Why Should We Be Worried About Rubric Reliability?

Lisa A. Pufpaff, Ph.D., Laura Clarke, Ed.D
& Ruth Jones, Ed.D
Ball State University, Muncie IN

Rationale

- Performance-based assessments for accreditation and accountability rely on rubric-based evaluations of performance
- The literature contains little evidence about reliability of rubric outcomes
- There are no mandates to evaluate inter-rater reliability of rubric outcomes
- Yet rubric outcomes are used to:
  - Grade students
  - Determine teacher candidate status
  - Evaluate faculty teaching
  - Evaluate courses
  - Evaluate programs

- Given the extent to which rubrics are used to evaluate candidate performance in teacher preparation programs, the issue of inter-rater reliability and effects of rater training must be explored (Reddy & Andrade, 2010)

Research Questions

- What is the inter-rater reliability on three different performance assessments across multiple, untrained raters?
- What is the effect of rater training on inter-rater reliability?
- What is the social validity of rater training materials?

Procedures – Phase 1

- Selected 3 different assessments completed across the first three years of an undergraduate special education teacher preparation program
- Identified 2 different exemplars from each of the 3 assessments
- The exemplars had varying rubric row and overall outcomes from one another (as evaluated by course instructor = original rating)

Assessments

- Digital portfolio completed freshman year during *Introduction to Special Education*; rubric consisted of 7 rows
- Research paper completed sophomore year during *Special Education Law*; rubric consisted of 4 rows
- Case study completed junior year during *Assessment in Special Education*; rubric consisted of 6 rows
- All rubrics evaluated performance across four levels: Unsatisfactory, Basic, Proficient, Distinguished

Participants

- 10 faculty within the Department of Special Education volunteered to rate assessments
- Raters included tenured, tenure-line, and full-time contract faculty
- Years of experience in their current position ranged from 0-32
- Level of experience using, creating, and analyzing data from rubrics varied with 3 experts, 6 experienced, and 1 novice
- Raters were not given assessments for courses that they typically taught

Method

- Each rater was given 6 assessments (2 exemplars each from the 3 assessments) and their corresponding rubrics
- No directions were provided beyond using the rubrics to evaluate the assessments
- Feedback was solicited on the assessments, the rubrics, and the process

Procedures - Phase Two

- Two types of rater training materials were created
  - Expanded Rubric – information was added to the original rubric:
    - Basic requirements of the assignment
    - Knowledge, skills, dispositions and/or performances from the professional standards aligned to the rubric
    - Directions provided to students
    - Definitions of terms
  - Narrated PowerPoint – PowerPoint presentation with imbedded audio narration explaining the Expanded Rubric

Before and After Expansion

	Unsatisfactory	Basic	Proficient	Distinguished
Recom-mend-ations for instruction ; Section IX	Recommend-ations are missing or poorly developed, do not reflect the evaluation data, or are not appropriate to the strengths and needs of the student.	Recommend-ations reflect data but are general or not appropriate to the strengths and needs of the student.	Recommend-ations reflect data and are appropriate to the strengths and needs of the student.	Recommend-ations reflect data, are appropriate to the strengths and needs of the student, and are prescriptive.

Recommendations for instruction; Section IX	Recommendations are missing or poorly developed, do not reflect the evaluation data, or are not appropriate to the strengths and needs of the student.	Recommendations reflect data but are general or not appropriate to the strengths and needs of the student	Recommendations reflect data and are appropriate to the strengths and needs of the student	Recommendations reflect data, are appropriate to the strengths and needs of the student, and are prescriptive.
Section IX, Recommendations, should be practical and directed to both teacher and parent. Recommendations should be based on strengths (need for enrichment) AND on growth areas (need for intervention and/or remediation). Writers are instructed to give specific activities, websites, etc. to demonstrate possible instructional strategies. Recommendations must be numbered, organized (school/home; skill area, etc.), linked to test results, and prioritized. Vocabulary: Prescriptive: The recommendations should be targeted directly to the need or strength and at the appropriate level. For example, if the child exhibits reading comprehension deficits, the cause should be identified and addressed (fluency, literal comprehension, inferential comprehension, etc.).			Rater Comments:	

2013 Assessment Institute
Pufpaff, Clarke & Jones

11

Script of narration to accompany previous rubric row

"The final rubric row is dedicated to **Section IX, Recommendations**. Recommendations are to be clearly tied to data from previous sections. They should reflect strengths and weaknesses as identified in specific assessments. Prescriptive recommendations are the end of the breadcrumb trail so to speak that has been laid in succeeding information. They are targeted with regards to identified specific skill weaknesses and offer appropriate suggestions for remediation and practice."

2013 Assessment Institute
Pufpaff, Clarke & Jones

12

Participants

- The same raters were used for Phase Two

Method

- Raters were directed to view the Expanded Rubric and listen to the Narrated PowerPoint Presentation
- Raters were presented 6 new assessments (2 different exemplars each from the same 3 assessments) and their corresponding rubrics
- Feedback was solicited on the assessments, the rubrics, the process, and the training materials

Results

- Data were analyzed by calculating the percent of agreement for the most commonly given score per rubric row (i.e., number of raters who chose the most commonly given score divided by total number of raters).

Digital Portfolio

#1 & 2 Pre-training Compared to #3 & 4 Post-training

Rubric Row	Exemplar #1	Exemplar #2	Exemplar #3	Exemplar #4
Reflection	B (55%)	B (64%)	B (30%) P (30%)	P (60%)
Rationale	B (64%)	B (36%) P (36%)	U (70%)	B (60%)
Design	B (73%)	P (45%)	B (30%) P (30%)	B (40%)
Environment	B (64%)	B (64%)	B (60%)	P (40%) D (40%)
Mechanics	B (45%)	P (55%)	B (50%)	D (50%)
Professionalism	B (45%)	B (64%)	B (60%)	P (60%)
Overall	B (64%)	B (45%) P (45%)	U (30%) B (30%)	P (60%)
Average Agreement	55%	55%	50%	50%

Research Paper

#1 & 2 Pre-training Compared to #3 & 4 Post-training

Rubric Row	Exemplar #1	Exemplar #2	Exemplar #3	Exemplar #4
Row A	B 36(%)	U (36%) B (36%)	B (36%)	B (45%)
Row B	B (36%) P (36%)	B (36%)	U (45%)	P (45%)
Row C	B (45%)	U (55%)	B (45%)	B (55%)
Overall	B (45%)	B (55%)	B (36%)	P (36%)
Average Agreement	41%	45%	41%	45%

Case Study

#1 & 2 Pre-training Compared to #3 & 4 Post-training

Rubric Row	Exemplar #1	Exemplar #2	Exemplar #3	Exemplar #4
Row 1	P (80%)	B (40%)	D (45%)	P (36%)
Row 2	P (40%)	B (60%)	D (64%)	P (73%)
Row 3	B (50%)	B (80%)	D (64%)	P (36%)
Row 4	D (80%)	B (60%)	D (82%)	B (45%)
Overall	P (60%)	B (60%)	D (64%)	P (55%)
Average Agreement	62%	60%	64%	49%

Discussion & Implications

- Faculty rated six exemplars of three different assignments using assessment rubrics.
- Faculty repeated ratings of additional, equivalent exemplars after participating in rater training.
- Inter-rater reliability was compared for each rubric before and after training.
- Descriptive statistics revealed negligible improvement in inter-rater agreement after training.

Results could be attributed to several variables that should guide future research in this area:

- Faculty may have overestimated their competence and not processed the training materials carefully.
- Controlling the environment may be of value, i.e., ask rater to complete the ratings and/or trainings in a quiet room, all at one time, free of distractions and interruptions.
- Results may be clearer if fewer exemplars are evaluated by each rater, resulting in reduced time commitment for participants.

- Rater feedback indicated that concrete examples at each rubric level may have been instructive, and that some rubric rows appeared to be ambiguous (even with the training).
- Rubrics may need to be revisited to improve their objective measurement of student performance.
- Alternative rater training materials may be more appropriate.

Conclusion

- As the call for increased accountability promotes reliance on rubric-based outcomes of student learning, it is imperative for teacher educators to utilize rubrics wisely, fairly, and reliably.
- This study only emphasizes the need for research into how to more effectively utilize these assessment and accountability tools.

References & Contacts

- Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448.

- The authors may be contacted at
  - Lisa Pufpaff lapufpaff@bsu.edu
  - Laura Clarke lsclarke@bsu.edu
  - Ruth Jones rejones@bsu.edu