource Availability	Available → Partially available → Needs to be developed
Schedule	Easy & quick → Hard & extensive coordination
Historical operational performance data	Yes same factors → Yes but missing key factor(s) → None
Modeling and Simulation	Yes accurate and validated over time → Yes but not well understood and/or missing key factor(s) → None
Scale	Single component → Parallel systems → Series system

Considerations for Design & Evaluation of Tests

ACQUISITION INNOVATION RESEARCH CENTER

Consideration	Possible Levels (best case → worst case)
Safety	Minimal risk (not live projectile) → High risk (live fire)
Cost	$ → $$$
Resource Availability	Available → Partially available → Needs to be developed
Schedule	Easy & quick → Hard & extensive coordination
Historical operational performance data	Yes same factors →
Modeling and Simulation	Yes accurate and val and/or missing key f
Scale	Single component →

Variety of DoD programs means that a variety of analysis approaches may be appropriate!

Practical Challenge (for Academia, Anyway)

Real programmatic data is often restricted

- Hard to access
- Hard to engage students on
- Hard to publish/disseminate results

Solution: Synthetic data mimicking real challenges

- Benefit: We can evaluate methods because we know the “truth”

A Synthetic Case Study

Background: Synthetic Radar Models

IDA created[1] the following model of a synthetic counterfire radar:

$$Y = 79 - 6B + 4D - 7.5F + 5AF - 5.5BD + 4.5DF + 4D^2 - 9F^2$$

with the following factors:

Design Factor	Label	Type	Levels
Quadrant Elevation	A	Continuous	Low, High
Aspect Angle	B	Continuous	Incoming, Crossing
Munition Type	C	Categorical	Mortar, Rockets, Artillery
Shot Range	D	Continuous	Low, High
Operating Mode	E	Categorical	90, 360
Radar to Weapon Range	F	Continuous	Low, High

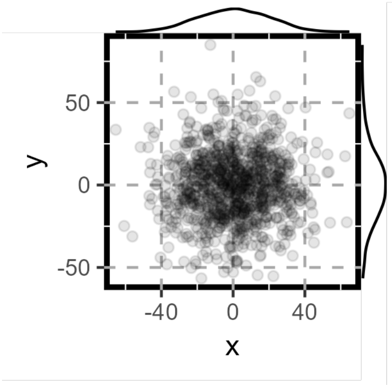
[1]Ahrens, Monica, Rebecca Medlin, Keyla Pagán-Rivera, and John W. Dennis. "Case Study on Applying Sequential Analyses in Operational Testing." *Quality Engineering* 35, no. 3 (December 12, 2022): 534–45. https://doi.org/10.1080/08982112.2022.2146510.

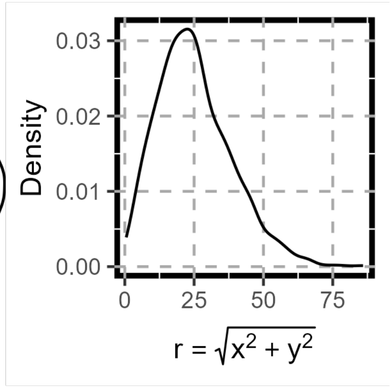
Using the Models

Scenario	Rationale	Formula
Operations ("Real Life")	Most complicated – full model	$79 - 6B + 4D - 7.5F - 5.5B * D + 4.5D * F + 5A * F + 4D^2 - 9F^2$
Operational Testing (OT)	Less fidelity than operations – drop quadratic terms	$79 - 6B + 4D - 7.5F - 5.5B * D + 4.5D * F + 5A * F$
Developmental Testing (DT)	Drop Quadrant Elevation (A)	$79 - 6B + 4D - 7.5F - 5.5B * D + 4.5D * F$
Modeling & Simulation (M&S)	Drop Radar to Weapon Range (F)	$79 - 6B + 4D - 5.5B * D$

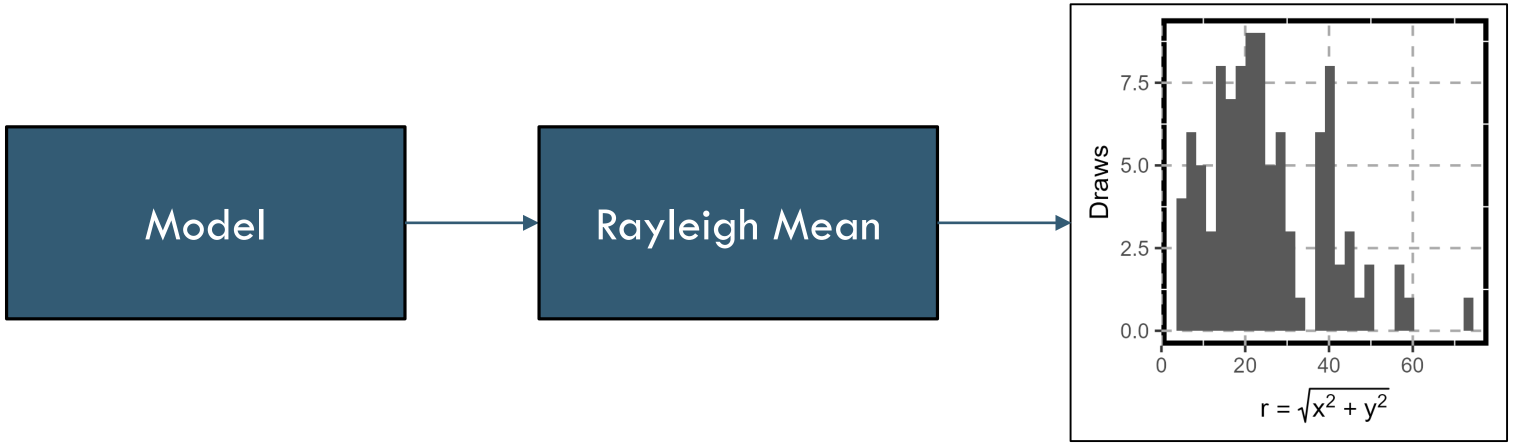
Generating Data

Assume location error is normally distributed in two dimensions: Rayleigh distribution





Models give distribution mean, which can then be used to generate data



Capabilities

Can compare:

- Analysis methods
- Design of experiments techniques (test designs via skpr package)

For problems with:

- Different numbers of test phases/data sources
- Varying data sizes, e.g., trials and reps by phase
- Evolving test factors
- Shifts/biases in test data (e.g., in M&S data)
- Different error/noise in measurements

Results

Comparing Methods

Five methods considered:

- Frequentist:
  - Using OT data only
  - All data, Blocking: With shift factors added for M&S and DT
  - All data, Without blocking: No shift factors for M&S and DT
- Bayesian informative priors w/ downweighting:
  - Resetting intercept uncertainty to prior value
  - Doubling intercept uncertainty

Key metric: RMSE between Rayleigh means

- Fitted model vs. Operational model
- Computed on full factorial dataset generated using the Operational model

Comparing Methods

Five methods considered:

- Frequentist:
  - Using OT data only
  - All data, Blocking: With shift factors added for M&S and DT
  - All data, Without blocking: No shift factors for M&S and DT
- Bayesian informative priors w/ downweighting:
  - Resetting intercept uncertainty to prior value
  - Doubling intercept uncertainty

Key metric: RMSE between Rayleigh means

- Fitted model vs. Operational model
- Computed on full factorial dataset generated using the Operational model

Benefit of working with synthetic data: We know the "truth"!

Example 1: Limited OT

Consider scenarios where OT is limited and M&S is quite a bit larger than DT

Intuition:

- Integrated testing should provide a benefit
- Challenge of managing different data sizes and changing test factors

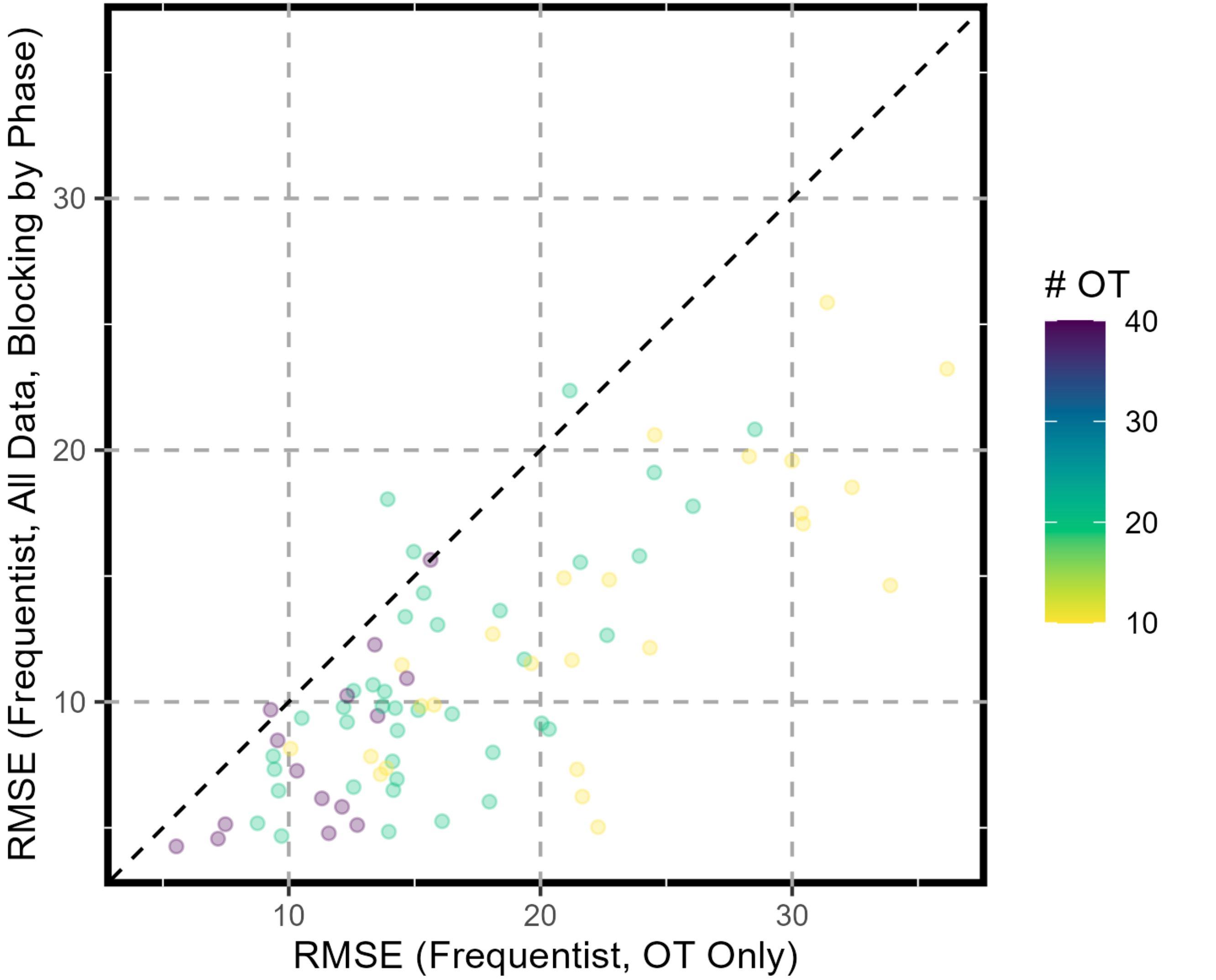
Parameter	Value
M&S Trials	Full factorial (9 trials)
M&S Reps	100
DT Trials	10, 20, 40
DT Reps	5, 10
OT Trials	10, 20
OT Reps	1, 2
DT Optimality	D
OT Optimality	D

(Additional assumption: No more than 200 DT datapoints.)

Example 1: Integrated vs. Single Phase

Error is lower for integrated model with blocking than for OT-only model

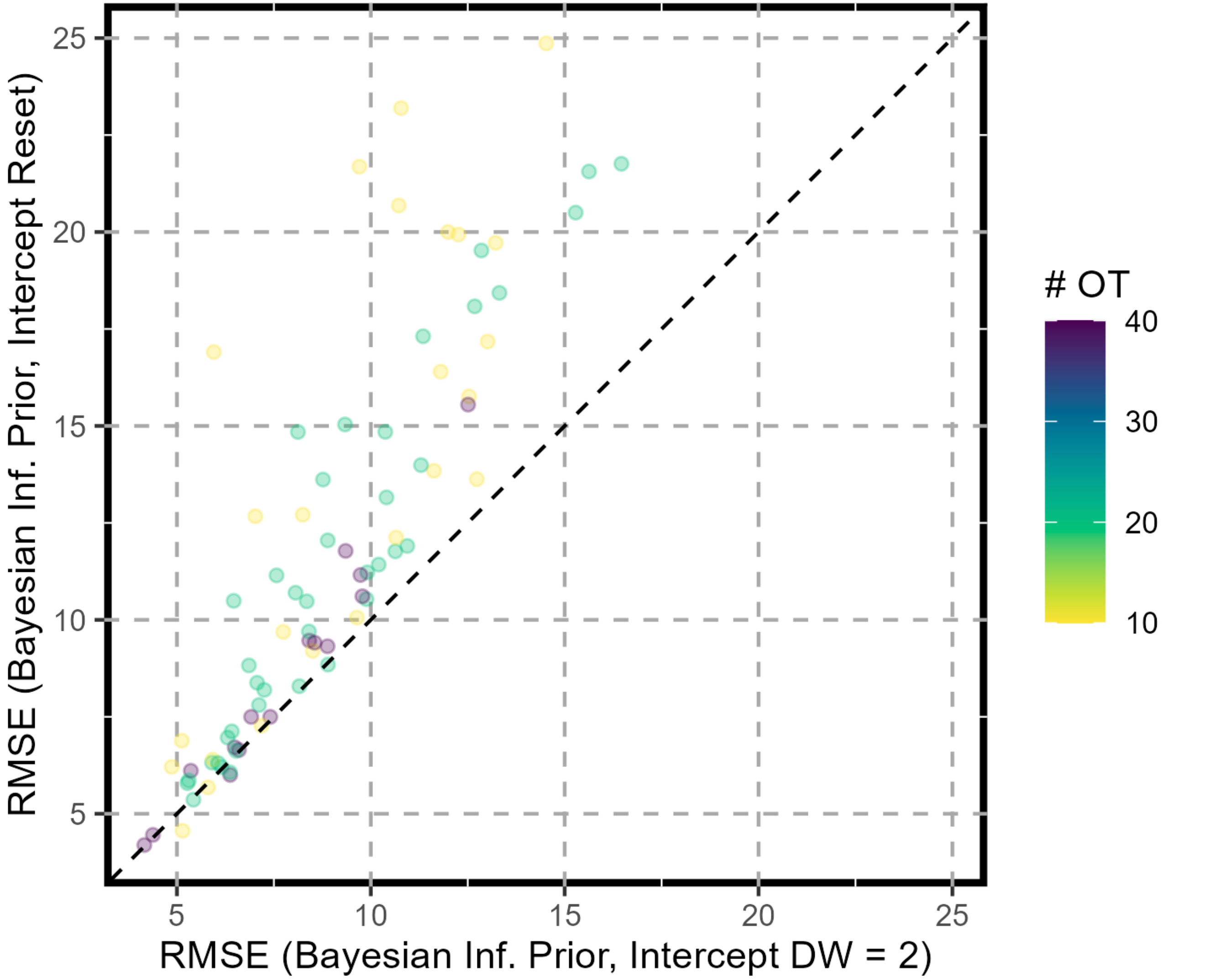
Takeaway: Benefit to integrating information



Example 1: Bayesian

Error is lower with less aggressive downweighting

Takeaway: Benefit to more aggressively integrating information



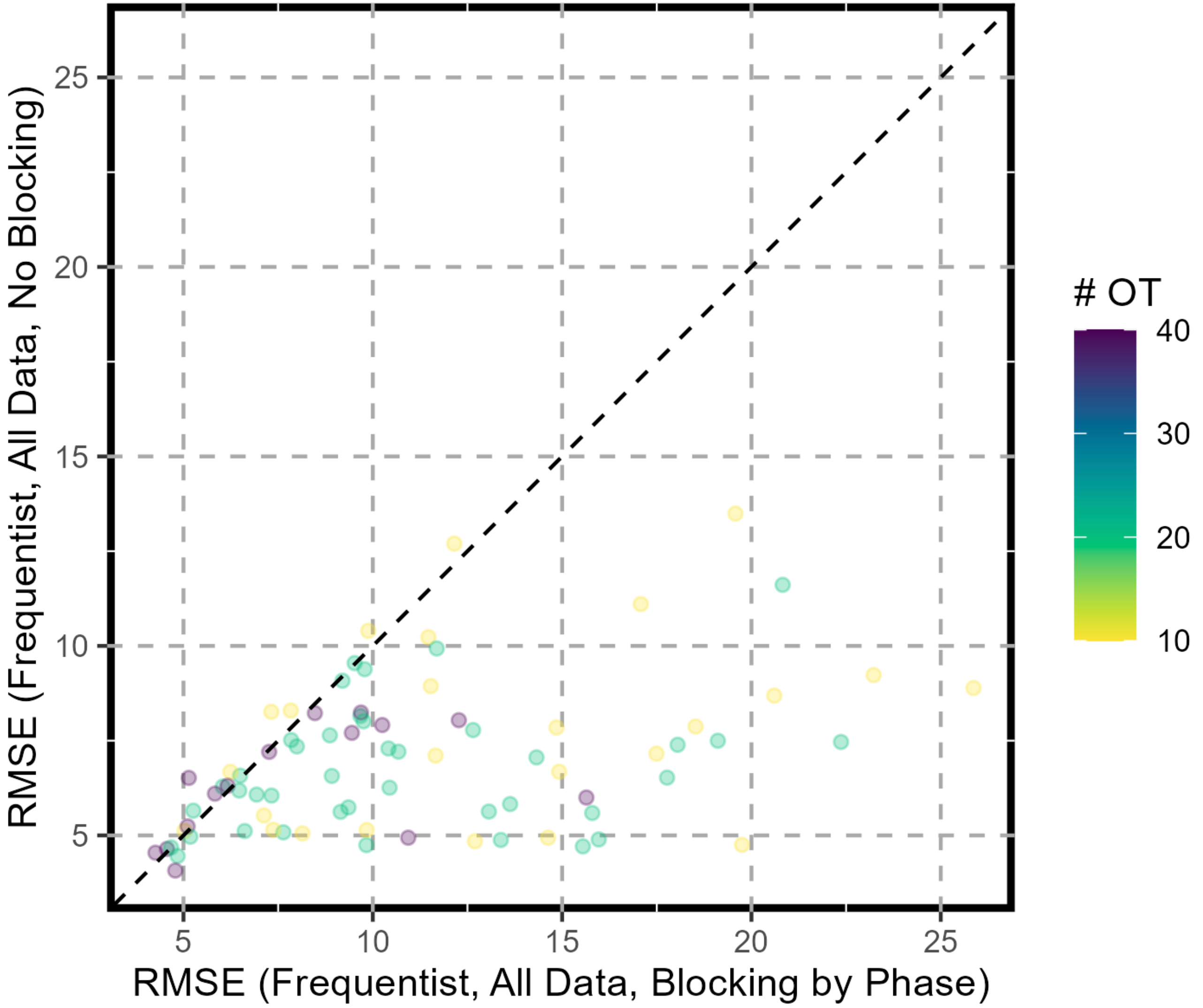
Example 1: Blocking

ACQUISITION INNOVATION RESEARCH CENTER

Error is lower for integrated model *without* blocking

Takeaways:

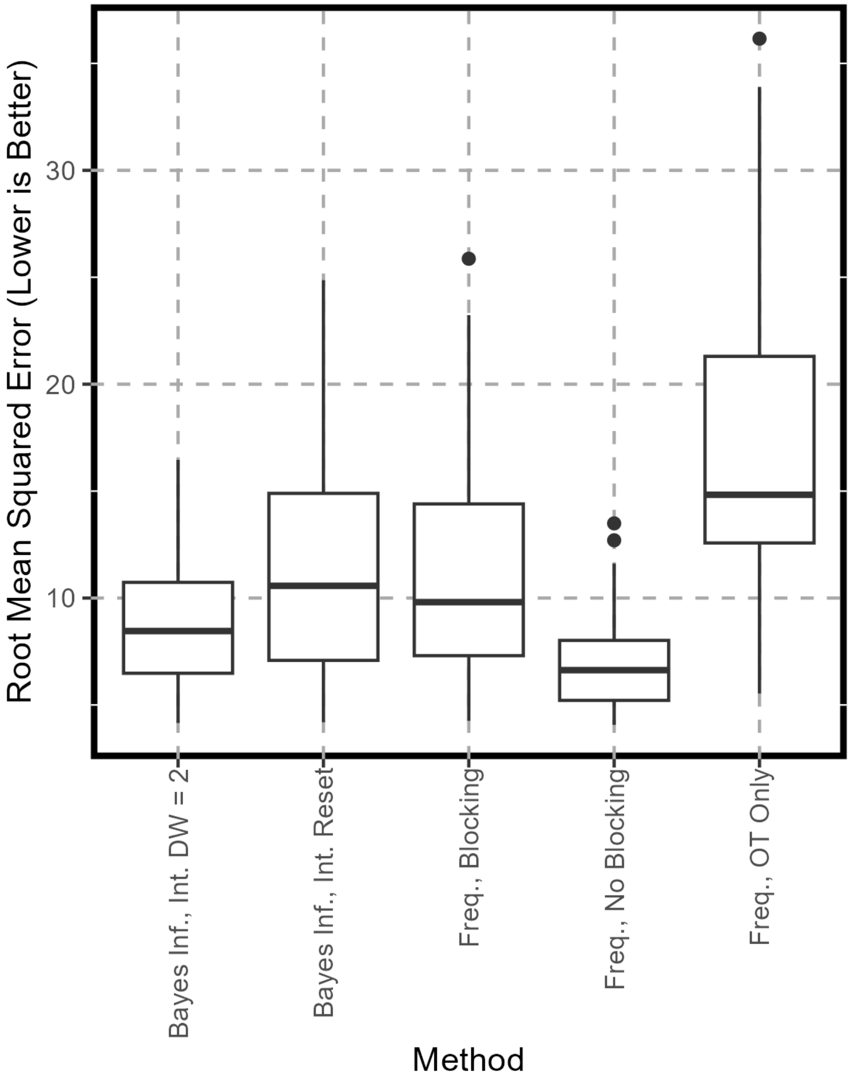
- Blocking shifts just fits bias due to noise in small datasets
- Blocking is bad?



Example 1: Summary

Takeaways:

- Integrated testing helps provide better models
- Without blocking seems to do a little better



Example 2: Limited OT with M&S, DT Shifts

ACQUISITION INNOVATION RESEARCH CENTER

Consider scenarios where OT is limited and M&S is quite a bit larger than DT

- Add random bias to M&S, DT model intercepts

	M&S	DT	OT
Intercept	99.6	86.0	79

Intuition:

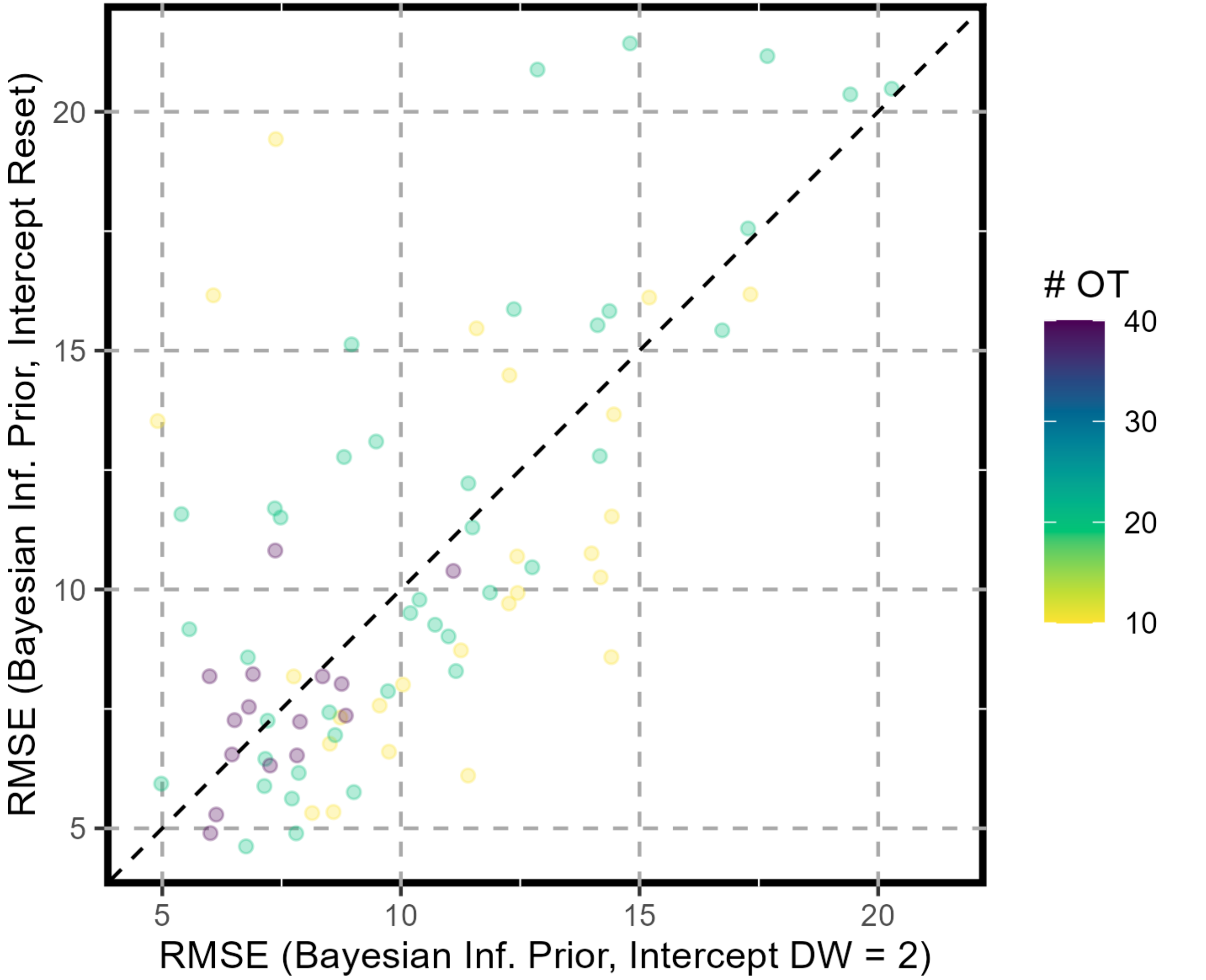
- Integrated testing should provide a benefit
- Biases in some of the data might change results?

Parameter	Value
M&S Trials	Full factorial (9 trials)
M&S Reps	100
DT Trials	10, 20, 40
DT Reps	5, 10
OT Trials	10, 20
OT Reps	1, 2
DT Optimality	D
OT Optimality	D

(Additional assumption: No more than 200 DT datapoints.)

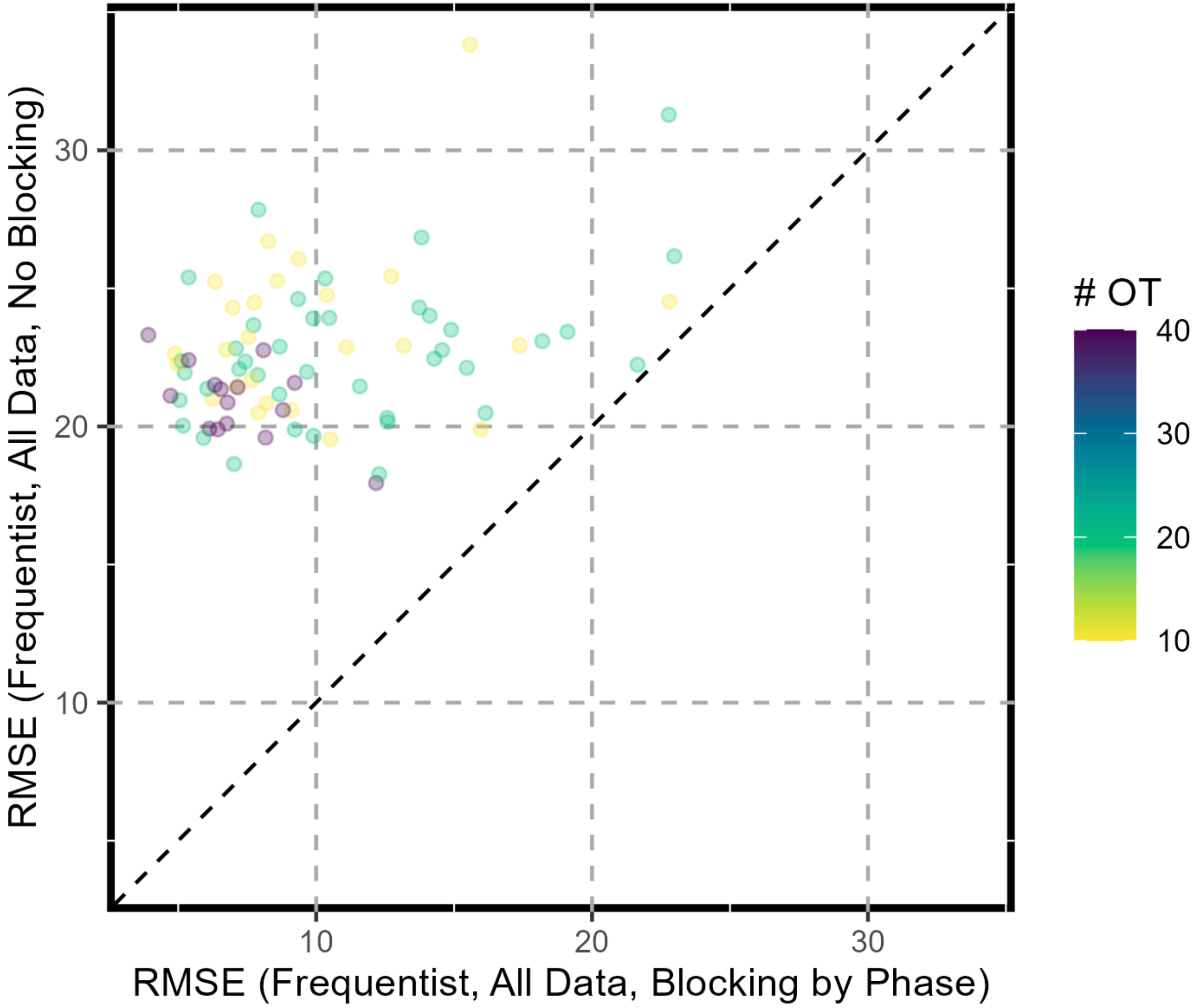
Example 2: Bayesian

Effect of downweighting is less clear



Example 2: Blocking

Error is dramatically lower for integrated model *with* blocking

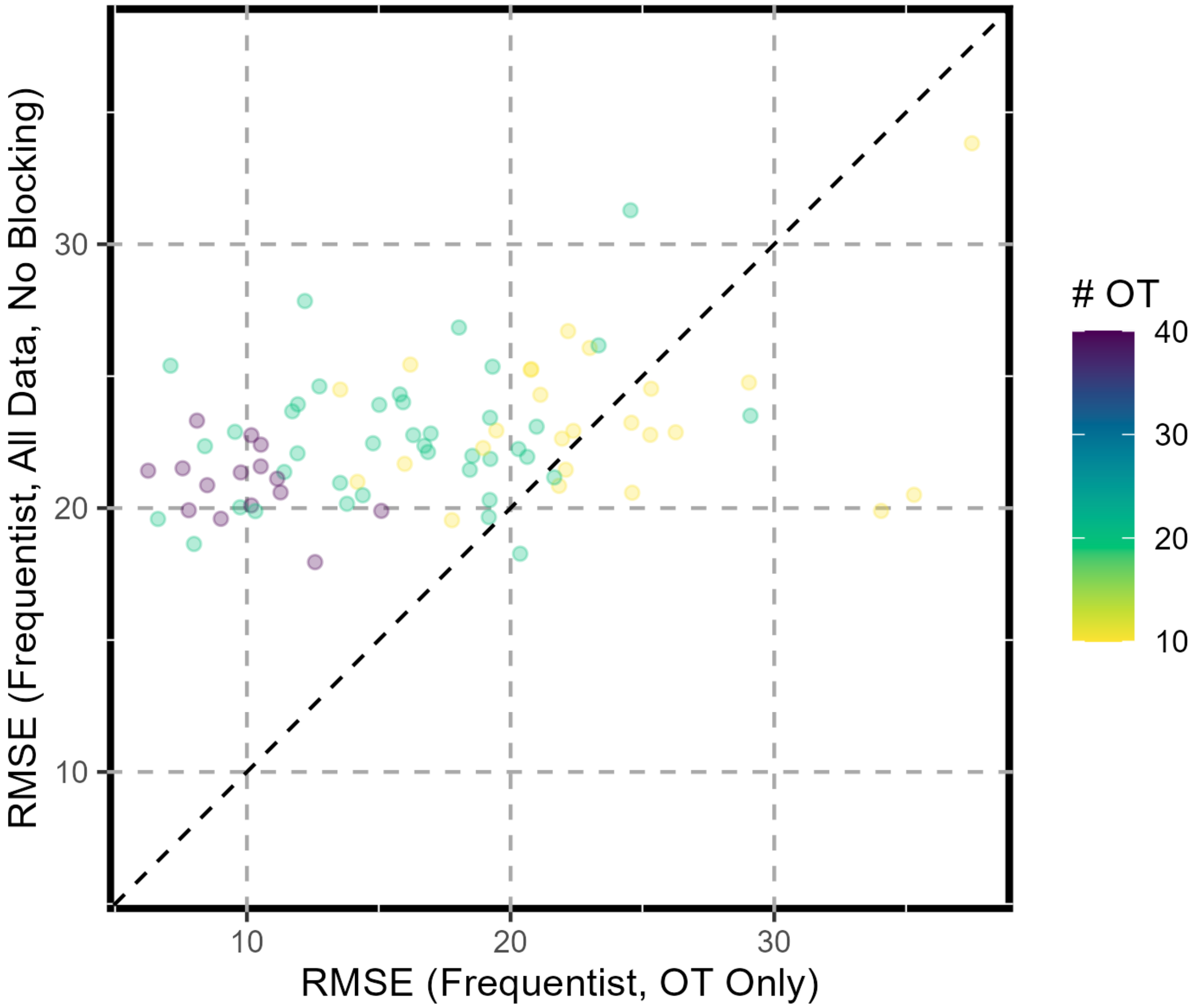


Takeaways:

- Blocking allows accounting for differences in data sources
- Blocking is good?

Example 2: Non-Blocking vs. Single-Phase

Non-blocking model actually makes estimates *worse* than single phase model



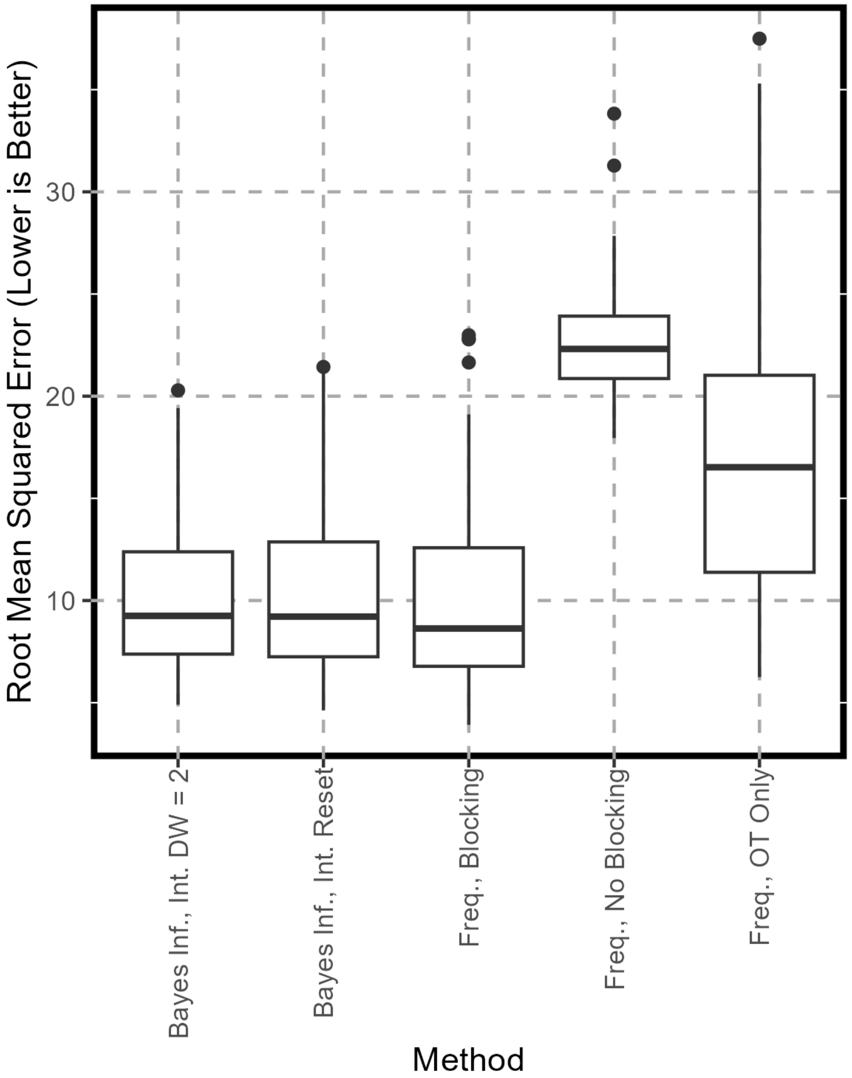
Takeaways:

- Integration of information can make analysis worse if not done carefully

Example 2: Summary

Integration without blocking makes estimates

- Worse than using OT only
- Much worse than other integration methods

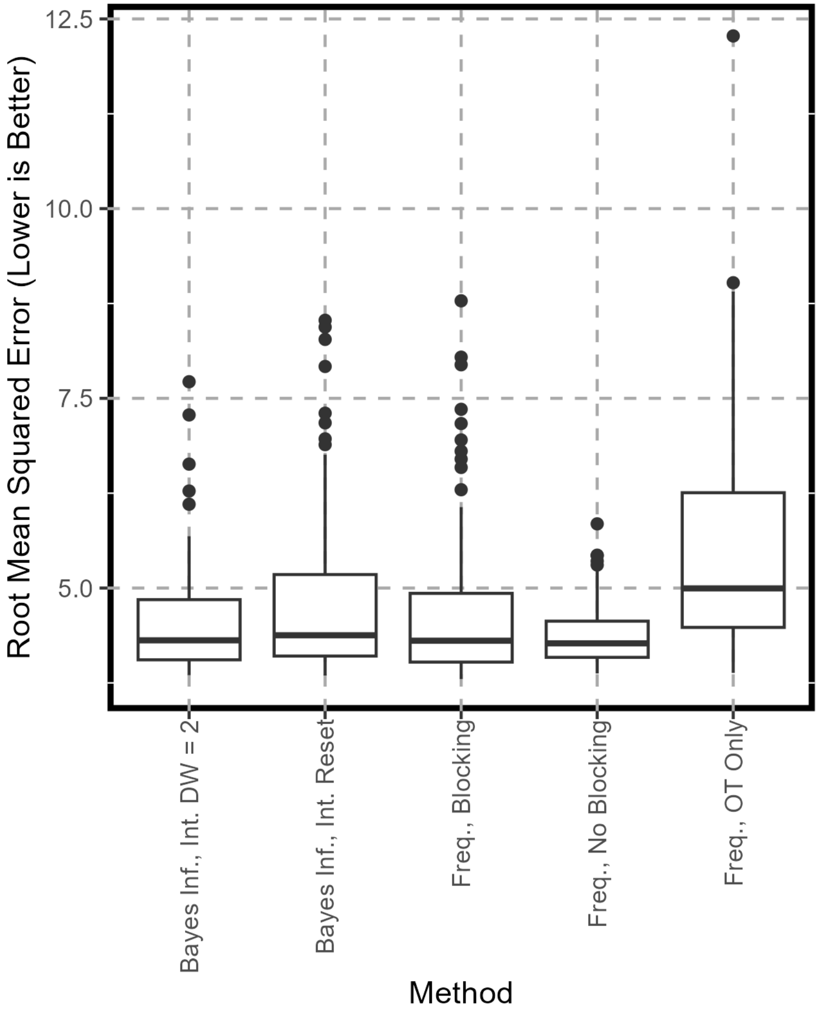


Takeaways:

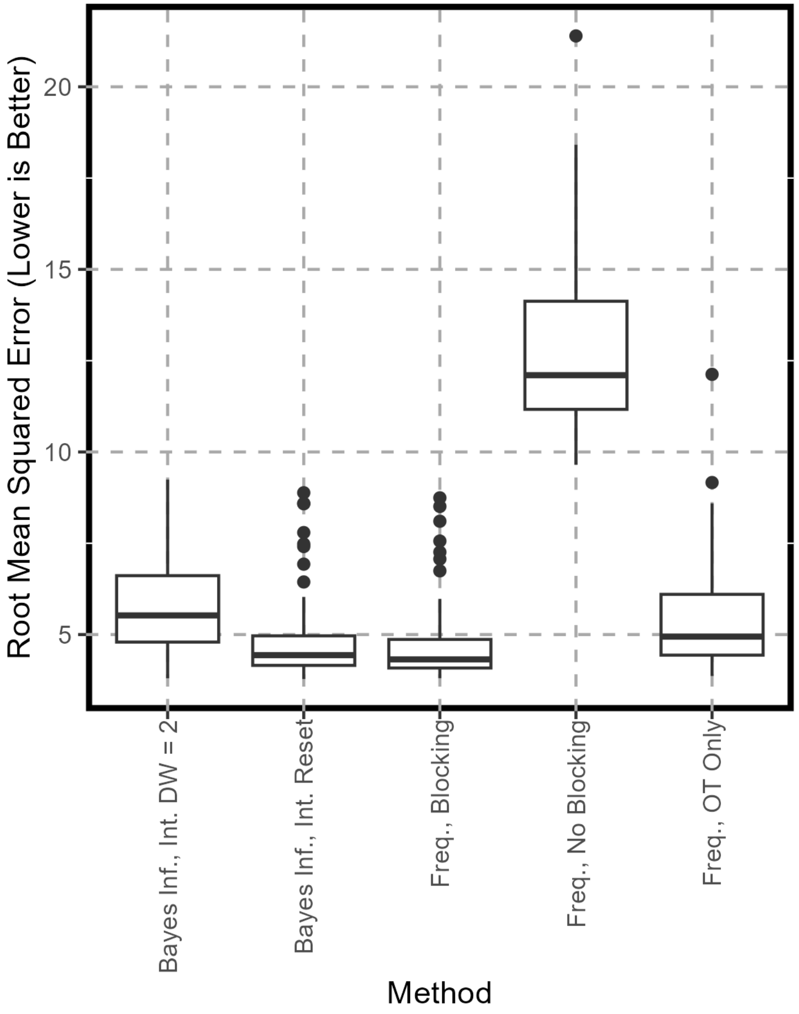
- Integrated testing *mostly* helps provide better models
- **But** some care must be taken in how the data is integrated

Example 3: Large OT

Unbiased M&S/DT



Biased M&S/DT



Conclusions

Conclusions

Testbed for evaluating methods for integrating T&E data, based on DoD-like model

Can mimic many challenges DoD programs face

- Different numbers of test phases/data sources
- Varying data sizes, e.g., trials and reps by phase
- Evolving test factors
- Shifts/biases in test data (e.g., in M&S data)
- Different error/noise in measurements

Results illustrate both the promise of integrating information but also some potential pitfalls

- Note: Importance of having enough information about data collection to integrate

Future Work

Expand to

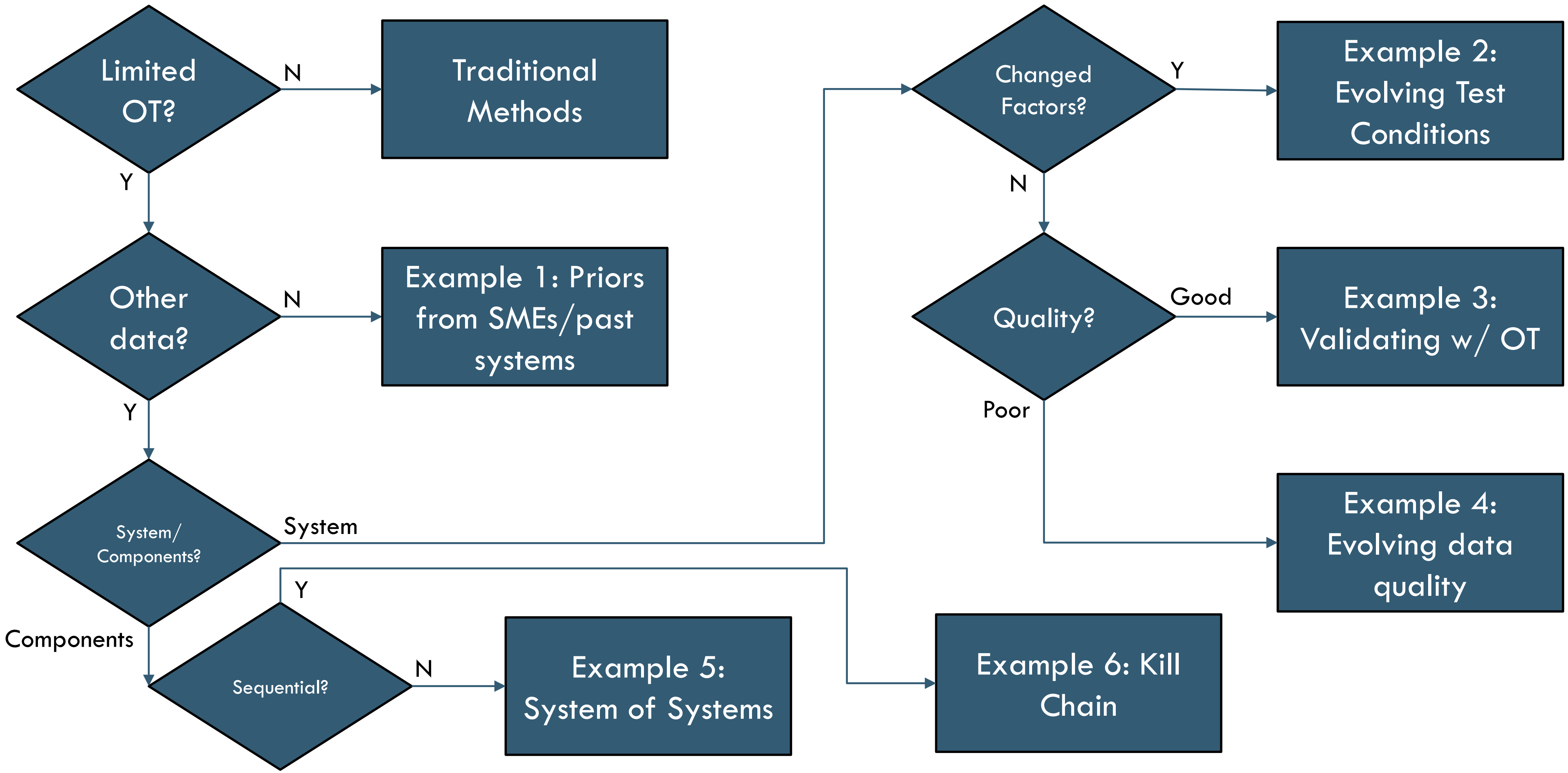
- Other points in "consideration-space"
- Other DoD-inspired models

Compare:

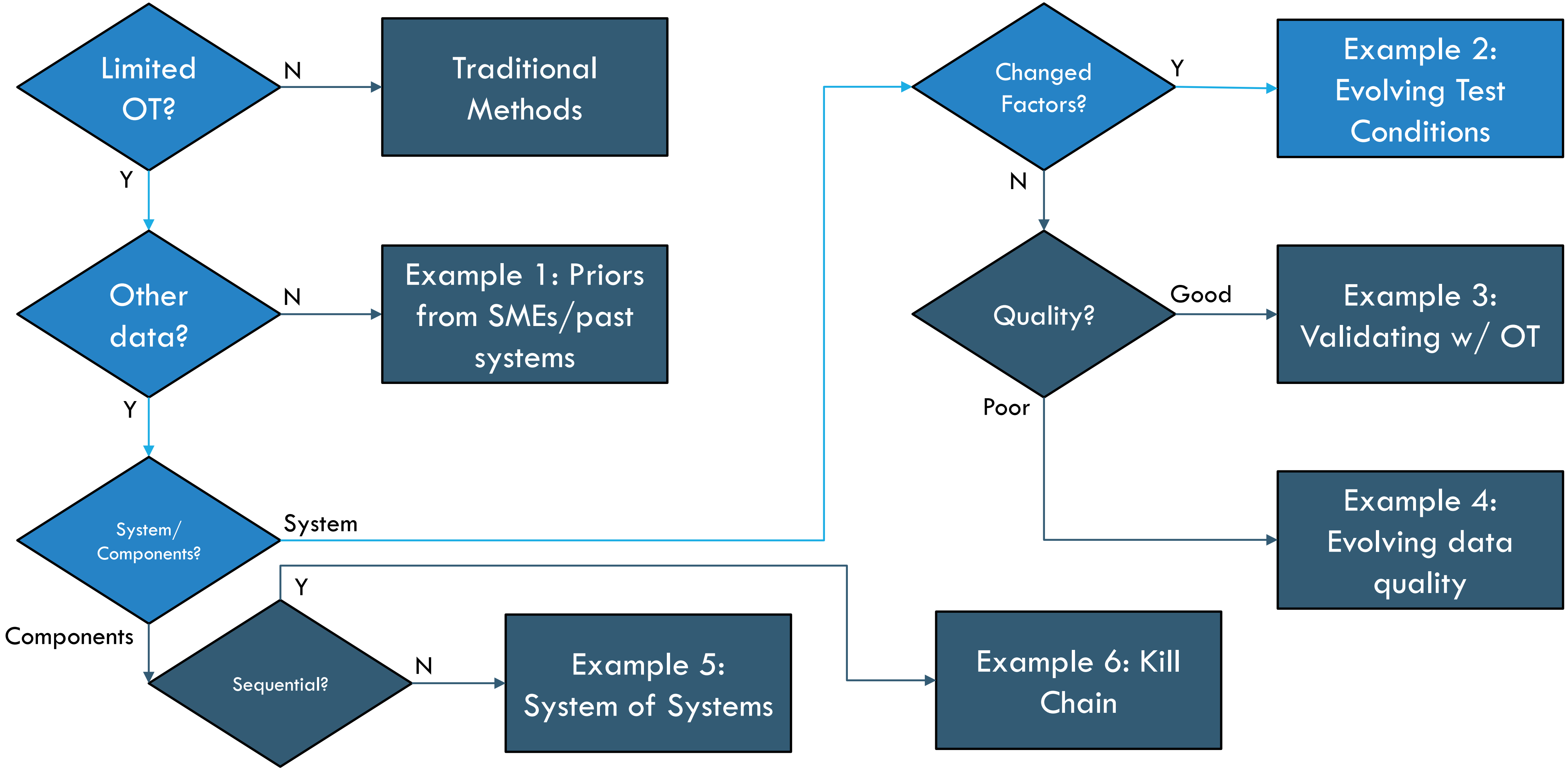
- Analysis methods
- Design of experiments techniques
  - Test designs via `skpr` package

Consideration
Safety
Cost
Resource Availability
Schedule
Historical operational performance data
Modeling and Simulation
Scale

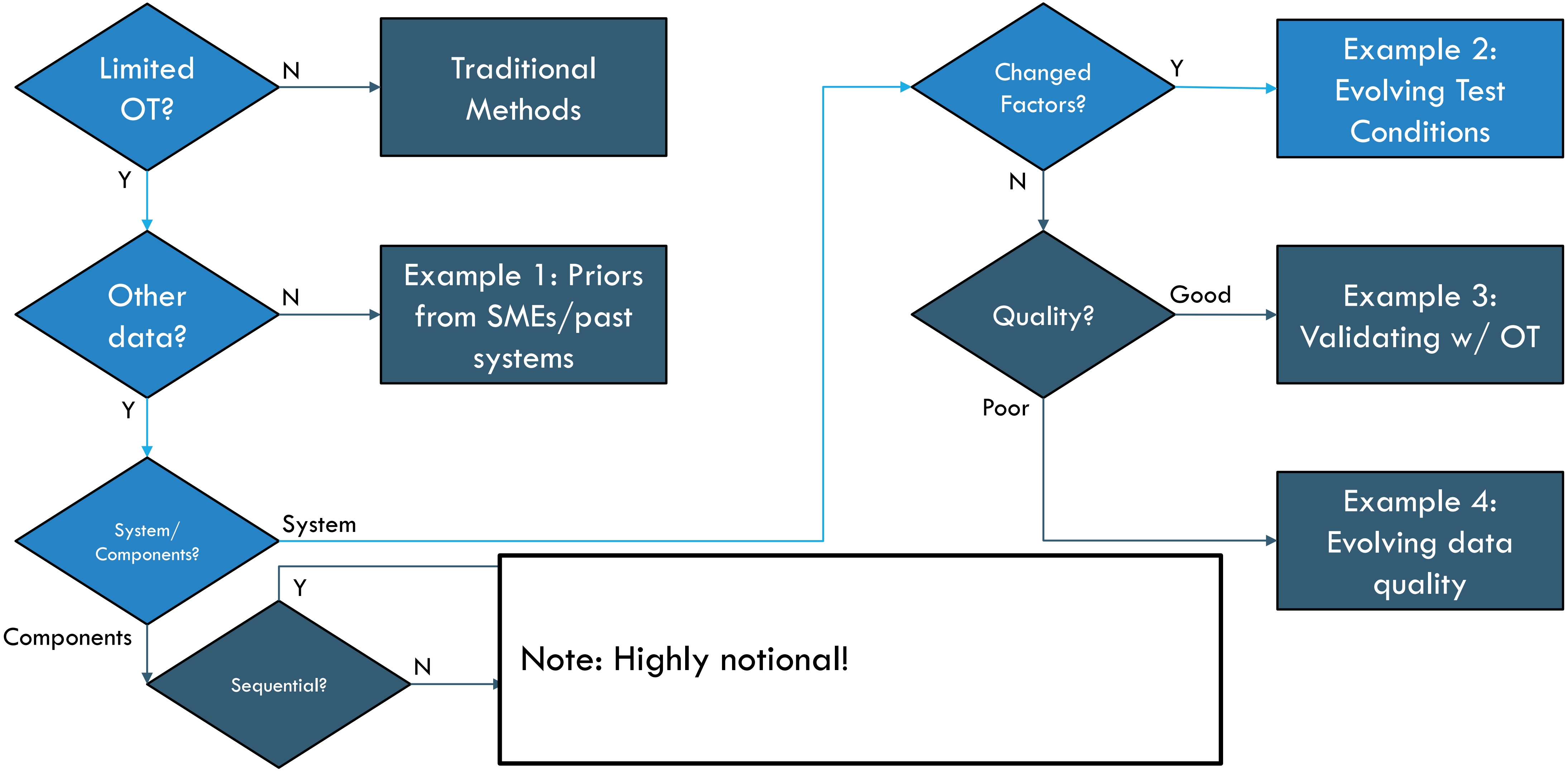
Long-Term Goal: Decision Tree for Best Practices



Long-Term Goal: Decision Tree for Best Practices



Long-Term Goal: Decision Tree for Best Practices



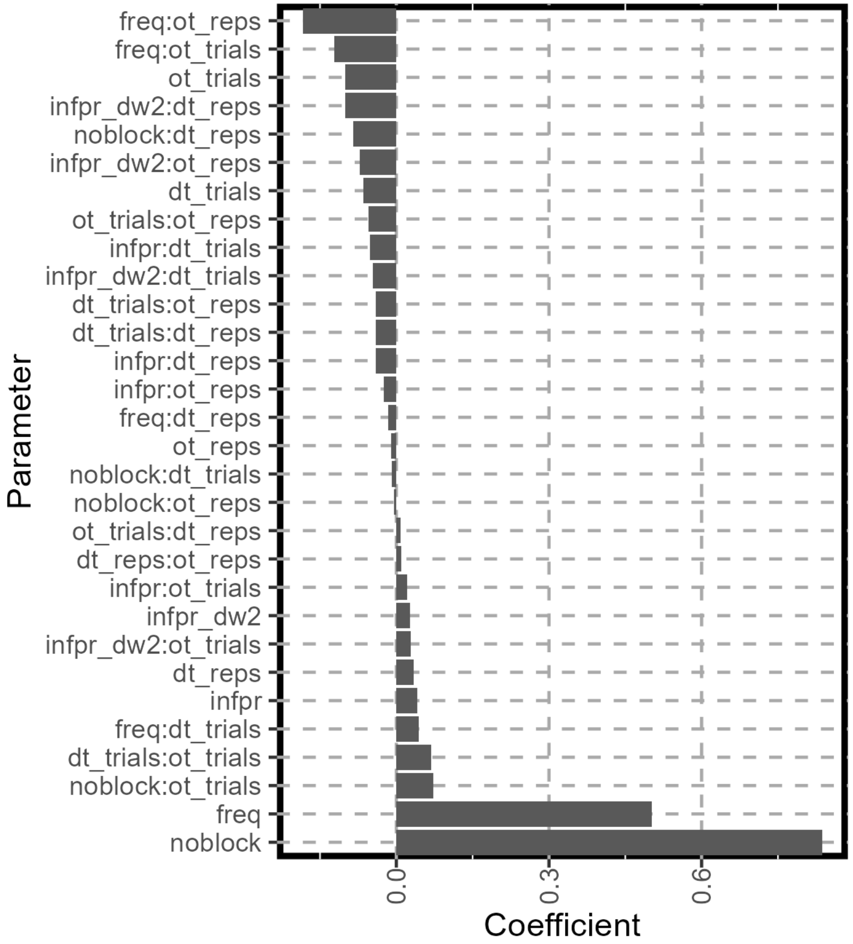
Note: Highly notional!

Questions?

Justin Krometis

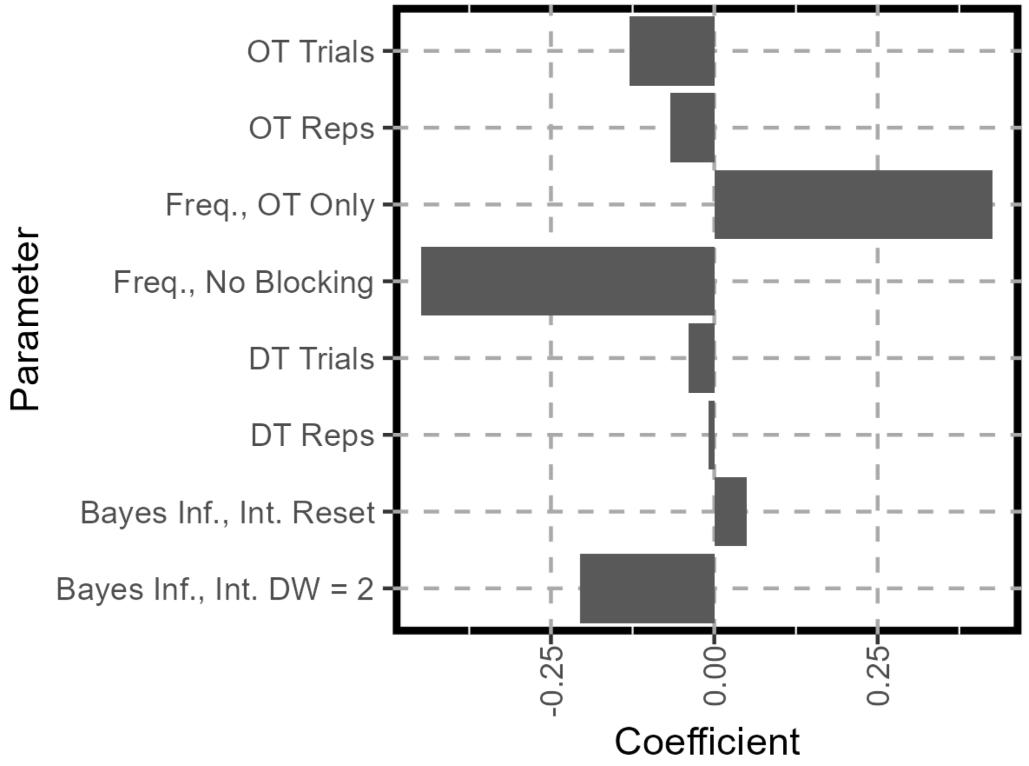
Virginia Tech National Security Institute

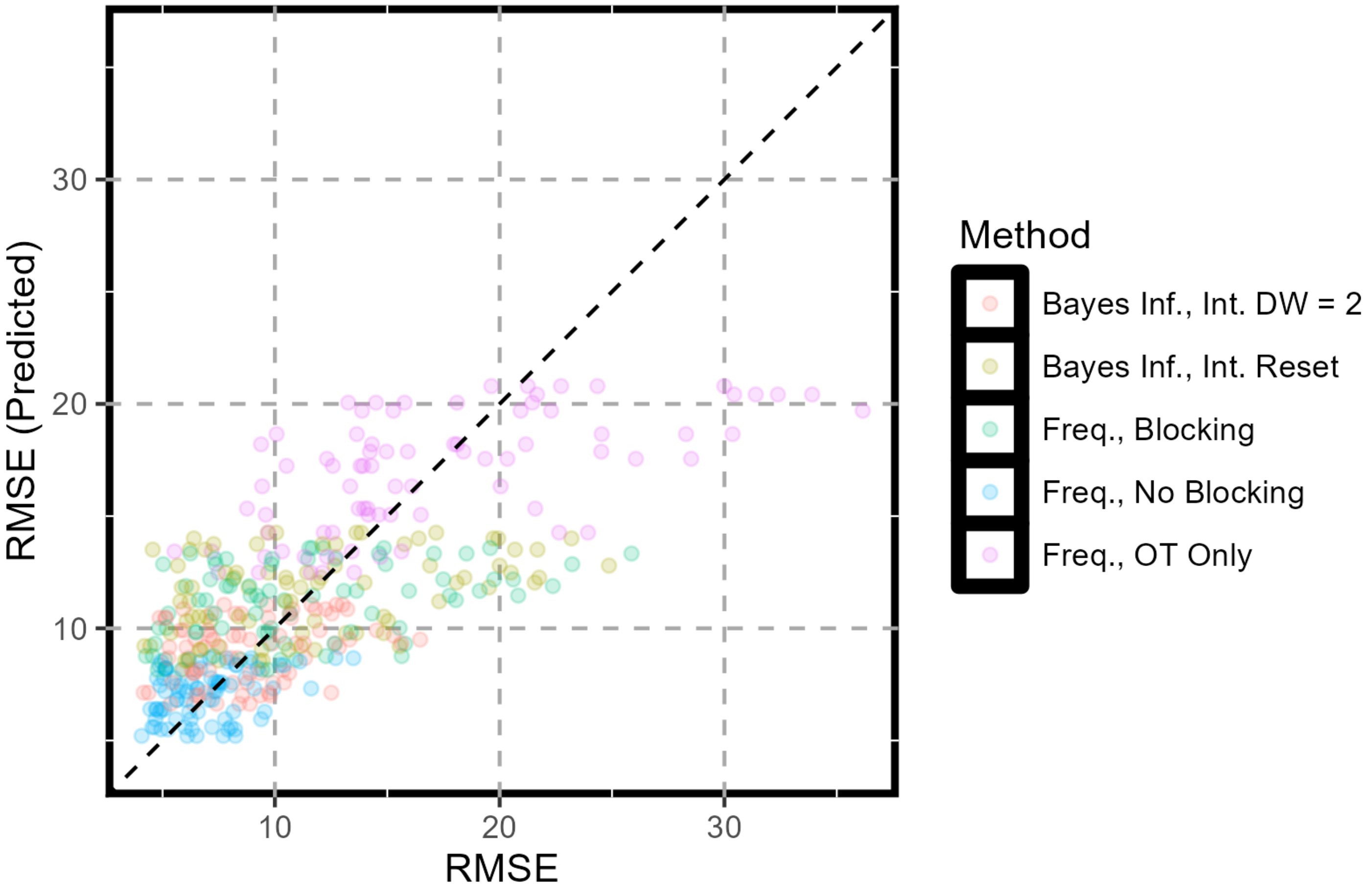
jkrometis@vt.edu



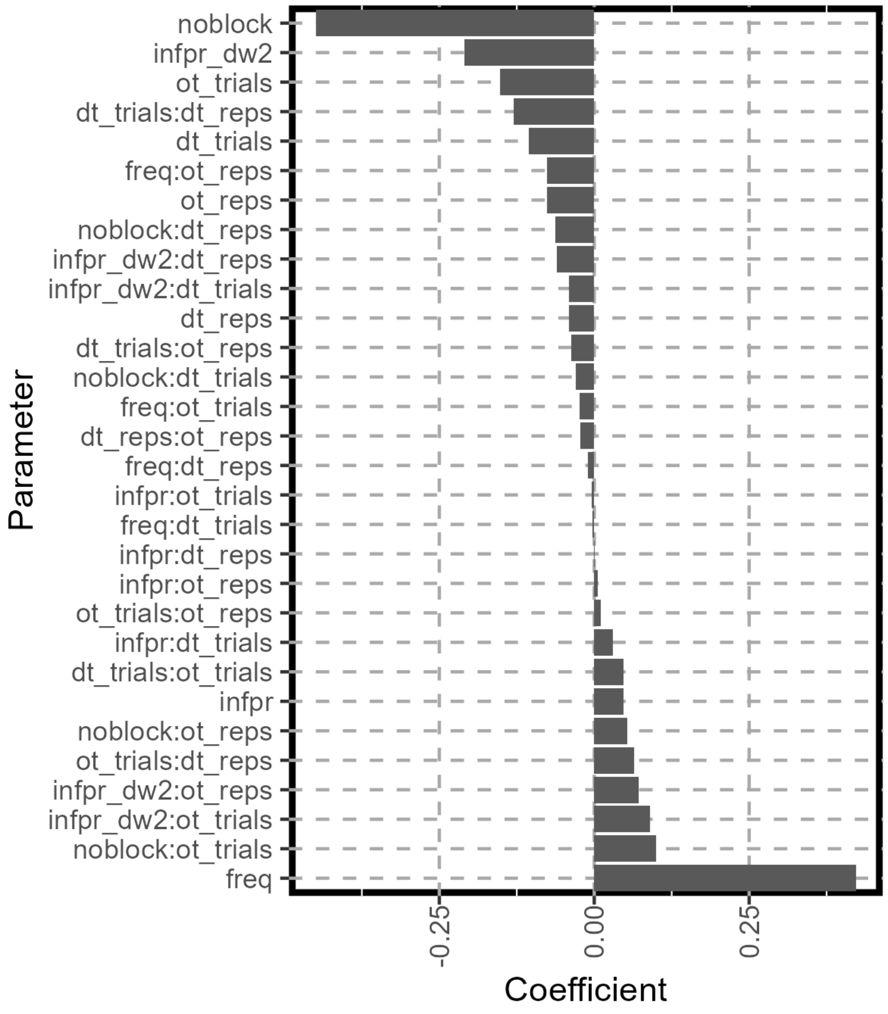
Backup Slides

Example 1: Factors Driving RMSE

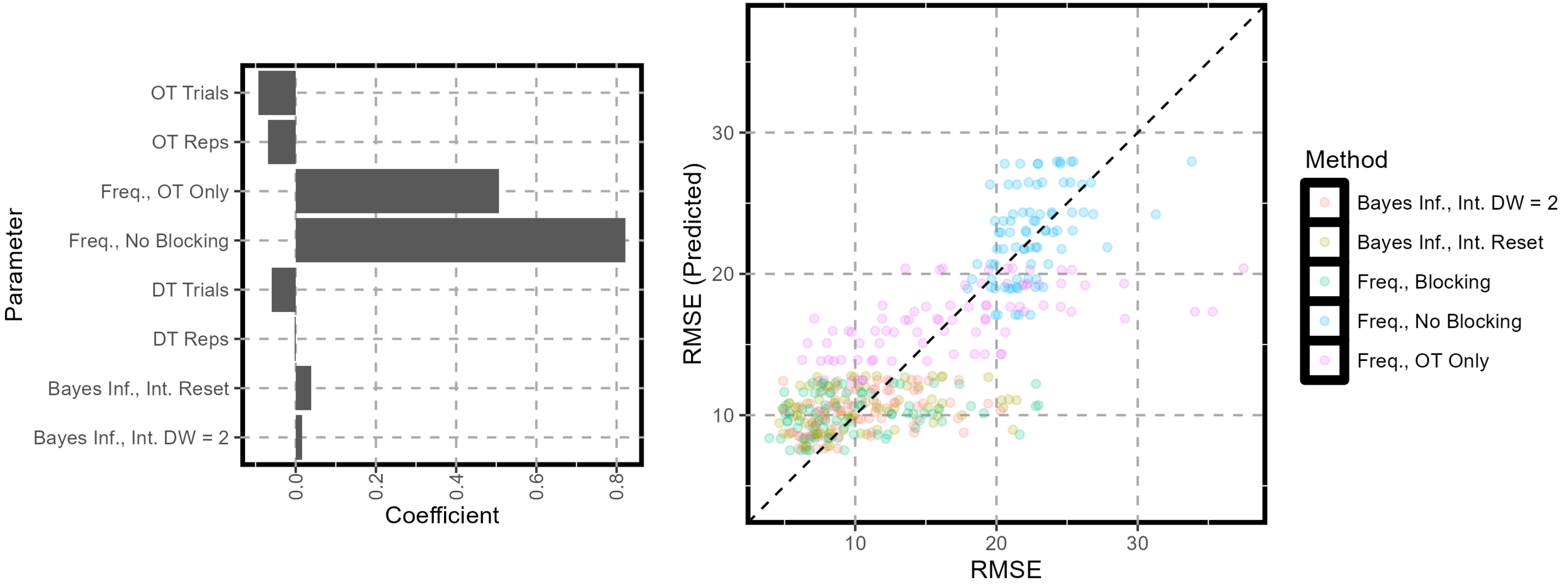




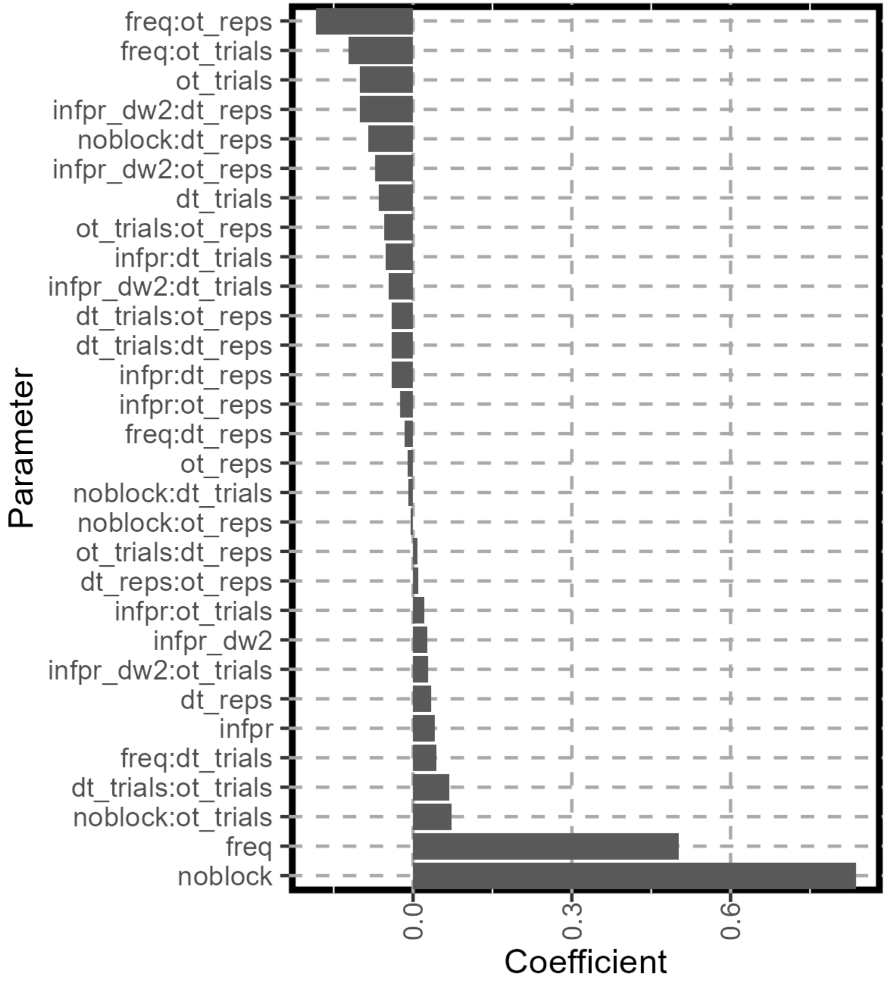
Example 1: Interactions



Example 2: Factors Driving RMSE

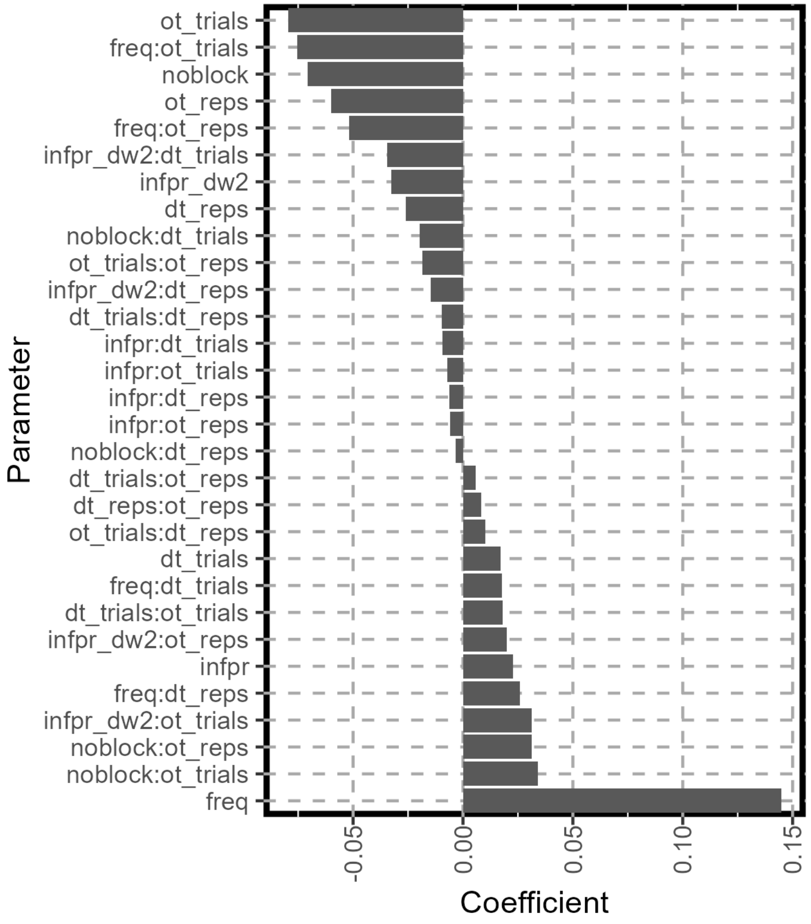


Example 2: Interactions



Example 3: Interactions

Unbiased M&S/DT



Biased M&S/DT