# Modeling paired-choice data to effectively predict human evaluations of individual performance

Josh Fiechter
Cognitive Data Scientist
Kairos Research



KAIROSRESEARCH

# General setup

- Human evaluators provide informed assessments of individuals that can be leveraged for picking optimal people for certain tasks and/or roles

# General setup

- Human evaluators provide informed assessments of individuals that can be leveraged for picking optimal people for certain tasks and/or roles
- But:
  - These evaluations are costly to implement
  - Human raters’ criteria might not be consistently enforced

# General setup

- Human evaluators provide informed assessments of individuals that can be leveraged for picking optimal people for certain tasks and/or roles
- But:
  - These evaluations are costly to implement
  - Human raters’ criteria might not be consistently enforced
- A reliable model of human evaluators would allow us to:
  - Assess individuals with greater speed and consistency
  - Minimize the burden on human raters

# The data

- **81 operators** spread across **8 nationwide sites**
- **76** attributes to serve as predictors
  - Demographic factors plus physical, intellectual, and personality traits
- **3 scenarios** on which operators are evaluated
- **3–4 evaluators** nested within each site
  - Evaluations are in the form of **paired choices**
  - E.g., Should <u>Bob</u> or <u>Tom</u> take part in this task?
- **3771 total choices**

# The data

| Site | Operator1 | Operator2 | win1 | win2 | Scenario | Evaluator | Age |
|---|---|---|---|---|---|---|---|
| Site A | 162 | 180 | 1 | 0 | 1 | 1 | 1 |
| Site A | 216 | 162 | 1 | 0 | 1 | 1 | 13 |
| Site A | 125 | 216 | 0 | 1 | 1 | 1 | 11 |
| Site A | 102 | 180 | 1 | 0 | 1 | 1 | 1 |
| Site A | 102 | 87 | 0 | 1 | 1 | 1 | -14 |

▪ ▪ ▪ *76 differential features*

⋮

*3771 rows*



KAIROSRESEARCH

# The data

| Site | Operator1 | Operator2 | win1 | win2 | Scenario | Evaluator | Age |
|---|---|---|---|---|---|---|---|
| Site A | 162 | 180 | 1 | 0 | 1 | 1 | 1 |
| Site A | 216 | 162 | 1 | 0 | 1 | 1 | 13 |
| Site A | 125 | 216 | 0 | 1 | 1 | 1 | 11 |
| Site A | 102 | 180 | 1 | 0 | 1 | 1 | 1 |
| Site A | 102 | 87 | 0 | 1 | 1 | 1 | -14 |

*76 differential features*

**Operator1** and **Operator2** are the two people being compared on a given trial

*3771 rows*

# The data

| Site | Operator1 | Operator2 | win1 | win2 | Scenario | Evaluator | Age |
|---|---|---|---|---|---|---|---|
| Site A | **162** | 180 | 1 | 0 | 1 | 1 | 1 |
| Site A | **216** | **162** | 1 | 0 | 1 | 1 | 13 |
| Site A | 125 | **216** | 0 | 1 | 1 | 1 | 11 |
| Site A | 102 | 180 | 1 | 0 | 1 | 1 | 1 |
| Site A | 102 | 87 | 0 | 1 | 1 | 1 | -14 |

▪ ▪ ▪ *76 differential features*

Note how individuals move between these two columns

⋮

*3771 rows*



KAIROSRESEARCH

# The data

**win1** is our outcome of interest

| Site | Operator1 | Operator2 | win1 | win2 | Scenario | Evaluator | Age |
|---|---|---|---|---|---|---|---|
| Site A | 162 | 180 | 1 | 0 | 1 | 1 | 1 |
| Site A | 216 | 162 | 1 | 0 | 1 | 1 | 13 |
| Site A | 125 | 216 | 0 | 1 | 1 | 1 | 11 |
| Site A | 102 | 180 | 1 | 0 | 1 | 1 | 1 |
| Site A | 102 | 87 | 0 | 1 | 1 | 1 | -14 |

... *76 differential features*

⋮

*3771 rows*

# The data

Data contain 76 **differential features** (e.g., $Age_{Operator1} - Age_{Operator2}$)

| Site | Operator1 | Operator2 | win1 | win2 | Scenario | Evaluator | Age |
|---|---|---|---|---|---|---|---|
| Site A | 162 | 180 | 1 | 0 | 1 | 1 | 1 |
| Site A | 216 | 162 | 1 | 0 | 1 | 1 | 13 |
| Site A | 125 | 216 | 0 | 1 | 1 | 1 | 11 |
| Site A | 102 | 180 | 1 | 0 | 1 | 1 | 1 |
| Site A | 102 | 87 | 0 | 1 | 1 | 1 | -14 |

. . . *76 differential features*

⋮

*3771 rows*

# Objectives

- Build a model that predicts human evaluations of individuals
- Evaluate which attributes most strongly influence evaluation
- Evaluate the predictive capabilities of the model
  - I.e., cross-validate the model on novel observations

# The model

- We fit a Bradley-Terry-Luce (BTL) model of paired choices

# The model

- We fit a Bradley-Terry-Luce (BTL) model of paired choices
- In its simplest from, the BTL model estimates log latent ability, $\lambda$, for every individual
    - $\text{logit}(P[win1]) \sim Bernoulli(\lambda_i – \lambda_j)$

# The model

- We fit a Bradley-Terry-Luce (BTL) model of paired choices
- In its simplest from, the BTL model estimates log latent ability, $\lambda$, for every individual
  - $\text{logit}(P[win1]) \sim Bernoulli(\lambda_i - \lambda_j)$
- BTL models can also be written as GLMs:
  - $\text{logit}(P[win1]) \sim Bernoulli(\beta_0 + \sum_{n=1}^{N} W_n \beta_n + \sum_{k=1}^{K} X_k \beta_k)$



KAIROSRESEARCH

# The model

- We fit a Bradley-Terry-Luce (BTL) model of paired choices
- In its simplest from, the BTL model estimates log latent ability, $\lambda$, for every individual
  - $\mathrm{logit}(P[win1]) \sim Bernoulli(\lambda_i - \lambda_j)$
- BTL models can also be written as GLMs:
  - $\mathrm{logit}(P[win1]) \sim Bernoulli(\boldsymbol{\beta_0} + \sum_{n=1}^{N} W_n \beta_n + \sum_{k=1}^{K} X_k \beta_k)$

Bias in favor of Operator1

# The model

- We fit a Bradley-Terry-Luce (BTL) model of paired choices
- In its simplest from, the BTL model estimates log latent ability, $\lambda$, for every individual
  - $\text{logit}(P[win1]) \sim Bernoulli(\lambda_i - \lambda_j)$
- BTL models can also be written as GLMs:
  - $\text{logit}(P[win1]) \sim Bernoulli(\beta_0 + \boldsymbol{\sum_{n=1}^{N} W_n \beta_n} + \sum_{k=1}^{K} X_k \beta_k)$

Operators 1 and 2 receive respective weights ($\boldsymbol{W_k}$) of 1 and -1

$\boldsymbol{\beta_n}$ is estimated latent ability



KAIROS RESEARCH

# The model

- We fit a Bradley-Terry-Luce (BTL) model of paired choices
- In its simplest from, the BTL model estimates log latent ability, $\lambda$, for every individual
  - $\text{logit}(P[win1]) \sim Bernoulli(\lambda_i - \lambda_j)$
- BTL models can also be written as GLMs:
  - $\text{logit}(P[win1]) \sim Bernoulli(\beta_0 + \sum_{n=1}^{N} W_n \beta_n + \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{X_k \beta_k})$

Coefficients ($\boldsymbol{\beta_k}$) are estimated for differential values of covariates ($\boldsymbol{X_k}$)



KAIROSRESEARCH

# The model

- $\text{logit}(P[win1]) \sim Bernoulli(\beta_0 + \sum_{n=1}^{N} W_n \beta_n + \sum_{k=1}^{K} X_k \beta_k)$
- We made three noteworthy modifications to the model above:
    1) Incorporated hierarchical model structure



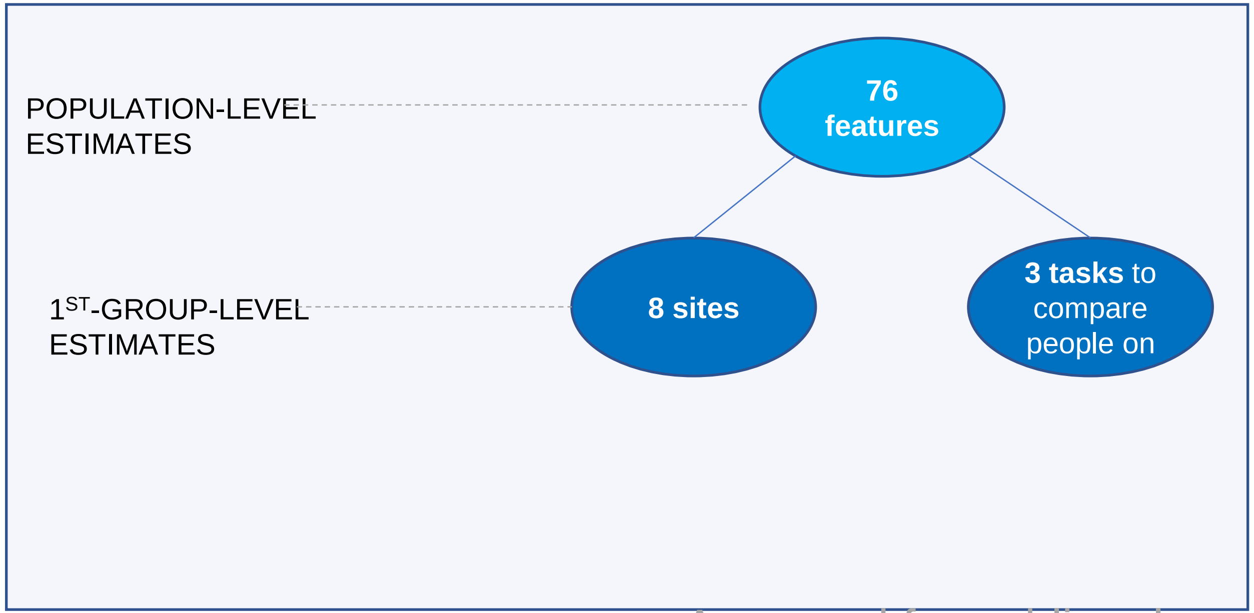
KAIROSRESEARCH

# The model

- $\text{logit}(P[win1]) \sim Bernoulli(\beta_0 + \sum_{n=1}^{N} W_n\beta_n + \sum_{k=1}^{K} X_k\beta_k)$
- We made three noteworthy modifications to the model above:
    1) Incorporated hierarchical model structure

# The model

- $\text{logit}(P[win1]) \sim Bernoulli(\beta_0 + \sum_{n=1}^{N} W_n \beta_n + \sum_{k=1}^{K} X_k \beta_k)$
- We made three noteworthy modifications to the model above:
    1) Incorporated hierarchical model structure

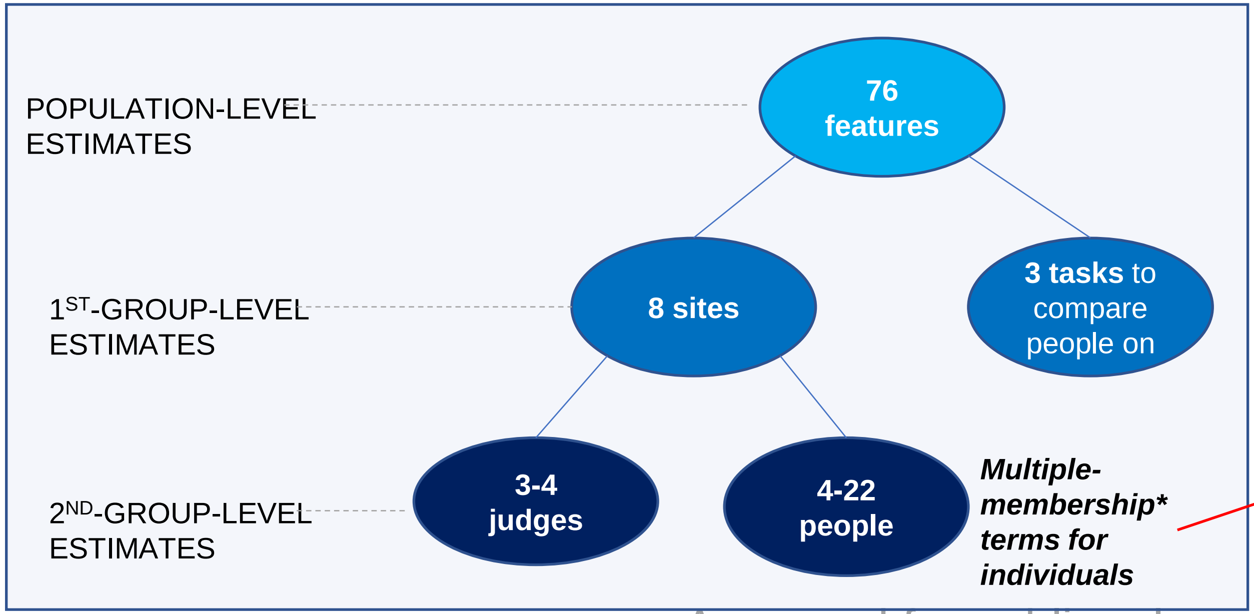





KAIROSRESEARCH

# The model

- $\text{logit}(P[win1]) \sim Bernoulli(\beta_0 + \sum_{n=1}^{N} W_n \beta_n + \sum_{k=1}^{K} X_k \beta_k)$
- We made three noteworthy modifications to the model above:
    1) Incorporated hierarchical model structure




*See Durrant, Vassallo, & Smith, 2018

| Site | Operator1 | Operator2 | win1 |
| --- | --- | --- | --- |
| Site A | 162 | 180 | 1 |
| Site A | 216 | 162 | 1 |
| Site A | 125 | 216 | 0 |
| Site A | 102 | 180 | 1 |
| Site A | 102 | 87 | 0 |

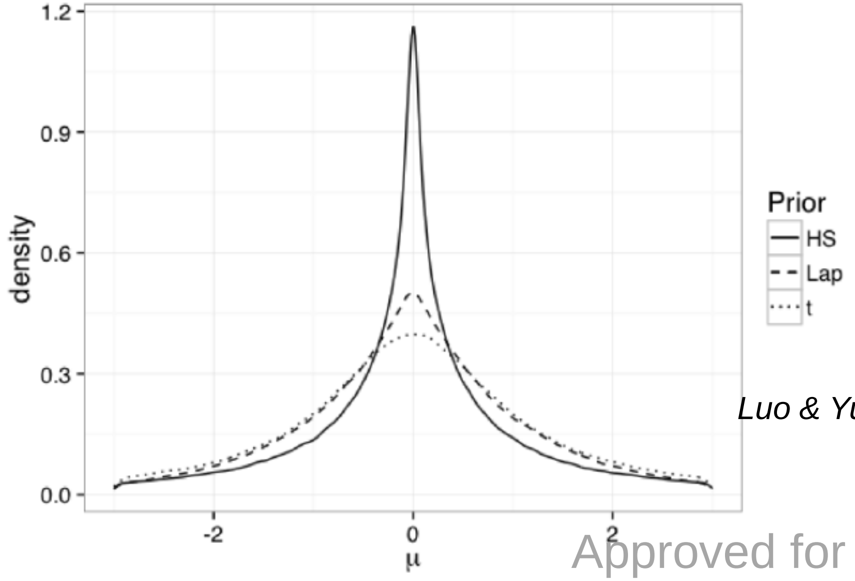
⋮

3771 r

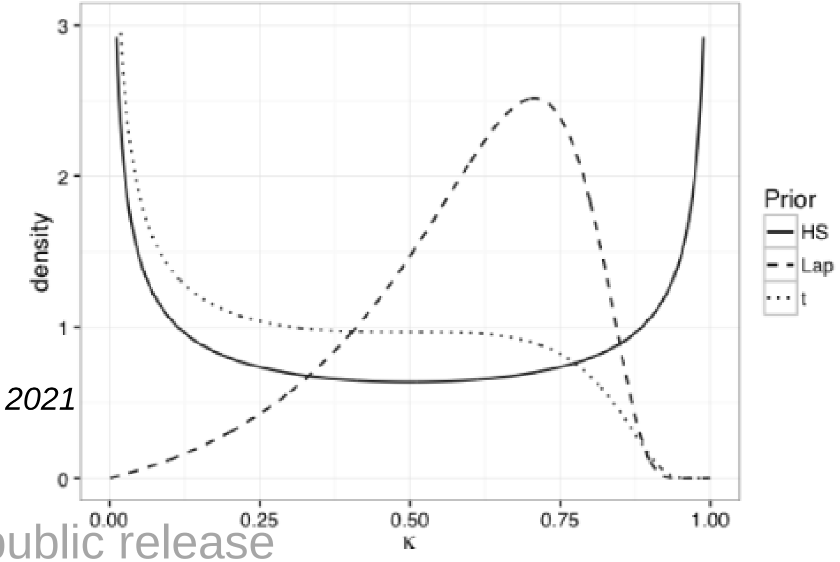
Note how individuals move between these two columns

# The model

- $\text{logit}(P[win1]) \sim Bernoulli(\beta_0 + \sum_{n=1}^{N} W_n\beta_n + \sum_{k=1}^{K} X_k\beta_k)$
- We made three noteworthy modifications to the model above:
    1) Incorporated hierarchical model structure
    2) Used *horseshoe priors* (Carvalho et al., 2010) for regularized estimates
        - Estimation conducted via MCMC sampling in Stan (Carpenter et al., 2017)




Luo & Yu, 2021



# The model

- $\text{logit}(P[win1]) \sim Bernoulli(\beta_0 + \sum_{n=1}^{N} W_n\beta_n + \sum_{k=1}^{K} X_k\beta_k)$
- We made three noteworthy modifications to the model above:
  1) Incorporated hierarchical model structure
  2) Used *horseshoe priors* (Carvalho et al., 2010) for regularized estimates
     - Estimation conducted via MCMC sampling in Stan (Carpenter et al., 2017)
  3) All inputs $X_k$ were *z*-transformed to remove any artifacts from differences in scale

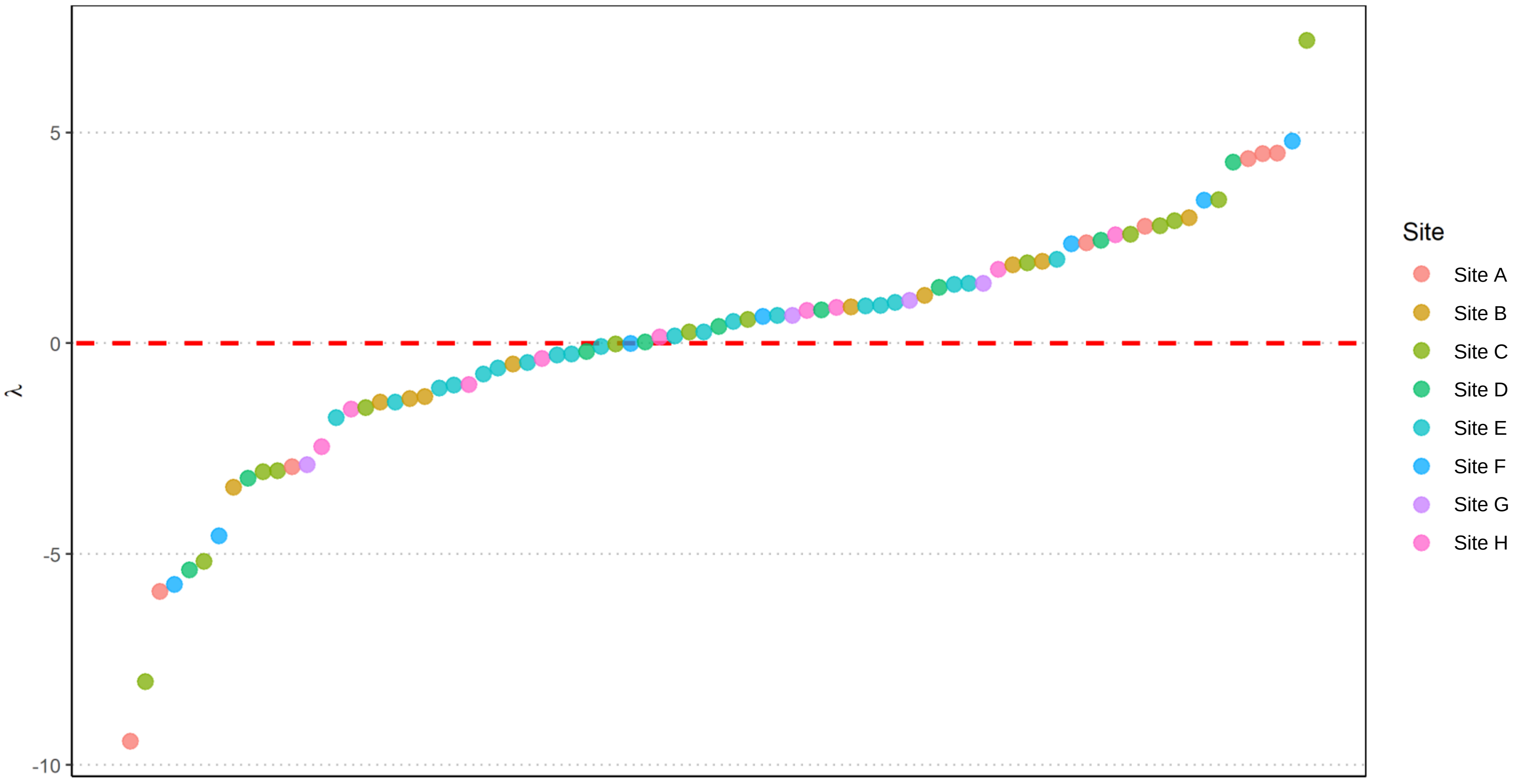
# Evaluating the model

- Top predictors:

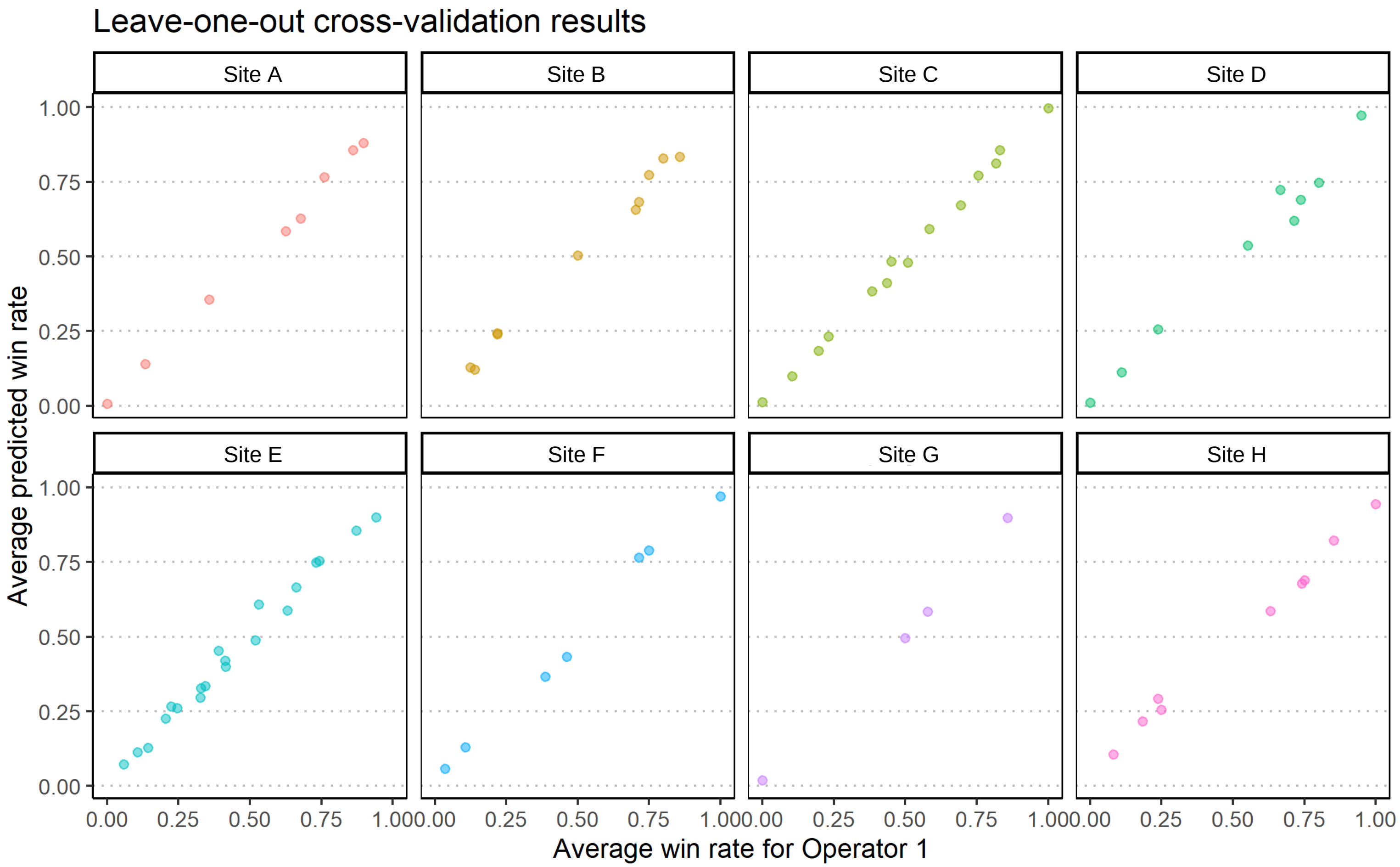
| Predictor | Estimate | Change in *P*(win1) |
|---|---|---|
| Values | 0.50 | 0.12 |
| Picture completion | 0.49 | 0.12 |
| Depth perception | -0.41 | -0.10 |
| Tender-mindedness | 0.33 | 0.08 |
| Wrist extension | -0.23 | -0.06 |
| Excitement seeking | 0.21 | 0.05 |
| Impulsivity | 0.20 | 0.05 |
| Assertiveness | 0.15 | 0.04 |
| Altruism | -0.13 | -0.03 |
| Contrast sensitivity | 0.13 | 0.03 |

For every 1 SD change in predictor, we'd see these corresponding changes to the win probability

# Evaluating the model

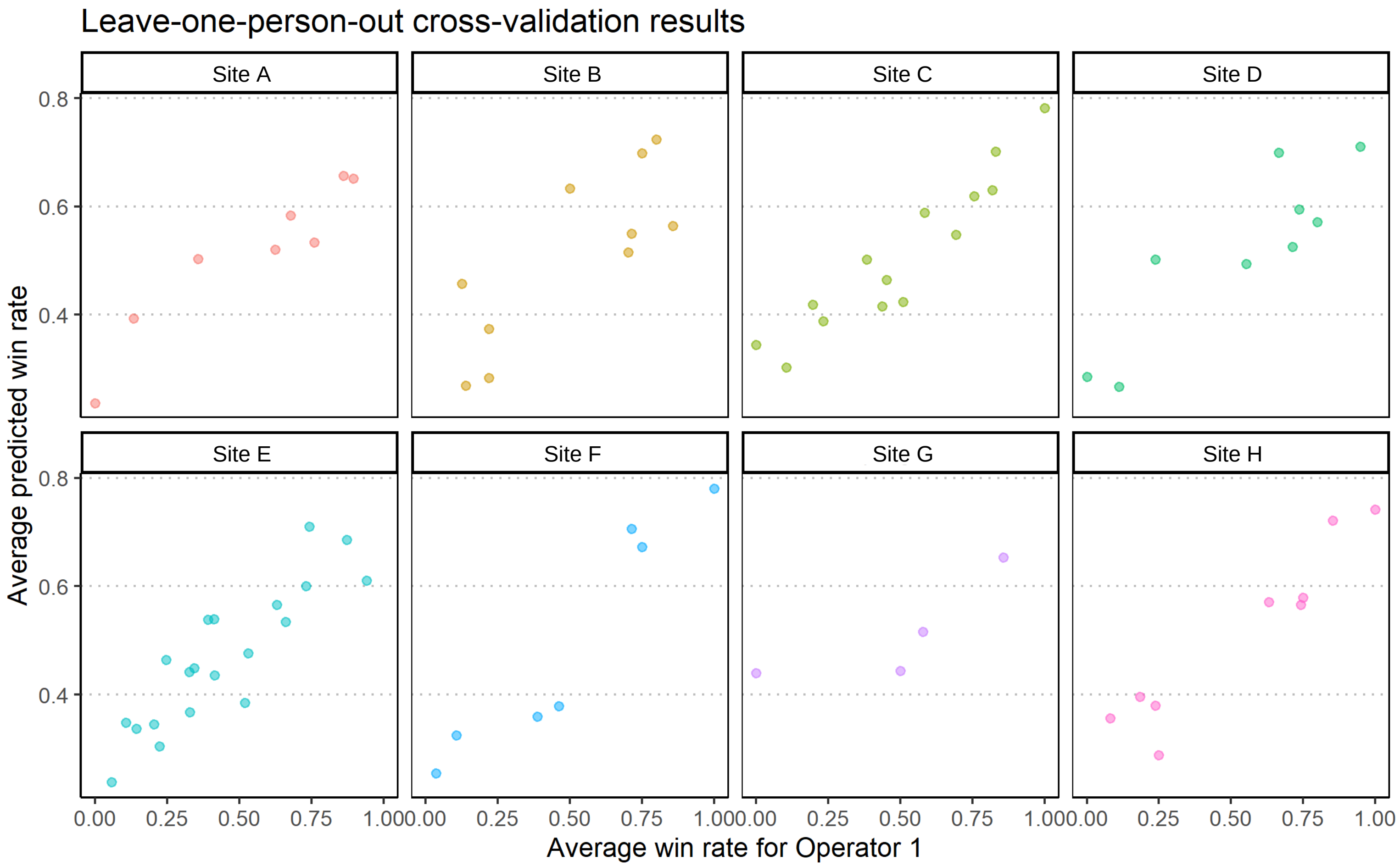


# Evaluating the model




Leave-one-out cross-validation yields highly accurate predictions
- AUC = 0.94
- Accuracy = 0.86

# Evaluating the model

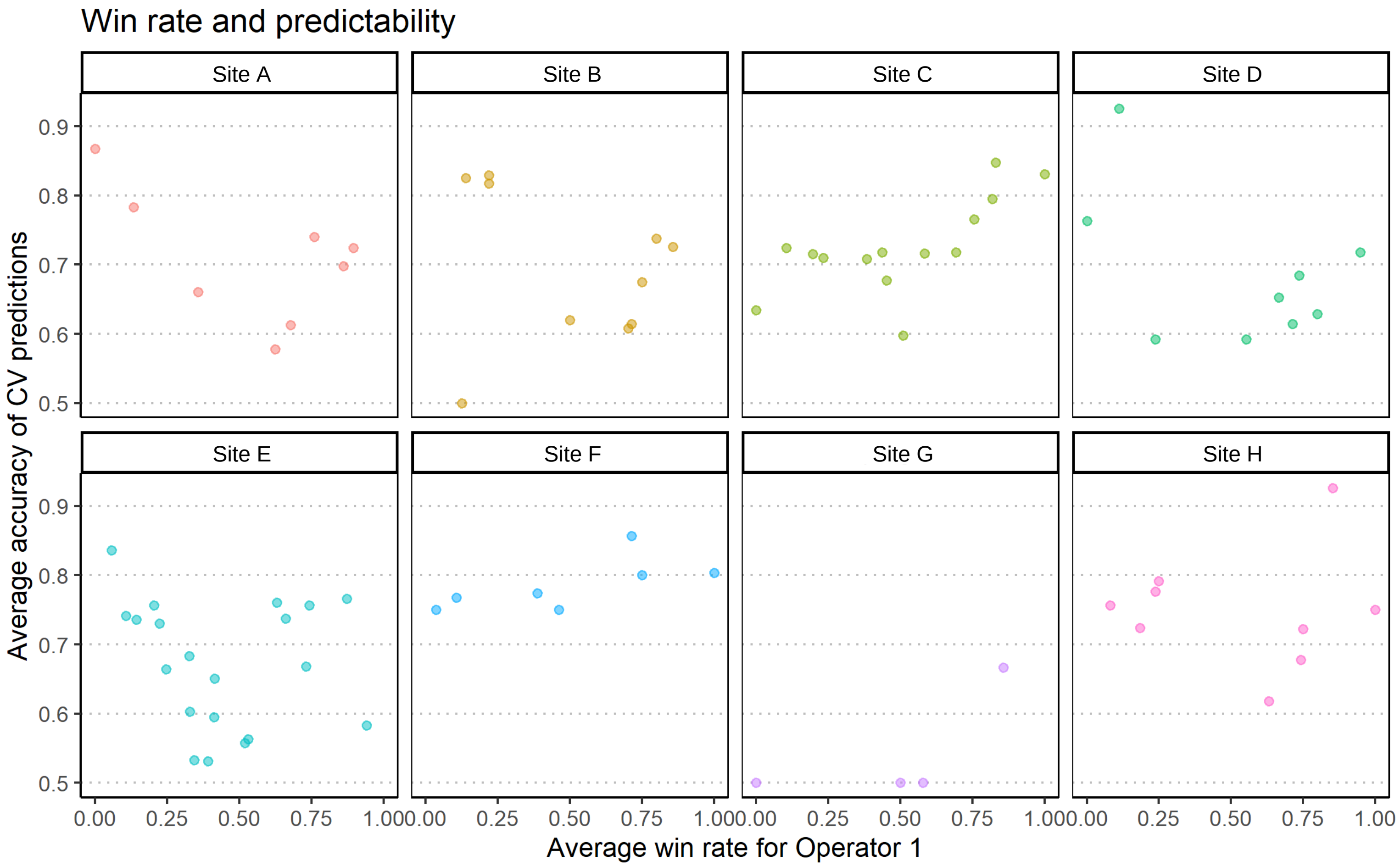



Leave-one-**person**-out cross-validation also yields accurate predictions

- AUC = 0.77
- Accuracy = 0.70

KAIROSRESEARCH

# Evaluating the model




There is **no overall trend** suggesting that low or high performance makes an individual easier to predict

- $r$ = -0.01

Approved for public release

# Takeaways

- The model performs well when predicting novel data
  - The model was **extremely accurate** at predicting novel instances of pairings (i.e., LOO CV)
  - Critically, the model was **accurate** at predicting novel people (i.e., LOPO CV)

# Takeaways

- The model performs well when predicting novel data
  - The model was **extremely accurate** at predicting novel instances of pairings (i.e., LOO CV)
  - Critically, the model was **accurate** at predicting novel people (i.e., LOPO CV)
- Prediction accuracy is independent of observed win rate
  - Worse- and better-performing individuals are all predicted with roughly the same accuracy

# Takeaways

- The model performs well when predicting novel data
  - The model was **extremely accurate** at predicting novel instances of pairings (i.e., LOO CV)
  - Critically, the model was **accurate** at predicting novel people (i.e., LOPO CV)
- Prediction accuracy is independent of observed win rate
  - Worse- and better-performing individuals are all predicted with roughly the same accuracy
- Our hierarchical BTL model is a promising step toward automating evaluations of individual performance