# Statistical Investigation of Scientific Review Group Ratings

Valen E. Johnson

Peer Review Advisory Committee

# Outline of talk

Scientific Questions

Example

Analyses

Summary

# Issues surrounding panel ratings

- Do rating criteria vary systematically between panel members within and between SRGs?
- If yes, how do these variations affect scoring ranges provided to non-reading members?
- Do personality traits (e.g., persuasiveness) of discussants differentially affect non-reader scores?
- How do such effects combine to influence the summary score of a proposal, and how might scoring procedures be changed or modified to minimize these effects?

# Hypothetical Panel Rating Data

| | Proposal | | | |
|---|---|---|---|---|
| Reviewer | A | B | C | D |
| 1 | 1.9 | | | 2.3 |
| 2 | 2.7 | | 2.8 | |
| 3 | | | 1.2 | 1.7 |
| 4 | 2.9 | 3.2 | | |
| 5 | | 1.2 | | 1.3 |
| 6 | 1.8 | | | 2.0 |
| 7 | | 1.9 | 2.0 | |
| 8 | 1.8 | 2.2 | 2.3 | 2.7 |
| 9 | | 2.2 | 2.3 | |
| SRG Mean | 2.22 | 2.14 | 2.12 | 2.0 |

- SRG mean assumes that non-reader ratings resulted in average score equal to average of readers' ratings.

- Order of merit of proposals based on "SRG Mean" is, from best to worst,

$$D > C > B > A$$

- Order of merit of proposals based on "SRG Mean" is, from best to worst,

$$D > C > B > A$$

- Correct ordering of proposals is exactly the opposite!

- Order of merit of proposals based on "SRG Mean" is, from best to worst,

$$D > C > B > A$$

- Correct ordering of proposals is exactly the opposite!

- All reviewers agree that

$$A > B > C > D$$

# Hypothetical Panel Rating Data

| | Proposal | | | |
|---|---|---|---|---|
| Reviewer | A | B | C | D |
| 1 | 1.9 | | | 2.3 |
| 2 | 2.7 | | 2.8 | |
| 3 | | | 1.2 | 1.7 |
| 4 | 2.9 | 3.2 | | |
| 5 | | 1.2 | | 1.3 |
| 6 | 1.8 | | | 2.0 |
| 7 | | 1.9 | 2.0 | |
| 8 | 1.8 | 2.2 | 2.3 | 2.7 |
| 9 | | 2.2 | 2.3 | |
| SRG Mean | 2.22 | 2.14 | 2.12 | 2.0 |

- SRG mean assumes that non-reader ratings resulted in average score equal to average of readers' ratings.

# Hypothetical Panel Rating Data (again)

| | Proposal | | | |
|---|---|---|---|---|
| Reviewer | A | B | C | D |
| 1 | 1.9 | | | 2.3 |
| 2 | 2.7 | | 2.8 | |
| 3 | | | 1.2 | 1.7 |
| 4 | 2.9 | 3.2 | | |
| 5 | | 1.2 | | 1.3 |
| 6 | 1.8 | | | 2.0 |
| 7 | | 1.9 | 2.0 | |
| 8 | 1.8 | 2.2 | 2.3 | 2.7 |
| 9 | | 2.2 | 2.3 | |
| Midpoint | 2.35 | 2.2 | 2.0 | 1.8 |

- Same result at midpoint of range.

# What happened?

| Reviewer | Proposal A | B | C | D | Reviewer Mean |
|---|---|---|---|---|---|
| 1 | 1.9 | | | 2.3 | 2.1 |
| 2 | 2.7 | | 2.8 | | 2.75 |
| 3 | | | 1.2 | 1.7 | 1.45 |
| 4 | 2.9 | 3.2 | | | 3.05 |
| 5 | | 1.2 | | 1.3 | 1.25 |
| 6 | 1.8 | | | 2.0 | 1.9 |
| 7 | | 1.9 | 2.0 | | 1.95 |
| 8 | 1.8 | 2.2 | 2.3 | 2.7 | 2.25 |
| 9 | | 2.2 | 2.3 | | 2.25 |
| Midpoint | 2.35 | 2.2 | 2.0 | 1.8 | |

- Raters used different thresholds or stringency in rating proposals.

# Explanations for reversal

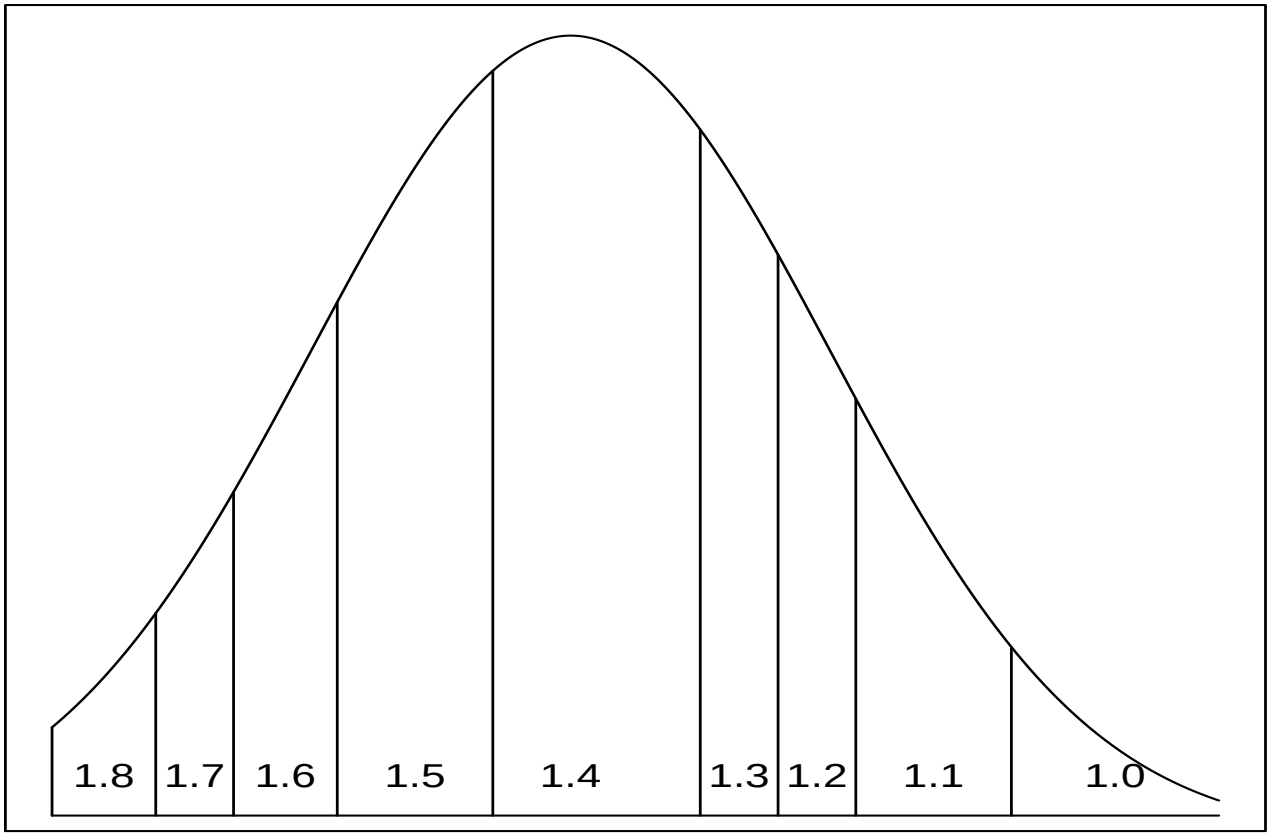
- Variation in "rater thresholds" is common to nearly all rating schemes. College and high school grading suffer from similar effects, as do most employee rating systems.
- If raters employed similar "thresholds" and had similar expertise, then it wouldn't be (as) necessary to have multiple raters evaluate the same proposal!
- Such effects are exacerbated if, say, the "persuasiveness" of raters varies systematically with a raters' critical tendencies.

# Statistical Modeling

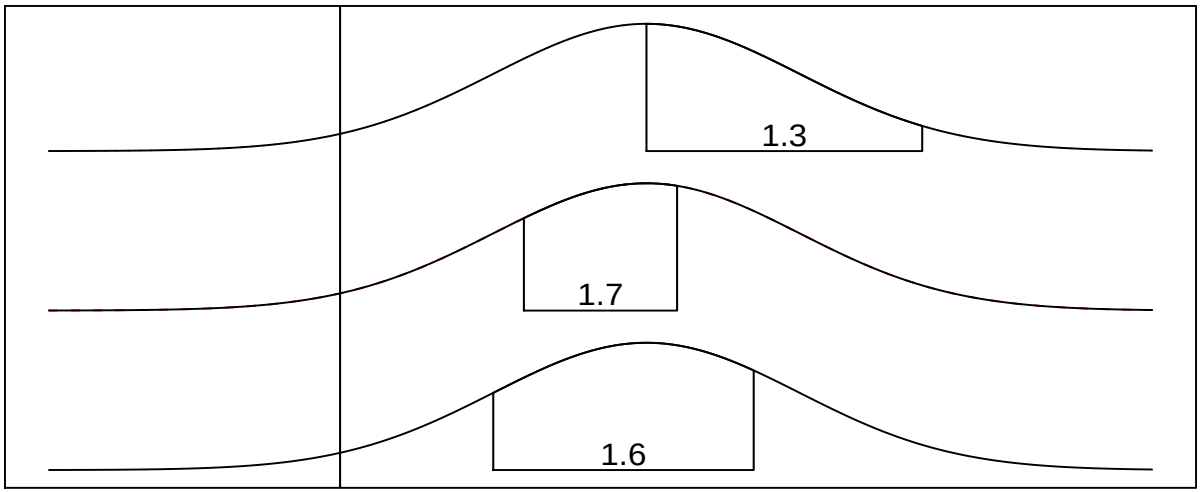
SRG rating data's primary purpose is the estimation of proposal merit. To better estimate a proposal's merit, it is necessary to also estimate

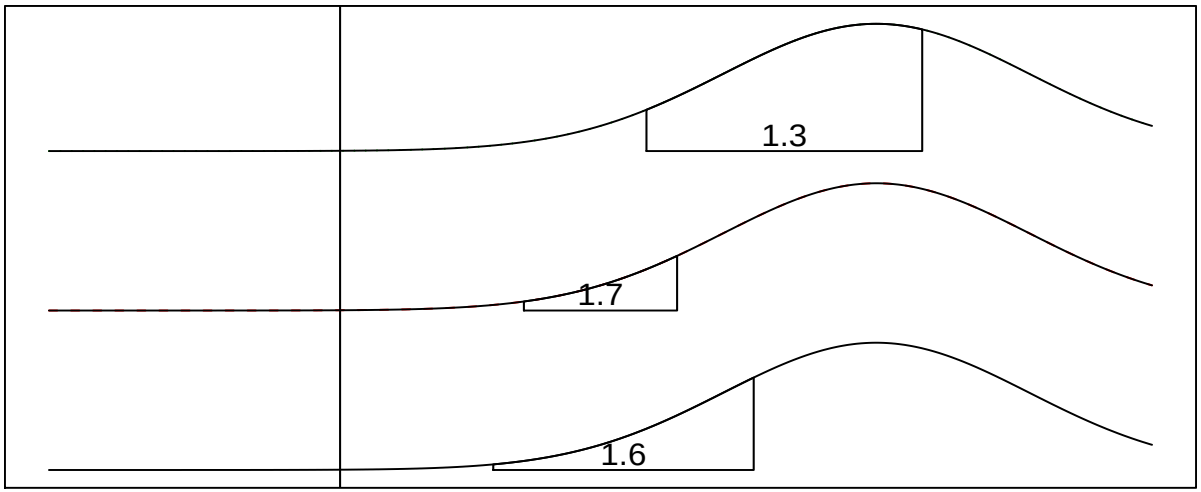
- Rater thresholds
- Rater precision
- Rater "persuasiveness"

# Baseline latent variable model

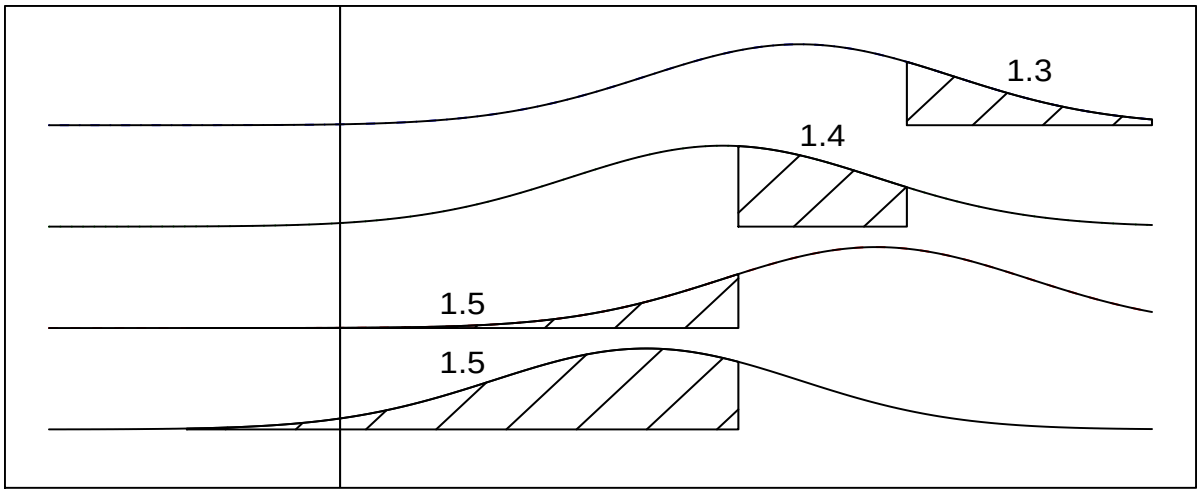


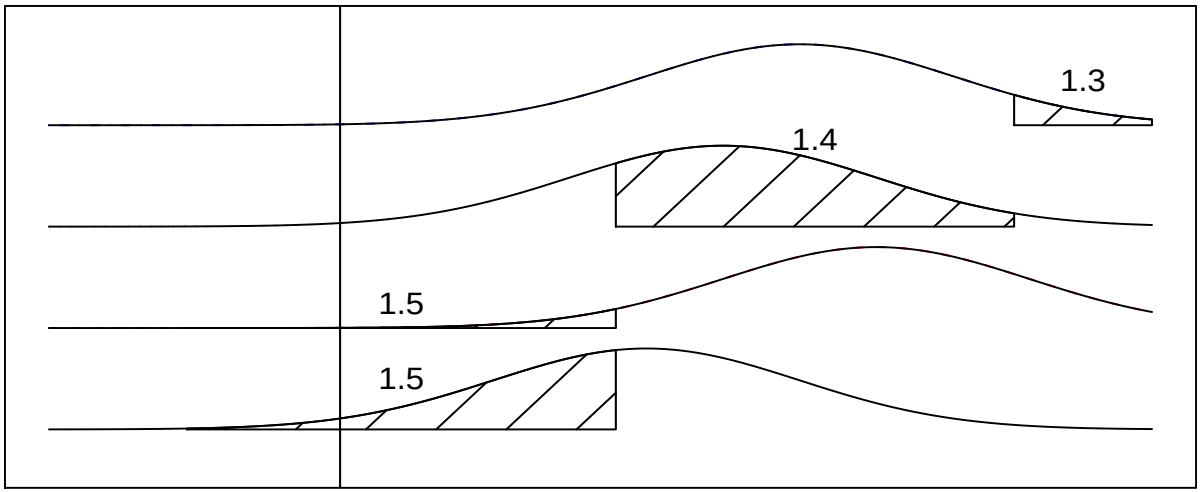
# Estimating proposal merit



# Estimating scoring thresholds



# Tentative model for reader scores

For example,....

- Notation:
  - $z_i$ = latent merit of proposal $i$
  - $z_{i,j}$ = merit of proposal $i$ as observed by reader $j$
  - $\sigma_j^2$ = variance of reader $j$ in observing proposal merit
  - $y_{i,j}$ = preliminary score assigned by reader $j$ to proposal $i$
  - $\gamma_j$ = scoring thresholds for reader $j$
- Then simple model for ratings is

$$z_{i,j} = z_i + \epsilon_{i,j} \qquad \epsilon_{i,j} \sim N(0, \sigma_j^2) \qquad z_i \sim N(0,1)$$

where

$$y_{i,j} = c \qquad \text{if and only if} \qquad \gamma_{j,c-1} < z_{i,j} \leq \gamma_{j,c}$$

# Tentative model for reader scores (cont)

- This simple model gives correct scoring of earlier example

# Tentative model for non-reader scores

Consider non-reader $k$'s interpretation of reader $j$'s score:

- $x_k$ = a value drawn from interval $(\gamma_{k,y_{i,j}-1}, \gamma_{k,y_{i,j}})$
- $\tau_j^2$ = group's perception of rater $j$'s variance in scoring proposals
- $z_{i,j,k}$ = non-reader $k$'s observation of proposal $i$'s merit based on reader $j$'s score
- Assume that

$$z_{i,j,k} \sim N(x_k, \tau_j^2)$$

# Tentative model for non-reader scores (cont.)

- $z_{ik}$ = non-reader $k$'s overall observation of proposal $i$'s merit based on $J$ readers' scores

- Combination of reader ratings leads to

$$z_{ik} \sim N\left(\frac{\sum_j z_{i,j,k}/\tau_j^2}{\sum_j 1/\tau_j^2}, \frac{1}{\sum_j 1/\tau_j^2}\right)$$

# Tentative model for non-reader scores (cont.)

- $a$ = minimum rating from any reader
- $b$ = maximum rating from any reader
- $y_{ik} = d$ if

$$\gamma_{k,d-1} < z_{i,k} \leq \gamma_{k,d} \qquad \text{and} \qquad a \leq d \leq b$$

- Otherwise,

$$y_{ik} = a \text{ or } b$$

with probability dependent on $z_{i,k}$, or a value outside of range.

# Model Extensions

- A combination of the reader and non-reader models can be specified to model reader scores after discussion
- Models can be expanded to account for tendency of non-readers to rate proposals closer to their mean scores
- Model assessment and sensitivity analyses can be performed to determine the importance of various model assumptions on final inference

# Summary

- Potentially serious and undetected biases may affect funding decisions.

# Summary

- Potentially serious and undetected biases may affect funding decisions.
- Such biases, if present, can be detected and quantified.

# Summary

- Potentially serious and undetected biases may affect funding decisions.
- Such biases, if present, can be detected and quantified.
- Statistical modeling may suggest mechanisms for improving the collection and interpretation of SRG scoring data.