Estimating Rater Consistency: Which Method Is Appropriate?

Robert L. Johnson
Grant Morgan
Min Zhu
Vasanthi Rao

University of South Carolina

2010 AEA Presentation, San Antonio, Texas

Methods for Examining Rater Consistency

Percent agreement (Andrade, Du, & Wang, 2008; Herman, Gearhart, & Baker, 1993; Johnson, McDaniel, & Willeke, 2000; Johnson, Penny, & Gordon, 2001; Koretz, Stecher, Klein, & McCaffrey, 1994; LeMahieu, Gitomer, & Eresh, 1995)

Pearson correlation (Herman, Gearhart, & Baker, 1993)

Spearman correlation (Johnson, McDaniel, & Willeke, 2000; Johnson, Penny, & Gordon, 2001; Koretz, Stecher, Klein, & McCaffrey, 1994; Supovitz, MacGowan, & Slattery, 1997)

Cronbach's alpha (van der Schaaf, Stokking, & Verloop, 2005)

Generalizability/dependability coefficient (Johnson, McDaniel, & Willeke, 2000; Johnson, Penny, & Gordon, 2001; Nie, Yeo, & Lau, 2007; Shavelson, Solano-Flores, & Ruiz-Primo, 1998; Yao, Thomas, Nickens, Downing, Burkett, & Lamson, 2008)

Questions

- Do we arrive at different conclusions when we use different methods of estimating interrater consistency?
- If so, which method results in a better estimate of interrater reliability?

The Relation between Agreement Levels and Correlation Estimates of Interrater Reliability

Rubric Scale	Agreement	Correlation between raters										
		0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
			Percent agreement between ratings using a 4- and 6-point rubric									
4	Exact	28	30	33	35	37	40	45	50	57	68	78
	Exact & Adjacent	73	77	80	83	85	89	92	95	98	100	100
6	Exact	26	28	30	32	34	35	40	46	52	64	75
	Exact & Adjacent	69	74	77	79	82	87	90	94	98	100	100

Johnson, R., Penny, J., & Gordon, B. (2009). *Assessing performance: Developing, scoring, and validating performance tasks*. New York: Guilford Publications.

- Empirical examination of interrater reliability estimates across methods – Min Zhu
- Monte Carlo simulation of interrater reliability estimates across methods – Grant Morgan

References

Andrade, H., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice, 27*(2), 3-13.

Herman, J., Gearhart, M. & Baker, E. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment, 1*(3),201-224.

Johnson, R., McDaniel, F., & Willeke, M. (2000). Using portfolios in program evaluation: An investigation of interrater reliability. *The American Journal of Evaluation*, *21*(1), 65-80.

Johnson, R., Penny, J., & Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication, 18*(2), 229-249.

Johnson, R., Penny, J., & Gordon, B. (2009). *Assessing performance: Developing, scoring, and validating performance tasks*. New York: Guilford Publications.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5-16.

LeMahieu, P., Gitomer, D., & Eresh, J. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice, 14*(3), 11-16, 25-28.

Nie, Y., Yeo, S., & Lau, S. (2007). Application of generalizability theory in the investigation of the quality of journal writing in mathematics. *Studies in Educational Evaluation 33*, 371-383.

Shavelson, R., Solano-Flores, G., & Ruiz-Primo, M. (1998). Toward a science performance assessment technology. *Evaluation and program planning, 21*(2), p. 171-184.

Supovitz, J., MacGowan, A., & Slattery, J. (1997). Assessing agreement: An examination of the interrater reliability of portfolio assessment in Rochester, New York. *Educational Assessment, 4*(3), 237-259.

Van der Schaaf, M., Stokking, K., & Verloop, N. (2005). Cognitive representations in raters' assessment of teacher portfolios. *Studies in Educational Evaluation 31*, 27-55.

Yao, Y., Thomas, M., Nickens, N., Downing, J., Burkett, R., & Lamson, S. (2008). Validity evidence of an electronic portfolio for preservice teachers. *Educational Measurement: Issues and Practice, 27*(1), 10-24,

Estimating Rater Consistency: How Do Methods Differ?

Min Zhu, Robert Johnson, Grant Morgan, & Vasanthi Rao
University of South Carolina
Department of Educational Studies

2010 AEA Presentation, San Antonio, Texas

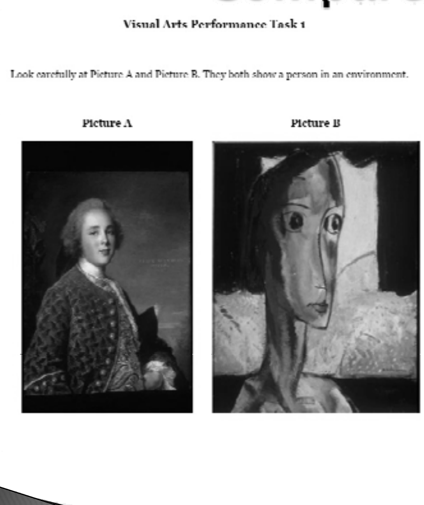
SCAAP Overview

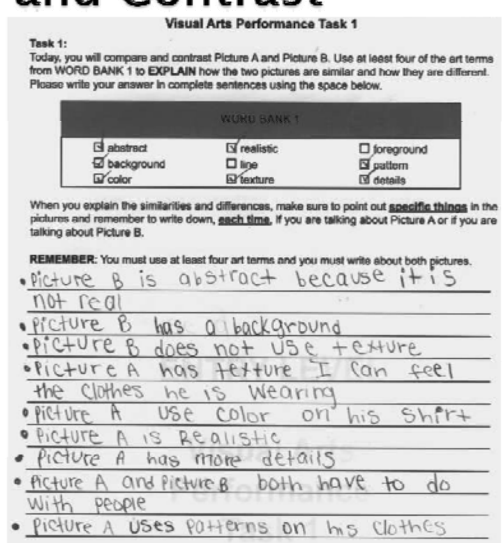
- The South Carolina Arts Assessment Program (SCAAP) was established by the SC Department of Education in 2000.
- Purpose: to provide arts educators and school administrators with a tool to measure their students' arts achievement and to objectively evaluate their schools' arts programs.
- Uniqueness: a web-based standardized arts assessment system
  - Include 6 assessments
  - Each assessment includes:
    - Two 45-item multiple-choice test forms
    - Two/three performance tasks
- Test developers –
  - South Carolina arts educators
  - Measurement specialists at the Office of Program Evaluation (OPE) at the University of South Carolina

Data Source

- 2007 SCAAP entry-level visual arts performance assessment results
  - Two tasks: one writing and one drawing
  - 8 raters and 4 paired-rater groups
  - 500 students in each group

SCAAP Visual Arts Task 1 -- Compare and Contrast




Visual Arts Performance Task 1

Task 1:
Today, you will compare and contrast Picture A and Picture B. Use at least four of the art terms from WORD BANK 1 to **EXPLAIN** how the two pictures are similar and how they are different. Please write your answer in complete sentences using the space below.

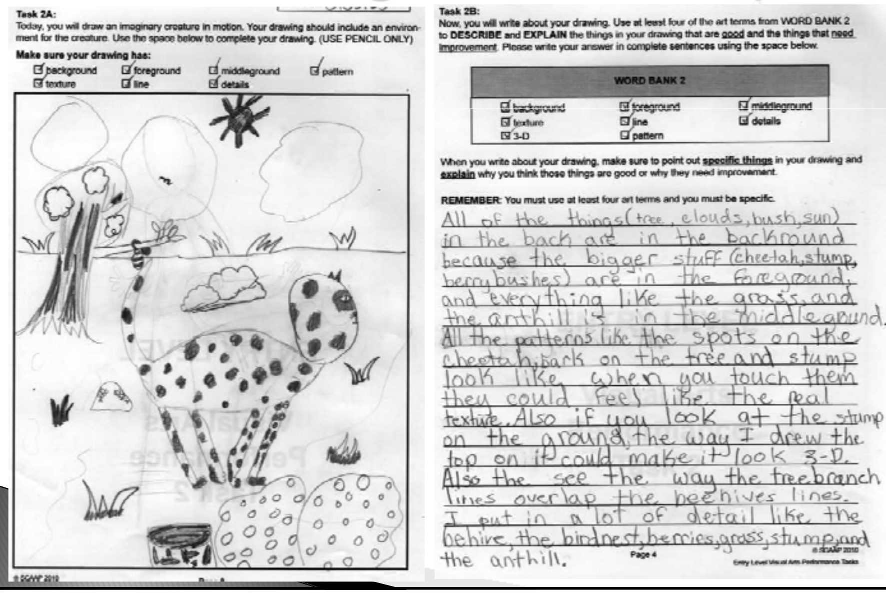
WORD BANK 1		
☑ abstract	☑ realistic	☐ foreground
☑ background	☐ line	☑ pattern
☑ color	☑ texture	☑ details

When you explain the similarities and differences, make sure to point out **specific things** in the pictures and remember to write down, **each time**, if you are talking about Picture A or if you are talking about Picture B.

REMEMBER: You must use at least four art terms and you must write about both pictures.

- Picture B is abstract because it is not real
- picture B has a background
- Picture B does not Use texture
- Picture A has texture I can feel the clothes he is Wearing
- picture A use color on his Shirt
- Picture A is Realistic
- Picture A has more details
- Picture A and Picture B both have to do With people
- Picture A uses patterns on his clothes

SCAAP Visual Arts Task 2 -- Drawing and Self-Critique



Task 2A:
Today, you will draw an imaginary creature in motion. Your drawing should include an environment for the creature. Use the space below to complete your drawing. (USE PENCIL ONLY)

Make sure your drawing has:
- ☑ background ☑ foreground ☑ middleground ☑ pattern
- ☑ texture ☑ line ☑ details

Task 2B:
Now, you will write about your drawing. Use at least four of the art terms from WORD BANK 2 to **DESCRIBE** and **EXPLAIN** the things in your drawing that are good and the things that need improvement. Please write your answer in complete sentences using the space below.

WORD BANK 2		
☑ background	☑ foreground	☑ middleground
☑ texture	☑ line	☑ details
☑ 3-D	☑ pattern	

When you write about your drawing, make sure to point out **specific things** in your drawing and **explain** why you think those things are good or why they need improvement.

REMEMBER: You must use at least four art terms and you must be specific.

All of the things(tree, clouds, bush, sun) in the back are in the backround because the bigger stuff (cheetah, stump, berry bushes) are in the foreground, and everything like the grass, and the anthill is in the middleground. All the patterns like the spots on the cheetah, bark on the tree and stump look like when you touch them they could feel like the real texture. Also if you look at the stump on the ground, the way I drew the top on it could make it look 3-D. Also the see the way the tree branch lines overlap the beehives lines. I put in a lot of detail like the behive, the birdnest, berries, grass, stump, and the anthill.

SCAAP Web-Based Rating System

- Raters: Trained arts professionals
- Rubrics:
  - Holistic rubrics for visual arts
  - Scale ranges from 0 to 4 with raters also being allowed to use augmentation (e.g. 2-, 2, 2+).
- Benchmarking:
  - Validation Committee members select student responses representative of each rubric level and use these as:
    - anchor responses
    - practice responses
    - qualifying responses
    - seed responses

SCAAP Web-Based Rating System (Cont')

- Rater Training
  - One-day training session at a central location
  - Anchor items are presented and explained.
  - Raters take a web-based practice test that provides detailed feedback.
  - Each rater is required to score at least 90% adjacent agreement on a 15-item, randomly generated qualifying test.
  - After passing the qualifying test, raters can score student responses.
  - Following the training, raters score student responses remotely via the SCAAP website – https://scaap.ed.sc.edu.

SCAAP Web-Based Rating System (Cont')

- Scoring & Monitoring
  - Raters are required to pass a randomly-generated 15-item refresher test after scoring every 100 student responses.
  - Seed responses are randomly distributed among unscored student performances to monitor rater accuracy.
  - Each student response is scored by two-raters. An expert rater is used for score resolution.

Rater Consistency Estimates in the Literature

Methods	1990s	2000s
Percent Agreement		
Exact	/	2
Adjacent	1	5
Kappa coefficient	/	1
Pearson product moment correlation coefficient (PPMCC)	2	3
Spearman rank-order	1	/
Cronbach's alpha	/	/
Intraclass correlation (ICC)	/	4
G-theory		
G-coefficient	1	2
Phi-coefficient	/	/
Multifaceted Rasch model (MFRM)	2	3
Others	1	/

Measures of Rater Agreement

- Percent Exact Agreement
- Percent Adjacent Agreement
- Advantage
  - Distribution-free estimate
  - Easy to compute
- Disadvantage
  - The small range of the scale in rubrics can inflate the estimate.
  - Chance agreement is not considered.

Sample: Percent Agreement --Exact and Adjacent

R1	R2					
	0	1	2	3	4	Total
0	**22 (4.41%)**	19 (3.81%)	100 (20.04%)	7 (1.4%)	0 (0%)	148 (29.66%)
1	1 (0.2%)	**4 (0.8%)**	101 (20.24%)	13 (2.61%)	4 (0.8%)	123 (24.65%)
2	0 (0%)	2 (0.4%)	**71 (14.23%)**	38 (7.62%)	4 (0.8%)	115 (23.05%)
3	0 (0%)	0 (0%)	17 (3.41%)	**26 (5.21%)**	14 (2.81%)	57 (11.42%)
4	0 (0%)	1 (0.2%)	4 (0.8%)	18 (3.61%)	**33 (6.61%)**	56 (11.22%)
Total	23	26	293	102	55	499

- *Note: Exact agreement 31.26%;*

  Adjacent agreement 73.34%

Measures of Association

- Pearson product-moment correlation coefficient (PPMCC)
- Spearman's rank-order correlation coefficient (SRCC)
- Polychoric correlation coefficient (PCC)

Measures of Association (Cont')

	Applications	Assumptions
PPMCC	Association between two continuous variables	✓ Bivariate normality ✓ No measurement error
Spearman Rank-order	Association between two ordinal variables	✓ Shape identity ✓ No measurement error
Polychoric	Association between two continuous latent variables grouped into ordered classes	✓ Latent bivariate normality ✓ No measurement error

G-coefficient and Phi-coefficient

- G-coefficient
  - When the ranking of individual or group scores is the focus
  - In a G-study with raters as a facet

$$\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(pr)}$$

- Phi-coefficient (index of dependability)
  - When examinees performance on a criterion-referenced test is of interest
  - With raters as the only facet, the phi-coefficient takes into account shifts in rater means and allows detection of raters who are overly severe or lenient.

$$\phi = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(r) + \sigma^2(pr)}$$

Questions to Answer

- How different are these interrater consistency estimates?
- How does the range of the rating scale (i.e. with and without augmentation) impact the difference among these interrater consistency estimates?
- How does the pattern differ across performance tasks?

SCAAP Visual Arts Task 1 Consistency
--Without Augmentation in Rating

Raters	N	R1		R2		Exact (%)	Adj (%)	PPMCC	SRCC	PCC	G-C	Phi-C
		Mean	SD	Mean	SD							
G1	499	1.50	1.32	2.28	0.90	31.26	73.34	0.66	0.65	0.76	0.61	0.57
G2	491	1.71	1.24	1.70	1.08	53.56	92.06	0.74	0.72	0.82	0.73	0.73
G3	496	1.26	1.26	1.77	0.88	40.52	84.07	0.70	0.70	0.86	0.66	0.64
G4	496	1.99	0.95	1.33	1.19	40.52	80.24	0.68	0.64	0.78	0.66	0.63

- *Note: Exact – Exact agreement*
  Adj – Adjacent agreement
  PPMCC – Pearson product-moment correlation coefficient
  SRCC – Spearman's rank-order correlation coefficient
  PCC – Polychoric correlation coefficient
  G-C – G-coefficient
  Phi-C – Phi-coefficient

SCAAP Visual Arts Task 1 Consistency --With Augmentation in Rating

Raters	N	R1		R2		Exact (%)	Adj (%)	PPMCC	SRCC	PCC	G-C	Phi-C
		Mean	SD	Mean	SD							
G1	499	1.48	1.31	2.31	0.90	14.03	27.46	0.68	0.68	0.75	0.63	0.59
G2	491	1.72	1.24	1.71	1.08	47.45	53.56	0.74	0.73	0.82	0.74	0.74
G3	496	1.28	1.27	1.75	0.89	33.47	40.32	0.73	0.75	0.86	0.68	0.67
G4	496	1.98	0.95	1.33	1.18	32.46	40.32	0.69	0.66	0.78	0.68	0.65

- *Note: Exact – Exact agreement*
  Adj – Adjacent agreement
  PPMCC – Pearson product-moment correlation coefficient
  SRCC – Spearman rank-order correlation coefficient
  PCC – Polychoric correlation coefficient
  G-C – G-coefficient
  Phi-C – Phi-coefficient

SCAAP Visual Arts Task 2A Consistency --Without Augmentation in Rating

Raters	N	R1		R2		Exact (%)	Adj (%)	PPMCC	SRCC	PCC	G-C	Phi-C
		Mean	SD	Mean	SD							
G1	495	2.02	0.68	2.20	1.02	49.29	94.95	0.63	0.62	0.75	0.58	0.57
G2	489	1.65	0.83	1.78	0.69	59.71	98.97	0.65	0.65	0.75	0.64	0.63
G3	491	1.90	0.78	1.96	0.83	58.04	98.16	0.63	0.62	0.72	0.63	0.63
G4	493	1.58	0.80	2.10	0.82	36.92	93.71	0.58	0.56	0.66	0.58	0.55

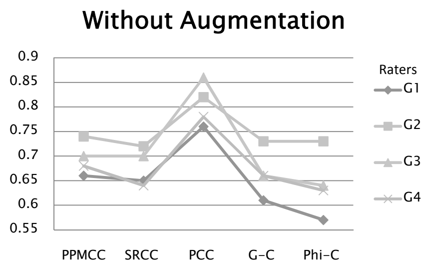
- *Note: Exact – Exact agreement*
  Adj – Adjacent agreement
  PPMCC – Pearson product-moment correlation coefficient
  SRCC – Spearman's rank-order correlation coefficient
  PCC – Polychoric correlation coefficient
  G-C – G-coefficient
  Phi-C – Phi-coefficient

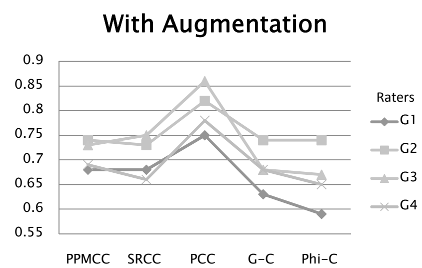
SCAAP Visual Arts Task 2A Consistency --With Augmentation in Rating

Raters	N	R1		R2		Exact (%)	Adj (%)	PPMCC	SRCC	PCC	G-C	Phi-C
		Mean	SD	Mean	SD							
G1	495	1.99	0.68	2.22	1.01	27.88	47.07	0.67	0.66	0.71	0.63	0.59
G2	489	1.65	0.83	1.78	0.69	52.56	59.72	0.67	0.67	0.75	0.74	0.74
G3	491	1.88	0.76	1.94	0.79	30.75	56.62	0.70	0.68	0.74	0.68	0.67
G4	493	1.58	0.78	2.09	0.81	26.17	36.52	0.63	0.62	0.69	0.68	0.65

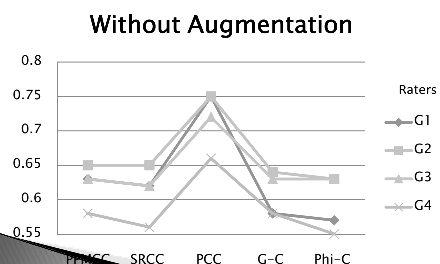
- *Note: Exact – Exact agreement*
  Adj – Adjacent agreement
  PPMCC – Pearson product-moment correlation coefficient
  SRCC – Spearman's rank-order correlation coefficient
  PCC – Polychoric correlation coefficient
  G-C – G-coefficient
  Phi-C – Phi-coefficient

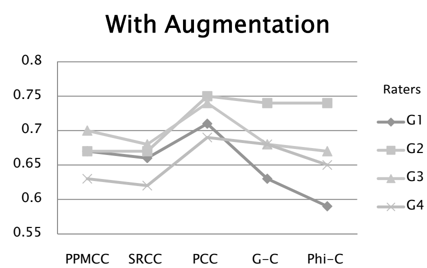
Task 1





Task 2A





Findings

- Consistent with previous studies, introducing augmentation scores does not result in large changes in mean scores, but increases some of the interrater reliability coefficient estimates (excluding polychoric correlation).
- As expected, phi-coefficients are slightly lower than G-coefficients in some instances, indicating the potential existence of a small rater effect.
- Polychoric correlations are always higher than other reliability estimates.
- In many cases, PPMCC, Spearman, and G-coefficients were very close.
- Such a pattern is quite consistent across the two tasks.

What's Next…

- Which reliability coefficient is closer to the truth?
- What should we consider when choosing a coefficient in our report?
- A simulation study will tell us more.

Which Measure Is Appropriate for Estimating Rater Consistency? A Simulation Study

Grant Morgan
Robert Johnson
Min Zhu
Vasanthi Rao

University of South Carolina

Evaluation 2010
American Evaluation Association – San Antonio, TX

Presentation Overview

- Select estimates of rater consistency
- What does "appropriate" mean?
  - Ease of communication
  - Estimates & data alignment
  - Accuracy of inferences
- Conclusions

Rater Consistency Estimates

	Applications	Assumptions
Pearson Product-Moment	Association between two continuous variables	✓ Bivariate normality ✓ No measurement error
Spearman	Association between two ordinal variables	✓ Shape identity ✓ No measurement error
Polychoric	Association between two continuous latent variables grouped into ordered classes	✓ Latent bivariate normality ✓ No measurement error
G-coefficient	Partition systematic and unsystematic error variation	✓ Randomly parallel tests sampled from the same population (i.e., universe)

Which measure is "appropriate"?

1) Ease of communication
- Pearson product-moment correlation coefficient
  - Proportion of explained variance when squared

"**Pearson's product-moment correlation** is the most commonly reported, even for those data for which it is superficially not a good match. Of course, the same is true of other familiar statistics, such as the mean and standard deviation" (Linacre, 2005, p.1028).

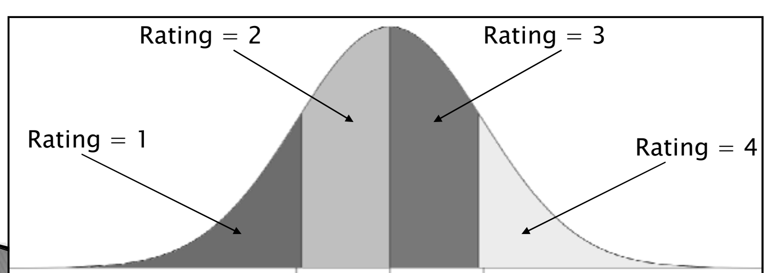
Which measure is "appropriate"?

1) Ease of communication
- Pearson product-moment correlation coefficient

2) Alignment between analysis and data
- Polychoric correlation coefficient
  - *Recall: Correlation between two latent continuous distributions that have been chunked into ordinal scales*

Performance Assessment Data

- Features:
  - Ability is a normally-distributed latent variable
  - Ability distribution is chunked into an ordinal scale (rubric rating scale)



Previous Research

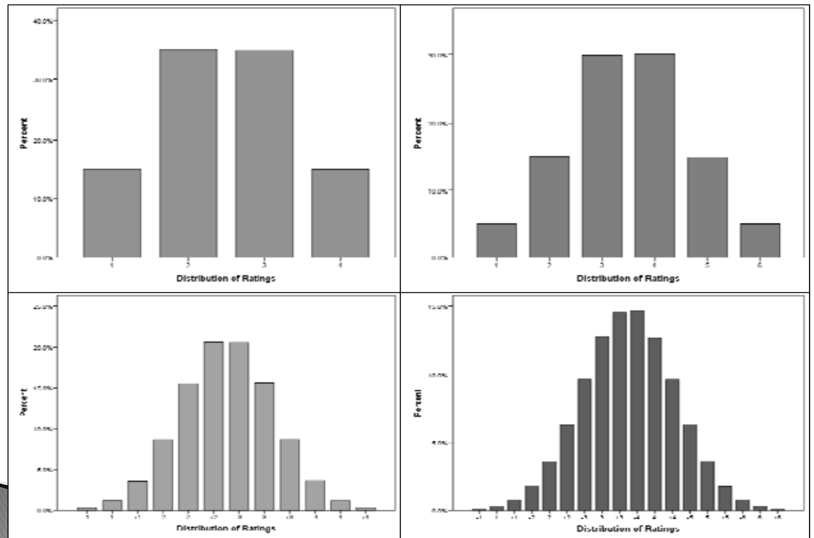
Problems with treating ordinal data as continuous

- No origins or units of measure (Joreskog, 1994)
- Increased likelihood of correlating error variances (Anderson & Gerbing, 1988)
- Standard error & chi-square tests are incorrect when using product-moment matrix with ordinal data (Bentler & Lee, 1983).

Design Factors

- Levels of inter-rater reliability
  - .70, .75, .80, .85, .90, .95
- Number of tasks
  - 25, 100, 250, 500, 2000
- Number of rating scale categories
  - 4, 6, 4 with augmentation (12), 6 with augmentation (18)
- 1,000 replications of each condition

Distributions

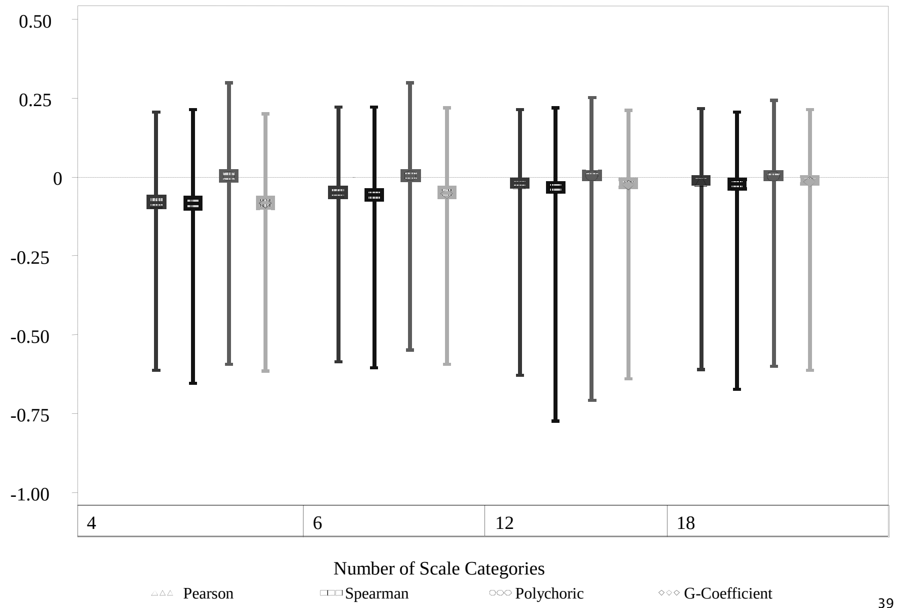


Estimated Bias

- Let simulated value of IRR = ρ
- E(ρ-hat) = ρ + Δ, where Δ = bias

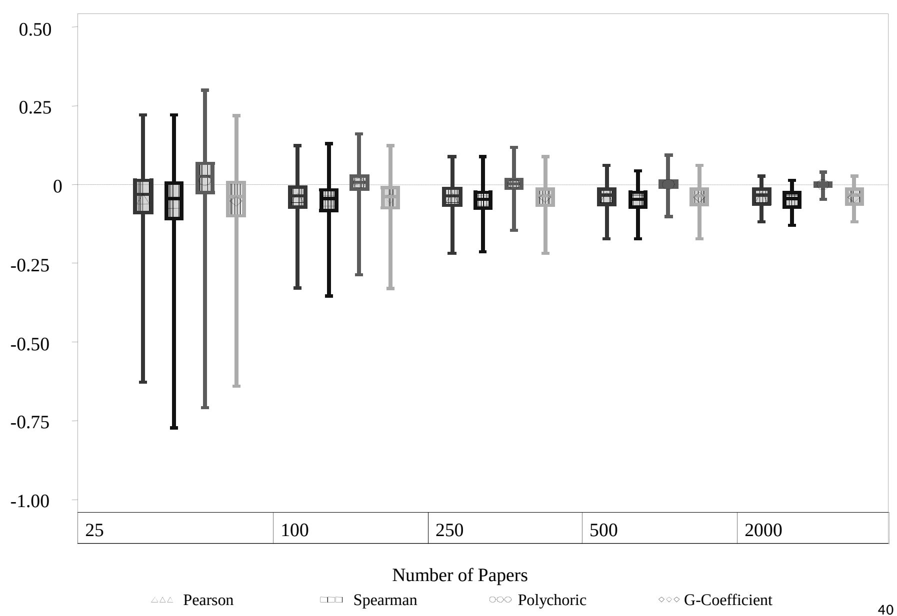
- We're interested in Δ!

Estimated Bias by Number of Scale Categories



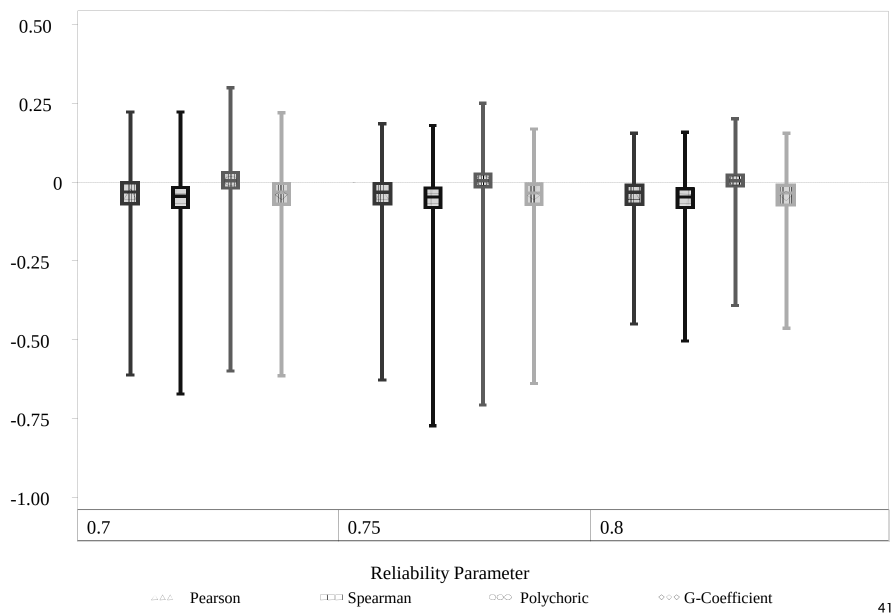
39

Estimated Bias by Number of Papers



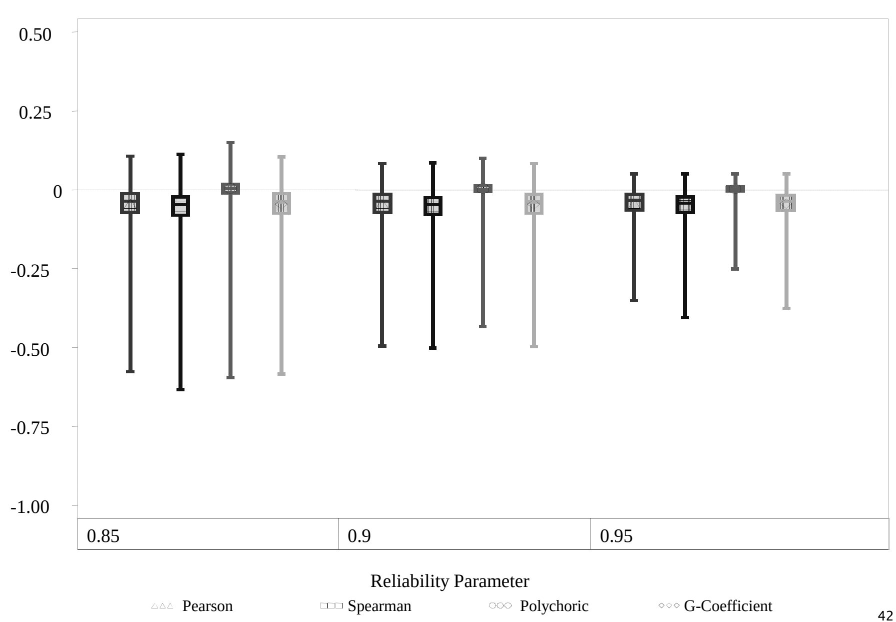
40

Estimated Bias by Reliability Parameter

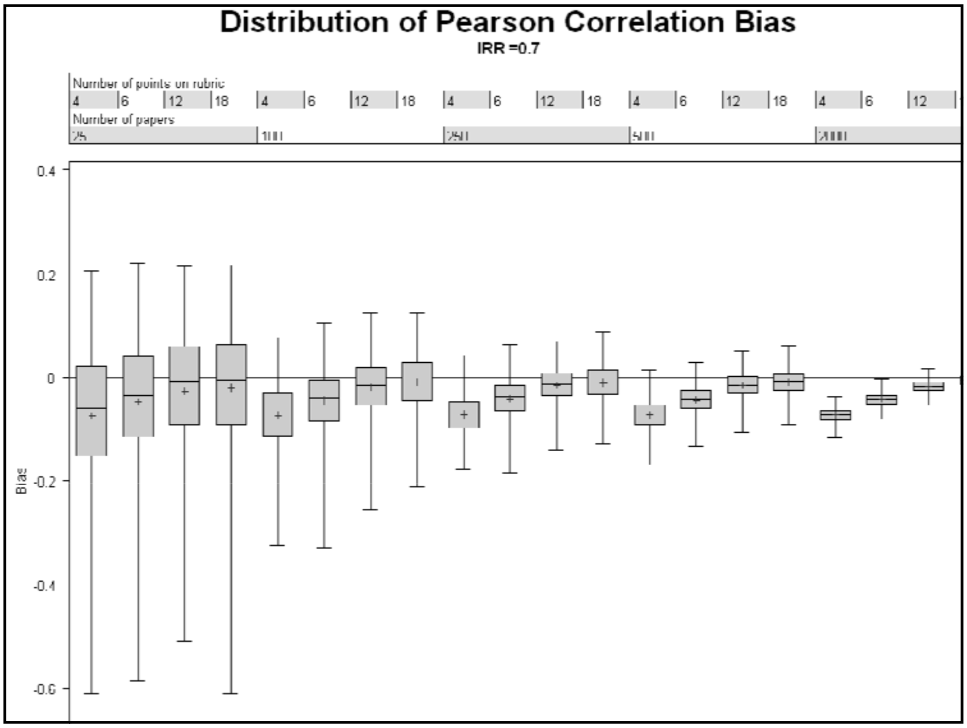


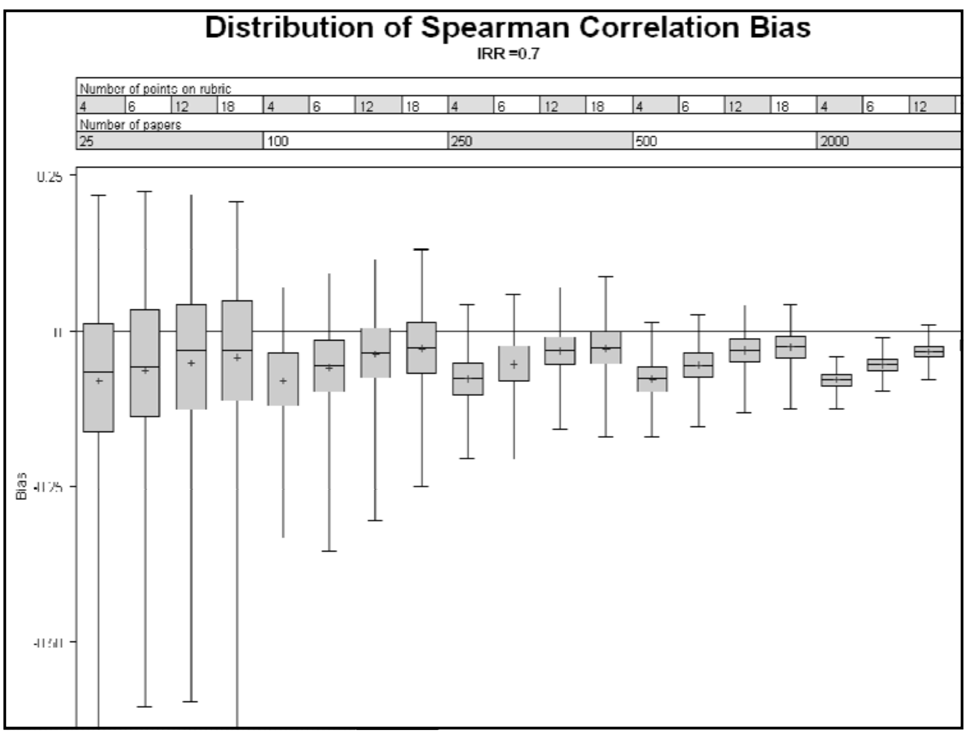
41

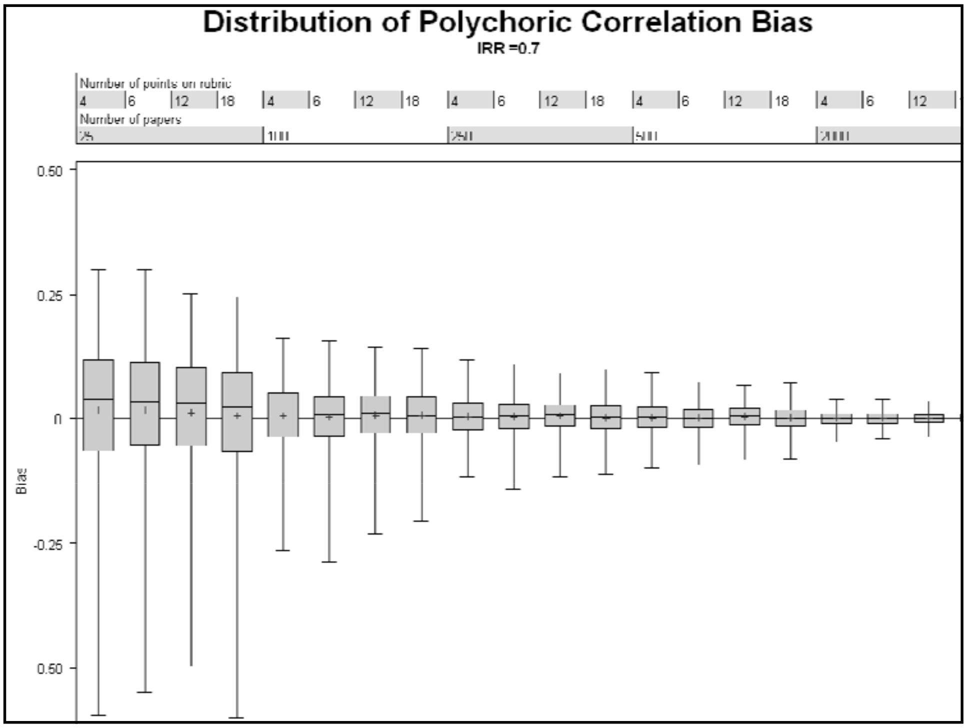
Estimated Bias by Reliability Parameter

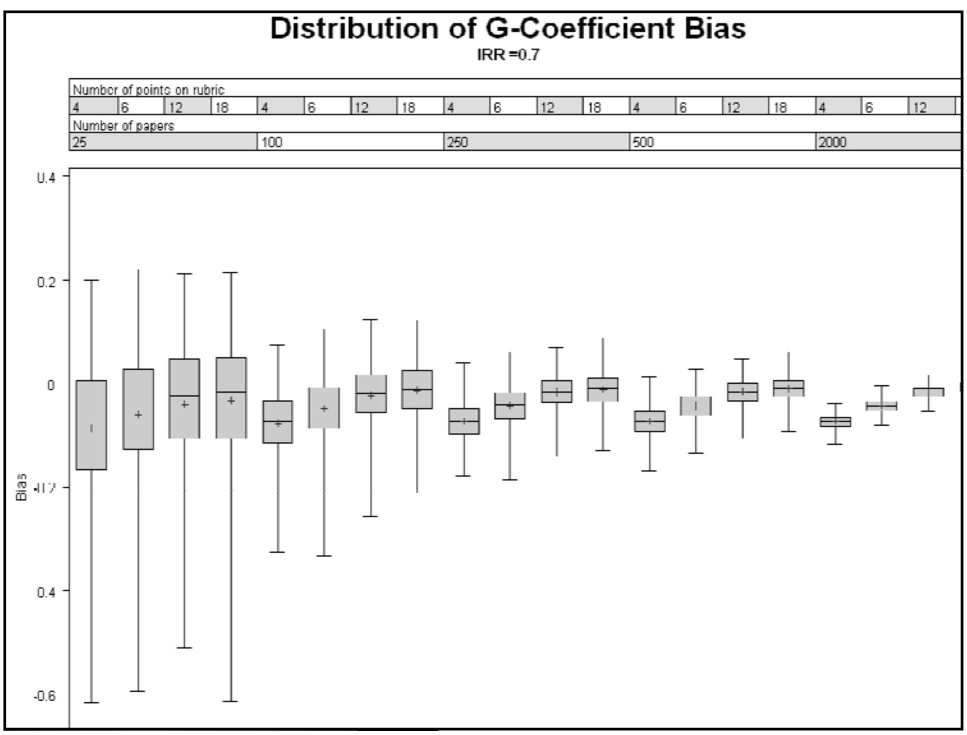


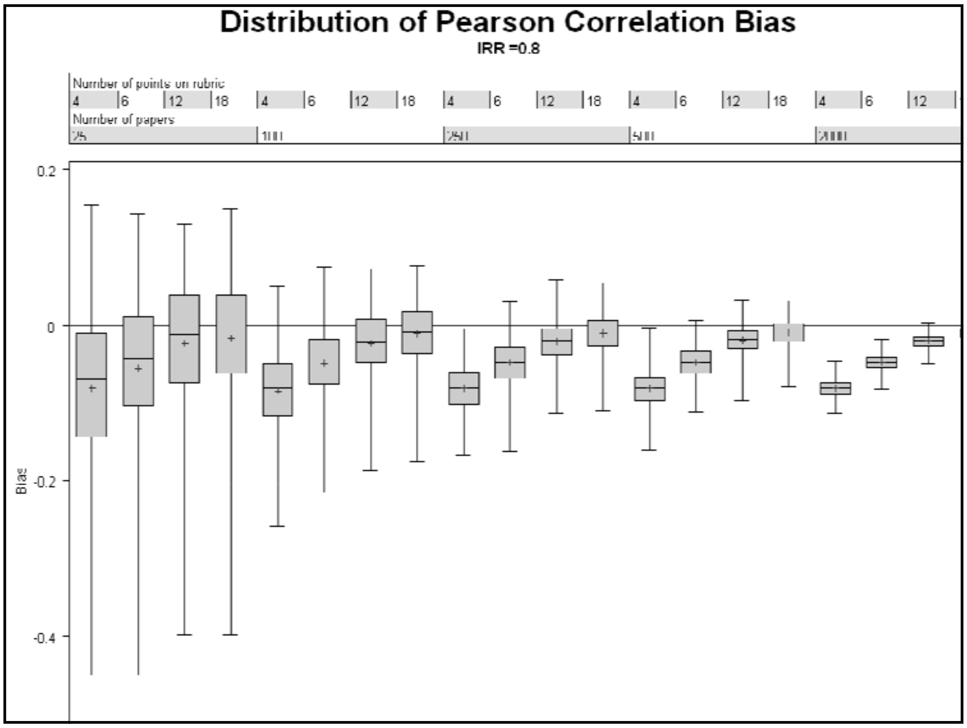
42

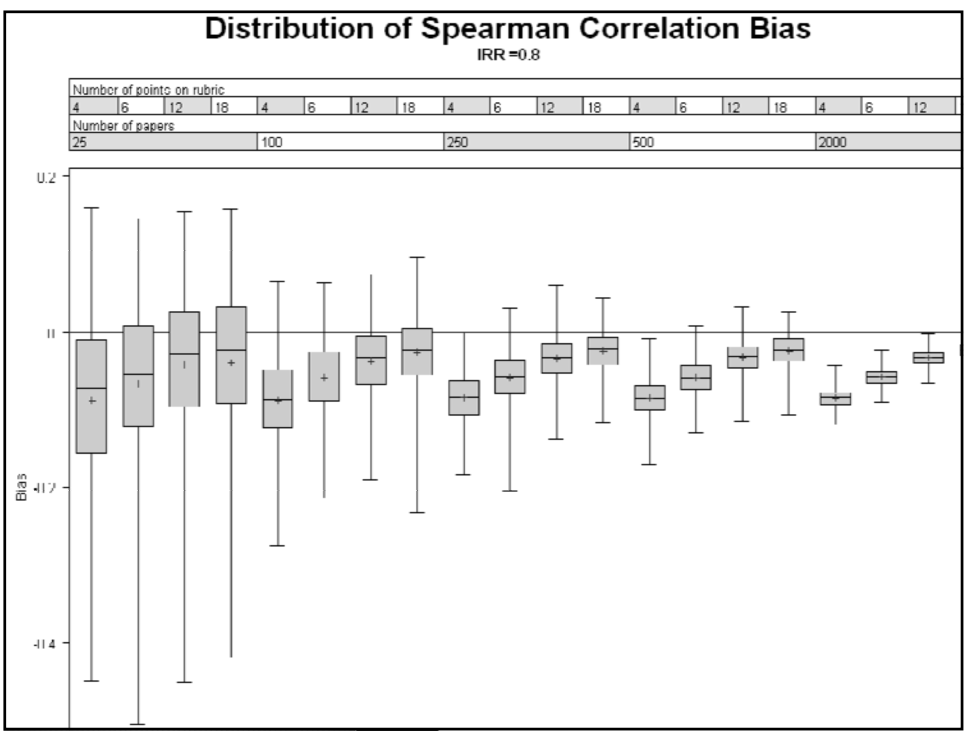

Distribution of Pearson Correlation Bias
IRR =0.7
Number of points on rubric
4 6 12 18 4 6 12 18 4 6 12 18 4 6 12 18 4 6 12
Number of papers
25 100 250 500 2000
Bias
0.4
0.2
0
-0.2
-0.4
-0.6

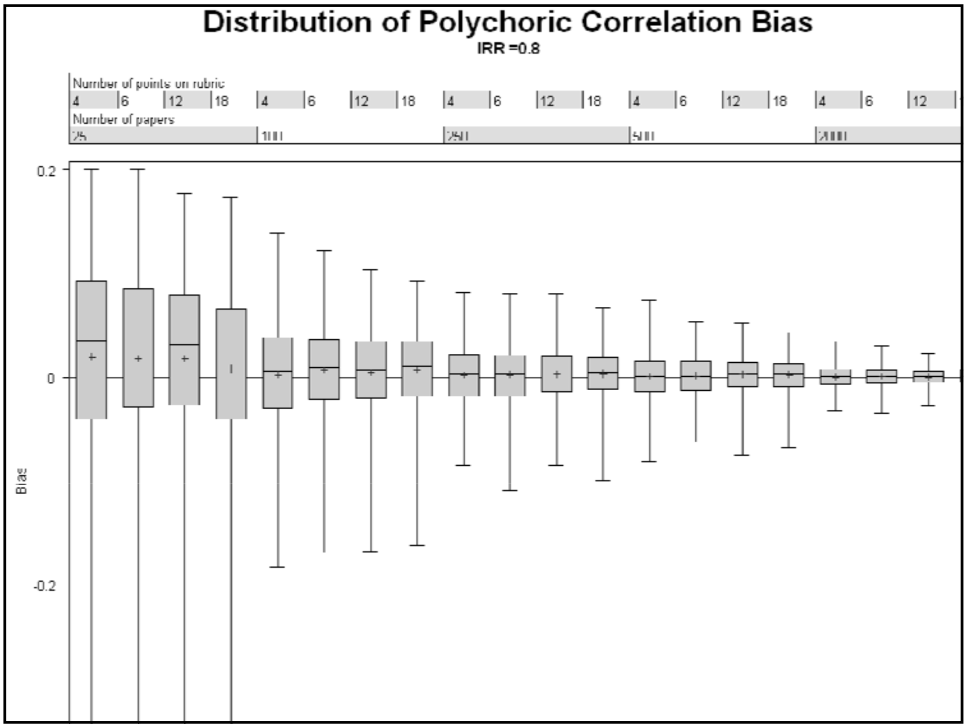

Distribution of Spearman Correlation Bias
IRR =0.7
Number of points on rubric
4 6 12 18 4 6 12 18 4 6 12 18 4 6 12 18 4 6 12
Number of papers
25 100 250 500 2000
Bias
0.25
0

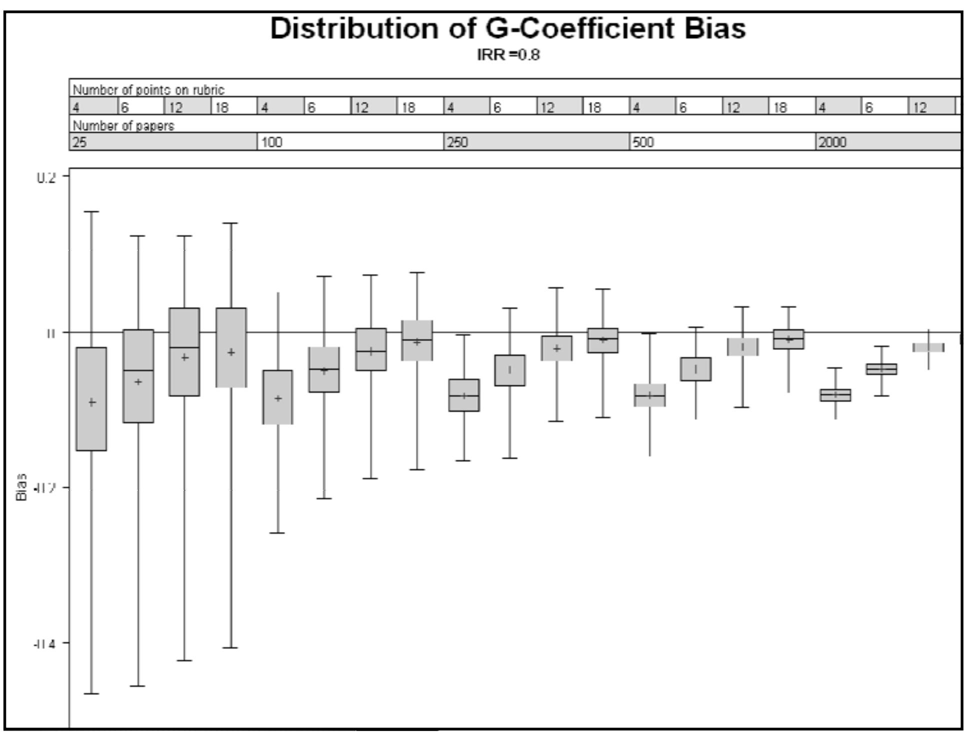

Distribution of Polychoric Correlation Bias
IRR =0.7
Number of points on rubric
4
6
12
18
Number of papers
25
100
250
500
2000
0.50
0.25
0
Bias
-0.25
0.50

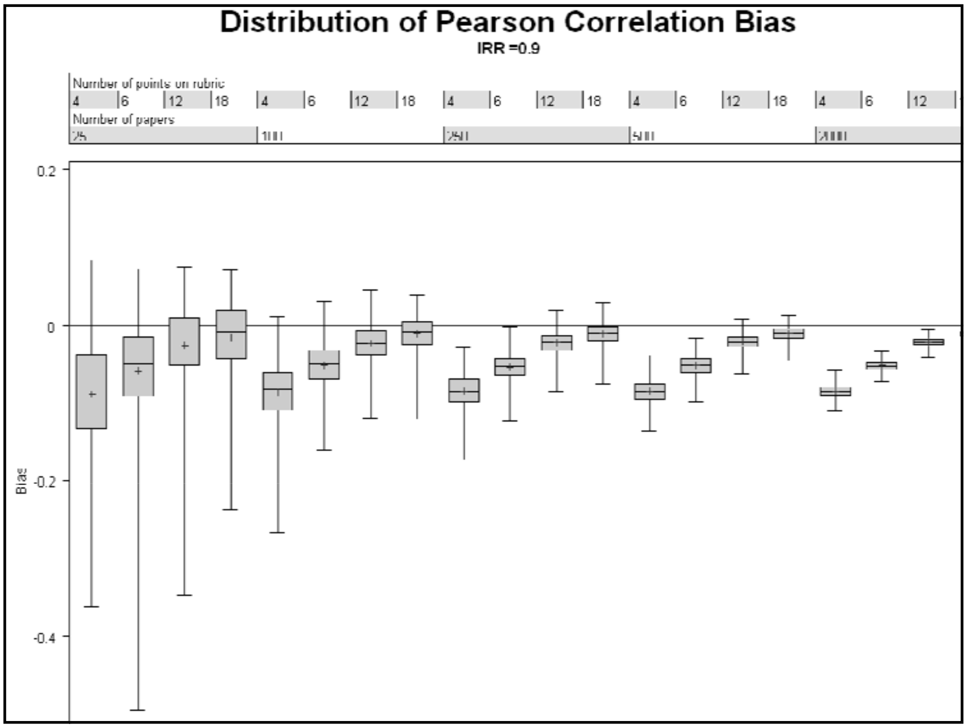

Distribution of G-Coefficient Bias
IRR =0.7
Number of points on rubric
4
6
12
18
Number of papers
25
100
250
500
2000
0.4
0.2
0
Bias
0.4
-0.6

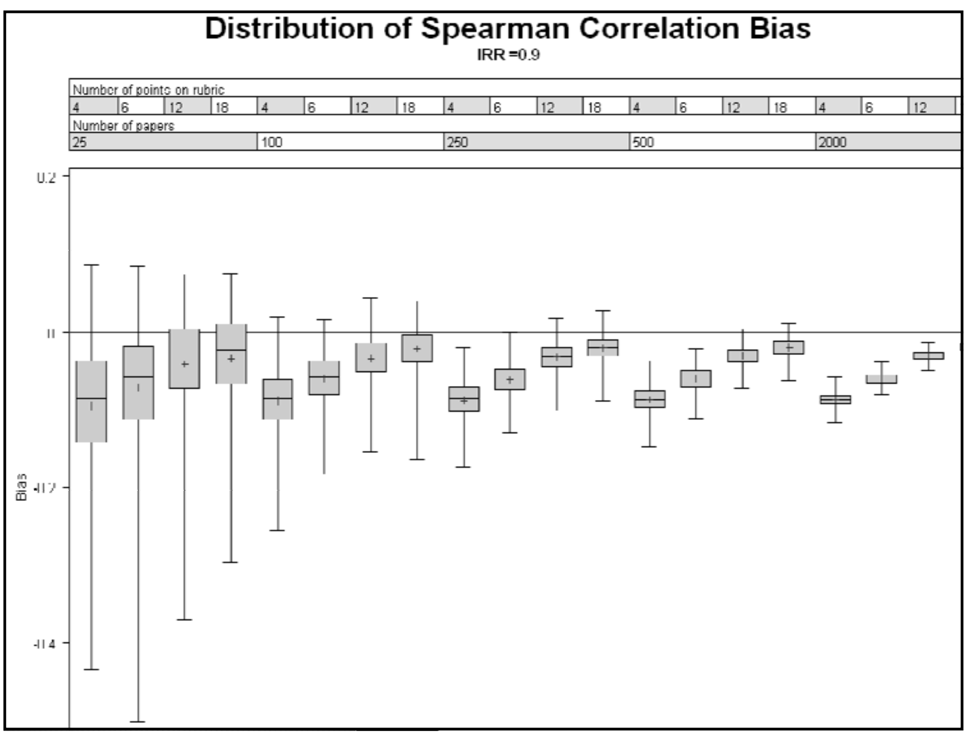

Distribution of Pearson Correlation Bias
IRR =0.8
Number of points on rubric
Number of papers
Bias

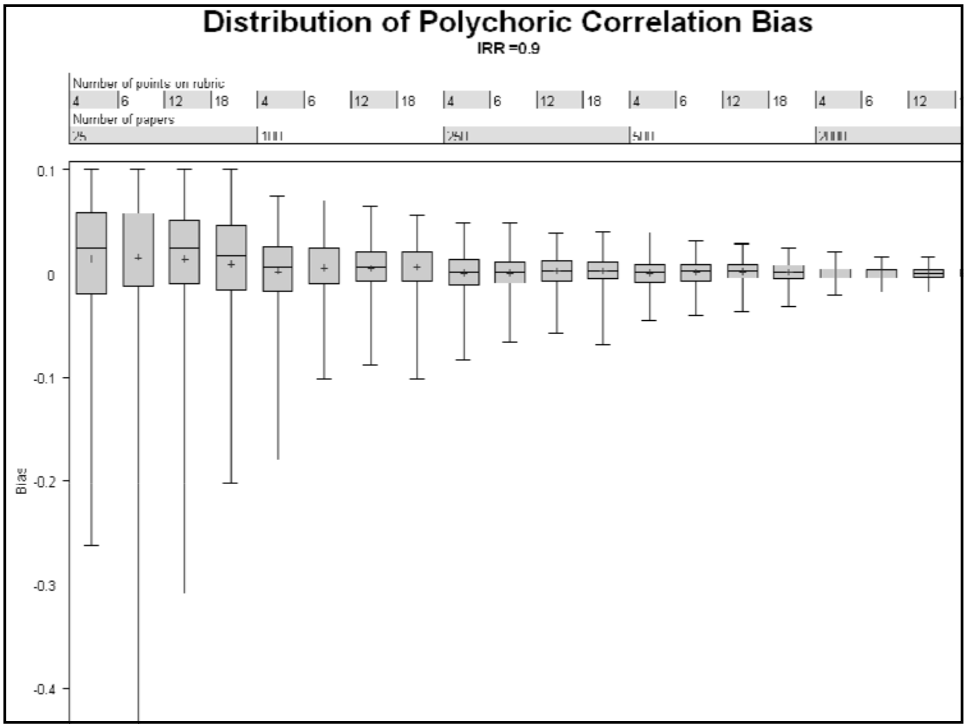

Distribution of Spearman Correlation Bias
IRR =0.8
Number of points on rubric
Number of papers
25
100
250
500
2000
Bias

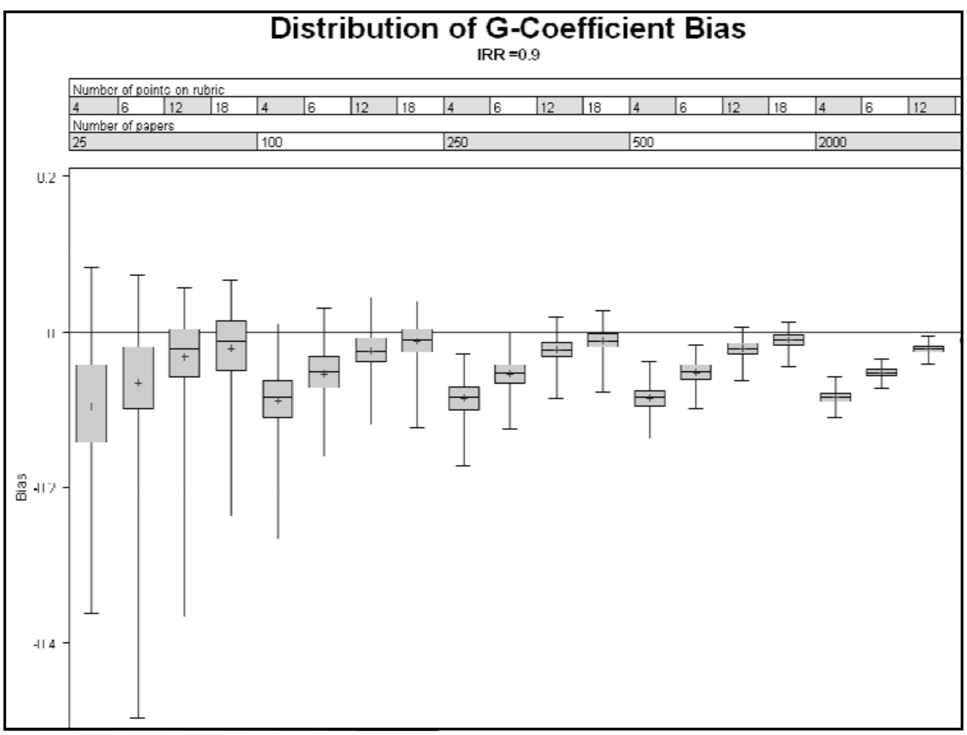

Distribution of Polychoric Correlation Bias
IRR =0.8
Number of points on rubric
Number of papers
Bias


Distribution of G-Coefficient Bias
IRR =0.8
Number of points on rubric
Number of papers
Bias


Distribution of Pearson Correlation Bias
IRR =0.9
Number of points on rubric
Number of papers
Bias


Distribution of Spearman Correlation Bias
IRR =0.9
Number of points on rubric
Number of papers
Bias


Distribution of Polychoric Correlation Bias
IRR =0.9
Number of points on rubric
Number of papers
Bias


Distribution of G-Coefficient Bias
IRR =0.9
Number of points on rubric
Number of papers
Bias

Accuracy of Estimates

Estimate	Mean	SD	Median	Min	Max
Pearson	–.04	.05	–.03	–.63	.22
Spearman	–.05	.06	–.05	–.77	.22
Polychoric	.00	.05	.00	–.71	.30
G-Coeff.	–.04	.05	–.04	–.64	.22

On average, all estimates were very close to the simulated parameter

Which measure is "appropriate"?

1) Ease of communication
   - Pearson product-moment correlation coefficient

2) Alignment between analysis and data
   - Polychoric correlation coefficient

3) Accuracy of estimates
   - Polychoric correlation coefficient

Conclusions

- Estimates approach simulated parameter as the number of scale categories increase.
- Range of coefficients decreases only slightly as scale categories increase.
- All coefficients become more precise as numbers of papers increase.

Conclusions

- Pearson tended to <u>underestimate</u> reliability *across conditions.*
- Spearman tended to <u>underestimate</u> reliability *across conditions.*
- G-coefficient tended to <u>underestimate</u> reliability *across conditions.*
- Polychoric tended to <u>overestimate</u> reliability *when the number of papers is smaller.*

Two Questions Answered

1) Should I use scale augmentation?
   When feasible, yes. Scale augmentation provides estimates closer to the parameter although there is not a major benefit for polychoric correlation.

2) How many papers (i.e., ratings) do I need to get good estimate of consistency?
   It depends on definition of "good" (i.e., one's desired level of confidence). Increasing the number of ratings increases precision. If one has a limited number of papers, polychoric correlation provides the least biased estimate on average.

NOTE: These answers based on results of this simulation. Generalizations to other conditions are not possible.

To Calculate Reliability Coefficients

- Pearson, Spearman, and Polychoric correlation coefficients
  - This study used SAS PROC FREQ with the PLCORR option on the TABLE line.
  - Mplus, R, PRELIS, SPSS also provide these estimates.
- G-coefficients & Phi-coefficients
  - This study used SAS PROC GLM (VARCOMP is also available in SAS).
  - SPSS, MATLAB
  - Specialized software
    - GENOVA, EduG

Using SAS PROC FREQ

- Set up data so each rater represents a column

	paper	rating1	rating2
1	1	3	2.67
2	2	3.67	3.67
3	3	3	3
4	4	2.67	2.33
5	5	2.67	3
6	6	3.33	3.33
7	7	3.33	3.33
8	8	3	2.67
9	9	4.67	4.33
10	10	4.33	4.33
11	11	4	4.67
12	12	3	3.67
13	13	3.33	3.33
14	14	4	4
15	15	1.67	2

Using SAS PROC FREQ

```
proc freq data=aeademo;
table rating1*rating2 / plcorr;
run;
```

Using SAS PROC FREQ

The FREQ Procedure

Statistics for Table of rating1 by rating2

Statistic	Value	ASE
Gamma	0.9065	0.0313
Kendall's Tau-b	0.8121	0.0373
Stuart's Tau-c	0.7760	0.0451
Somers' D C|R	0.8159	0.0389
Somers' D R|C	0.8083	0.0376
Pearson Correlation	0.9187	0.0220
Spearman Correlation	0.9032	0.0303
Polychoric Correlation	0.9398	0.0201
Lambda Asymmetric C|R	0.3500	0.0945
Lambda Asymmetric R|C	0.3684	0.0783
Lambda Symmetric	0.3590	0.0766
Uncertainty Coefficient C|R	0.5035	0.0447
Uncertainty Coefficient R|C	0.5022	0.0439
Uncertainty Coefficient Symmetric	0.5028	0.0433

Sample Size = 50

Evaluation 2010 63

Using PROC VARCOMP

- Set up data so every rating has its own row and is classified by paper and by rater

	paper	rater	rating
1	1	1	3
2	1	2	2.67
3	2	1	3.67
4	2	2	3.67
5	3	1	3
6	3	2	3
7	4	1	2.67
8	4	2	2.33
9	5	1	2.67
10	5	2	3

Evaluation 2010 64

Using PROC VARCOMP

```
proc varcomp data=aeademo2;
class paper rater;
model rating=paper|rater;
run;
```


Using PROC VARCOMP

Variance Components Estimation Procedure

Class Level Information

Class	Levels	Values
paper	50	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
rater	2	1 2

Number of Observations Read	100
Number of Observations Used	100

MIVQUE(0) SSQ Matrix

Source	paper	rater	paper*rater
paper	196.00000	0	98.00000
rater	0	2500.0	50.00000
paper*rater	98.00000	50.00000	99.00000
Error	98.00000	50.00000	99.00000

MIVQUE(0) SSQ Matrix

Source	Error	rating
paper	98.00000	1013.4
rater	50.00000	0.50000
paper*rater	99.00000	528.19000
Error	99.00000	528.19000

MIVQUE(0) Estimates

Variance Component	rating
Var(paper)	4.95102
Var(rater)	-0.0085714
Var(paper*rater)	0.13857
Var(Error)	0

Using PROC VARCOMP

- Using the estimates from previous slide to estimate the G coefficient for one rater:

$$G = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{p*r}^2}{k}} \qquad G = \frac{4.95102}{4.95102 + \frac{0.43857}{2}} = .958$$

Using PROC GLM (and 37 lines of code):

```
The MEANS Procedure

Analysis Variable : gencoef

     Mean
---------
0.9575875
---------
```


Future Research

- Need for more conditions
- Examinations of additional estimates
- Examine Winsorized distributions

For more information…

Grant Morgan – morgang@mailbox.sc.edu

Dr. Robert Johnson – rjohnson@mailbox.sc.edu

Min Zhu – zhum@mailbox.sc.edu

Vasanthi Rao – raov@mailbox.sc.edu

References

Anderson, J. C. & Gerbing, D. W. (1988) Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411-423.

Bentler, P.M., & Lee, S.Y. (1983). Covariance structures under polynomial constraints: Applications to correlation and alpha-type structural models. *Journal of Educational Statistics, 8*, 207-222.

Joreskog, K. G. (1994) On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59*(3), 381-389.

Linacre, J.M. (2005). Correlation coefficients: Describing relationships. *Rasch Measurement Transactions, 19*(3), 1028-1029.