Least square differences method for quantitative determination of rater bias

R. Clive Woods
Department of Electrical and Computer Engineering and Microelectronics Research Center
2128 Coover Hall, Iowa State University
Ames, Iowa 50011, U.S.A.
cwoods@iastate.edu

IOWA STATE UNIVERSITY

Grading a large number of projects

- Large panel of raters (faculty?)
- Each rater grades a small number of projects
- Each project graded by ≥2 raters
- Average all grades for each project

Grading a large number of projects

- Large panel of raters (faculty?)
- Each rater grades a small number of projects
- Each project graded by ≥2 raters
- Average all grades for each project
- Sounds familiar?

Typical grade table

Name		Mark 1		Mark 2	Average mark
Project 1	Rater A	$M_{11}^{(1)}$	Rater B	$M_{12}^{(1)}$	$(M_{11}^{(1)} + M_{12}^{(1)}) / 2$
Project 2	Rater C	$M_{23}^{(1)}$	Rater A	$M_{21}^{(1)}$	$(M_{23}^{(1)} + M_{21}^{(1)}) / 2$
Project 3	Rater D	$M_{34}^{(1)}$	Rater C	$M_{33}^{(1)}$	$(M_{34}^{(1)} + M_{33}^{(1)}) / 2$
Project 4	Rater B	$M_{42}^{(1)}$	Rater D	$M_{44}^{(1)}$	$(M_{42}^{(1)} + M_{44}^{(1)}) / 2$
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮

Rater Bias = Rater returns grades deviating from required established impartial standards, so all project(s) rated by that assessor are systematically advantaged or disadvantaged

Determination of rater bias

- Distribute “standard examples” – cumbersome

Determination of rater bias

- Distribute “standard examples” – cumbersome
- Simple statistics: examine each rater’s mean grade – but abnormal average can be caused by both rater bias and/or abnormal batch of projects

Effect of rater bias

- Poor batch of proposals $\Rightarrow$ low average
- Low-marking rater $\Rightarrow$ low average

Effect of rater bias

- Poor batch of proposals $\Rightarrow$ low average
- Low-marking rater $\Rightarrow$ low average

- Good batch of proposals $\Rightarrow$ high average
- High-marking rater $\Rightarrow$ high average

Effect of rater bias

- Poor batch of proposals $\Rightarrow$ low average
- Low-marking rater $\Rightarrow$ low average

- Good batch of proposals $\Rightarrow$ high average
- High-marking rater $\Rightarrow$ high average

- Simple statistics inadequate

IOWA STATE UNIVERSITY

Does it matter?

- At borderlines, <1% can make the difference between grades or funding

Does it matter?

- At borderlines, <1% can make the difference between grades or funding
- Can we grade to this precision?

Does it matter?

- At borderlines, <1% can make the difference between grades or funding
- Can we grade to this precision?
- Traditionally, the answer is “yes”, and we average two raters’ grades to get final result

Determination of rater bias (continued)

- Seek a method of determining rater bias from the grades list alone

Determination of rater bias (continued)

- Seek a method of determining rater bias from the grades list alone
- Each project graded by ≥2 raters; so: compare the grades from each rater to find each rater's rater bias for each project, then average for each rater

Examples

- Rater's own average is low but rater agrees closely with others ("paired raters") on all projects graded

Examples

- Rater's own average is low but rater agrees closely with others ("paired raters") on all projects graded $\Rightarrow$ poor projects

Examples

- Rater's own average is low but rater agrees closely with others ("paired raters") on all projects graded $\Rightarrow$ poor projects
- Rater returns grades consistently lower than others grading the same projects

Examples

- Rater's own average is low but rater agrees closely with others ("paired raters") on all projects graded $\Rightarrow$ poor projects
- Rater returns grades consistently lower than others grading the same projects $\Rightarrow$ rater has negative rater bias (or all the others have positive rater bias)

IOWA STATE UNIVERSITY

Assumptions

- Each rater grades accurately and professionally the relative quality of the proposals seen but habitually with a high/low average, and with a high/low standard deviation

Assumptions

- Each rater grades accurately and professionally the relative quality of the proposals seen but habitually with a high/low average, and with a high/low standard deviation
- Hence, each rater's mean and standard deviation can be adjusted to achieve "best agreement" with paired assessors

ECE

IOWA STATE UNIVERSITY

Algorithm

- Shift all a rater's grades so that summed squared differences between grades from each rater with the corresponding paired grades are minimized

Algorithm

- Shift all a rater's grades so that summed squared differences between grades from each rater with the corresponding paired grades are minimized
- Do the same with standard deviations

Algorithm

- Shift all a rater's grades so that summed squared differences between grades from each rater with the corresponding paired grades are minimized
- Do the same with standard deviations
- Apply to all raters in turn

Algorithm

- Shift all a rater's grades so that summed squared differences between grades from each rater with the corresponding paired grades are minimized
- Do the same with standard deviations
- Apply to all raters in turn
- Adjustments to each rater affect adjustments to other raters, so go back to top and iterate until converged

Limitations

- Each rater must grade a significant number of projects, so means and standard deviations are accurate

Limitations

- Each rater must grade a significant number of projects, so means and standard deviations are accurate
- If a rater grades just one or two projects, algorithm produces exact agreement with the paired raters

Limitations

- Each rater must grade a significant number of projects, so means and standard deviations are accurate
- If a rater grades just one or two projects, algorithm produces exact agreement with the paired raters
- Each rater must be paired with a representative sample of the other raters for the “network” to operate satisfactorily

Results

- Algorithm programmed

Results

- Algorithm programmed
- Program tested and applied to several sets of genuine grades

Results

- Algorithm programmed
- Program tested and applied to several sets of genuine grades
- Convergence typically takes ~500 iterations (a few seconds on a fast PC)

Results

- Algorithm programmed
- Program tested and applied to several sets of genuine grades
- Convergence typically takes ~500 iterations (a few seconds on a fast PC)
- Typical shifts of grades indicate precision of grades

Results

- Algorithm programmed
- Program tested and applied to several sets of genuine grades
- Convergence typically takes ~500 iterations (a few seconds on a fast PC)
- Typical shifts of grades indicate precision of grades
- Identify raters with abnormally high or low averages and standard deviations

Results

- Algorithm programmed
- Program tested and applied to several sets of genuine grades
- Convergence typically takes ~500 iterations (a few seconds on a fast PC)
- Typical shifts of grades indicate precision of grades
- Identify raters with abnormally high or low averages and standard deviations
- Identify projects advantaged or disadvantaged as a result of rater bias

ECpE

IOWA STATE UNIVERSITY

Use of program

- Identify projects needing grading by an extra rater(s)

Use of program

- Identify projects needing grading by an extra rater(s)
- Identify raters applying incorrect standards

ECpE

IOWA STATE UNIVERSITY

Use of program

- Identify projects needing grading by an extra rater(s)
- Identify raters applying incorrect standards
- Typically, report unadjusted marks

Use of program

- Identify projects needing grading by an extra rater(s)
- Identify raters applying incorrect standards
- Typically, report unadjusted marks
- Each rater must be linked with all others though pairings, directly or indirectly, for the “network” to operate – but this is also true if the algorithm is not used

Conclusions

- Simple statistics (e.g. calculating raters' means) inadequate for determining rater bias

Conclusions

- Simple statistics (e.g. calculating raters' means) inadequate for determining rater bias
- Rater bias is a real effect and ignoring it leads to trouble

Conclusions

- Simple statistics (e.g. calculating raters' means) inadequate for determining rater bias
- Rater bias is a real effect and ignoring it leads to trouble
- Rater bias may be determined quantitatively and self-consistently from a list of grades

Conclusions

- Simple statistics (e.g. calculating raters' means) inadequate for determining rater bias
- Rater bias is a real effect and ignoring it leads to trouble
- Rater bias may be determined quantitatively and self-consistently from a list of grades
- Many faculty have insignificant rater biases (<2%) but some have >5%

ECpE

IOWA STATE UNIVERSITY

Conclusions

- Simple statistics (e.g. calculating raters' means) inadequate for determining rater bias
- Rater bias is a real effect and ignoring it leads to trouble
- Rater bias may be determined quantitatively and self-consistently from a list of grades
- Many faculty have insignificant rater biases (<2%) but some have >5%
- When choosing paired raters, all raters must be linked with all others, directly or indirectly

ECpE

IOWA STATE UNIVERSITY

The end

IOWA STATE UNIVERSITY