Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 14, Number 11, May 2009 ISSN 1531-7714

Combining Dual Scaling with Semi-Structured Interviews to Interpret Rating Differences

Ruth A. Childs, Anita Ram, & Yunmei Xu

Ontario Institute for Studies in Education, University of Toronto

Dual scaling, a variation of multidimensional scaling, can reveal the dimensions underlying scores, such as raters' judgments. This study illustrates the use of a dual scaling analysis with semi-structured interviews of raters to investigate the differences among the raters as captured by the dimensions. Thirty applications to a one-year post-Bachelor's degree teacher education program were rated by nine teacher educators. Eight of the raters were subsequently interviewed about how they rated the responses. A three-dimensional model was found to explain most of the variance in the ratings for two of the questions and a two-dimensional model was most interpretable for the third question. The interviews suggested that the dimensions reflected, in addition to differences in raters' stringency, differences in their beliefs about their roles as raters and about the types of insights that were required of applicants.

Whenever more than one person rates a response, there is opportunity for disagreement; indeed, numerous studies (e.g., Conway, Jako, & Goodman, 1995; Hoyt & Kerns, 1999) have showed that raters in a wide variety of contexts, even when provided with training and with detailed rubrics, rarely agree perfectly. This study illustrates the use of a dual scaling analysis of ratings, combined with semi-structured interviews of the raters, to investigate patterns of agreement and disagreement among a group of raters and to suggest reasons for the disagreements. The data are from a study of teacher educators rating applications to an initial teacher education program.

The study of rater accuracy and agreement is hardly new. In fact, such phenomena as the halo effect and leniency-severity have been studied since the 1920s (Saal, Downey, & Lahey, 1980). More recently, systematic differences in raters' judgments of handwritten versus computer-printed documents and the effect of training on those differences have been examined by Russell and Tao (2004a, 2004b). Many studies have quantified the agreement among raters (interrater reliability; see Stemler, 2004, for a review of methods) and investigated ways to minimize disagreement (see Rudner, 1992, for a summary). Fewer studies have directly investigated raters' accuracy, as external criteria are not often available; however, many researchers have posited rater biases (including the halo effect and leniency or stringency) as indirect evidence that many raters are systematically inaccurate.

The difficulty is that raters are both essential to a rating process and intractably idiosyncratic. Hoyt (2000) summarized the problem succinctly: "Raters may interpret scale items differently or have unique reactions to particular targets so that the obtained ratings reflect characteristics of the raters to some extent, in addition to reflecting the target characteristics that are of interest" (p. 64).

Numerous researchers have investigated how to minimize the disagreement among raters or, if that fails, the effect of the disagreement. Two meta-analysis studies are particularly notable. The first, by Conway,

Jako, and Goodman (1995), analyzed the interrater reliability of interview raters, including interviews for jobs and for admission to academic programs. The data were drawn from 82 sources. They found that interviewer training contributed to higher interrater reliability, as did requiring raters to rate each question separately, rather than make a holistic rating.

The second meta-analysis was performed by Hoyt and Kerns (1999), who analyzed generalizability studies of 79 datasets involving ratings of essays, performances, and clinical assessments. They found that, on average, about a third of the variance could be explained by the rater and rater by trait effects, but that these effects could be significantly reduced by increasing rater training and by making the required judgments less subjective. However, because they found that even highly trained raters differed significantly in their ratings, Hoyt and Kerns concluded that combining ratings from multiple raters is the best way to reduce the effect of rater differences.

Another approach to minimizing the effects of rater differences was suggested by Lunz, Stahl, and Wright (1994). In a study of ratings of student portfolios, they argued for statistically adjusting ratings based on analyses of rater patterns (they used the Rasch-based Facets analysis), because, while they supported training of raters, they also believed that raters "are unique and will remain unique regardless of the amount of training and grading experience acquired" (p. 924).

In the literature on rater differences, attempts to understand the raters' perspectives are surprisingly rare. A recent exception is Murphy, Cleveland, Skattebo, and Kinney's (2004) investigation of whether course evaluations were influenced by students' goals in providing the evaluations. Studying students in five university courses, they found that differences in the goals students cited for providing the ratings (e.g., to rate the instructor fairly, to improve the instructor's confidence, to identify areas where the instructor needs more training) accounted for a small but significant proportion of the differences in their ratings.

The preceding brief review of the rating literature suggests that disagreement among raters is very common. The purpose of this study is to explore a way to interpret these differences among raters. Using ratings of applications to an initial teacher education program as an example, we combined a dual scaling (Nishisato, 1994) analysis with semi-structured interviews of the raters.

It is no surprise that there should be disagreement among raters of the responses of applicants to an initial teacher education program. As Fenstermacher and Richardson's (2005) discussion of the importance of distinguishing good teaching from successful teaching illustrates, there has been little consensus among educators about what precisely it means to be a good teacher. Defining the experiences, insights, and attitudes that are needed by an applicant to an initial teacher education program is likely to be even more contentious. It is hardly surprising that training and detailed rubrics do not result in complete agreement among raters who may have very different beliefs about teaching and teacher education.

METHOD

All applicants for September 2008 admission to a large one-year post-Bachelor's degree teacher education program were required to provide a three-part written profile in the Fall of 2007. Admission to the program is highly competitive: In the year studied, almost 5,500 applications were received for fewer than 1,300 spots. The first part of the profile asked applicants to describe three experiences that had helped them prepare for a career as a teacher and what they learned from one of those experiences; the second part asked applicants to describe their social background and experiences that have prepared them to work with diverse students and families; the third part asked them to describe an experience of advantage or disadvantage and what they learned from that experience that prepared them to work with students and families. The questions are provided in Appendix A.

Applicants' responses to each part were rated on a three-point scale – INSUFFICIENT EVIDENCE, PASS, and HIGH PASS – based on detailed rubrics (see Appendix B). All raters were instructors in the program or educators associated with the program (e.g., mentor teachers) and received four hours of training in the rating process and the use of the rubrics, plus a 33-page handbook about the rating process.

The applicants submitted their profiles through a secure on-line system and the profiles were presented to the raters in batches of 30 using a similar system. For this study, one randomly-selected batch of profiles from the Intermediate/Senior (Grades 7-12) program was evaluated by nine raters (instead of the usual two raters), selected at random from among those raters who were assigned to read profiles from that program and who had received their training by the beginning of the reading period. The batches themselves were created by

randomly drawing applications that had not yet been rated twice. All raters in this study were instructors in the teacher education program. The study ratings were completed during the regular rating period and the raters did not know they were in the study until after they had completed their ratings. Informed consent for the use of their ratings in this study was obtained from each rater after the completion of the ratings.

The ratings were analyzed using the dual scaling approach to modeling categorical data (Nishisato, 1994), as implemented in the DUAL3 computer program (Nishisato & Nishisato, 1998). Dual scaling is a variation of multidimensional scaling and was used in this study because it permitted us to fully explore the complex structure of the data – especially the disagreement among raters that could not be accounted for by differences in leniency-severity. Each rater's ratings on each part were converted to rankings, and then analyzed using the dual scaling method for rank order data.

For Parts 1, 2, and 3 separately, three solutions (analogous to dimensions) were extracted (a fourth solution was also extracted, but accounted for very little variance, so is not reported here). Unlike in factor analysis, the extraction of additional solutions does not affect the weights of the first solutions, so that it is possible to choose not to interpret later solutions; this decision is typically based on the relative percentages of variance accounted for (particularly where the percentage of variance decreases dramatically between solutions) and the interpretability of the solutions. Based on the former criterion, three solutions were provisionally chosen. As recommended by Nishisato (1994), for each solution, the raters' normed weights and the profiles' projected weights (i.e., the normed weights multiplied by the solution's maximum correlation) were plotted. Dual scaling was chosen for these analyses because the small number of rating levels (3) limited the usefulness of generalizability theory approaches. In addition, the design of the profile, with each of the three questions designed to measure very different constructs, made a scaling approach such as Facets analysis inappropriate.

The results of the dual scaling revealed complex patterns in the ratings. To help us understand the dimensions, we contacted the nine raters to request follow-up structured interviews. Eight of the raters consented to be interviewed and for their interviews to be used as part of this study. Each rater was interviewed for approximately an hour. The raters were asked detailed questions about how they interpreted the rubrics and what they were looking for in the responses. They were also asked to think aloud as they re-rated three of the 30 profiles in the study batch; these profiles were selected because of the wide disagreement among the raters on their original ratings. Most raters preferred to have their responses summarized by the interviewer (the first author) in notes taken during the interview, rather than being tape-recorded. The notes from each interview were typed up shortly after the interview, along with summaries.

Some of the dimensions were easily interpretable based on their relationship to the mean rating received by each profile or the mean rating given by each rater. For each remaining dimension, we ordered the raters by their placement on that dimension and reviewed their interview responses for patterns of increasing or decreasing attention to particular features of the profiles or systematic differences in the characteristics they associated with strong and weak profiles.

RESULTS

Table 1 provides the average rating given by each rater across profiles and the average rating received by each profile across raters. For the purpose of these analyses, ratings of INSUFFICIENT EVIDENCE, PASS, and HIGH PASS were coded 1, 2, and 3, respectively.

Part 1. Experience

From Table 1, it is clear that Rater R8 gave, on average, the highest ratings on Part 1, while Rater R5 gave the lowest ratings. Profile IS19 received, on average, the lowest ratings on Part 1, while Profile IS14 received the highest average rating.

The three dimensions extracted for Part 1 account for 30.5%, 17.5%, and 15.2%, respectively, of the variance among the profiles and raters, for a total of 63.2% (additional dimensions did not account for significant amounts of the variance).

Figures 1a and 1b show the distribution of both the raters (R1 to R9) and the profiles (IS1 to IS30) in relation to the three dimensions for Part 1 (analogous figures can be created for Parts 2 and 3). For readability, only the raters, the highest and lowest rated profiles, and the three profiles that were used in the think-alouds are labeled. Both visual inspection of the plot and the correlation between the dimension weights and the mean ratings

Table 1: Mean Ratings and Dimension Weights for Each Rater and Each Profile

	Part 1				Part 2				Part 3			
	M (*SD*)	Dimension Weights			*M* (*SD*)	Dimension Weights			*M* (*SD*)	Dimension Weights		
		1	2	3		1	2	3		1	2	3
Raters												
R1	1.87(0.51)	1.03	-0.40	-0.64	1.77(0.63)	1.05	-0.44	-0.94	1.77(0.50)	-0.55	0.47	-2.35
R2	1.77(0.57)	1.09	1.10	0.33	1.57(0.73)	1.11	-0.74	0.30	1.50(0.57)	-1.29	-0.07	0.97
R3	1.90(0.66)	0.97	1.55	-0.28	1.50(0.57)	0.82	1.61	-0.95	1.70(0.70)	-1.02	1.43	0.73
R4	2.27(0.58)	0.08	1.34	1.66	2.23(0.73)	1.00	-0.30	2.12	2.30(0.60)	-1.02	-0.52	-0.22
R5	1.60(0.50)	1.33	-0.44	-0.85	1.40(0.50)	0.95	0.56	-0.89	1.40(0.50)	-0.95	-0.90	0.69
R6	2.17(0.59)	1.38	0.27	-0.40	1.73(0.58)	0.98	-0.59	0.35	1.87(0.51)	-0.40	1.75	0.73
R7	2.23(0.57)	0.57	-1.38	1.60	2.07(0.69)	0.96	-1.65	-0.86	2.10(0.61)	-0.96	-1.44	0.21
R8	2.43(0.57)	1.07	-0.59	1.44	2.13(0.78)	1.21	0.58	-0.14	2.33(0.92)	-1.27	-0.27	-0.59
R9	1.77(0.43)	0.81	-0.96	-0.40	1.70(0.53)	0.86	1.37	0.96	1.83(0.59)	-1.17	0.67	-0.75
Profiles												
IS1	1.89(0.78)	-0.19	0.30	0.34	1.44(0.53)	-0.29	0.16	0.08	1.78(0.83)	-0.10	-0.10	0.26
IS2	1.67(0.50)	-0.29	0.08	-0.17	1.33(0.50)	-0.40	-0.18	-0.09	1.22(0.44)	-0.53	0.24	-0.12
IS3	2.11(0.60)	0.24	-0.02	-0.20	2.33(0.50)	0.46	0.11	0.07	2.33(0.71)	0.39	-0.26	0.08
IS4	1.56(0.53)	-0.31	-0.03	-0.30	1.33(0.50)	-0.39	-0.01	0.11	1.44(0.53)	-0.39	-0.19	0.15
IS5	2.22(0.67)	0.20	-0.35	0.02	2.00(0.71)	0.19	-0.30	0.20	2.11(0.78)	0.19	0.28	0.14
IS6	2.00(0.00)	0.04	0.02	-0.21	1.67(0.50)	-0.09	-0.13	-0.16	1.11(0.33)	-0.64	-0.11	-0.20
IS7	1.78(0.97)	-0.34	-0.02	0.57	1.78(0.67)	0.00	-0.42	-0.11	2.22(0.67)	0.32	0.04	0.04
IS8	1.56(0.53)	-0.32	-0.14	-0.14	2.56(0.53)	0.62	-0.09	0.02	2.11(0.60)	0.28	0.27	0.01
IS9	2.44(0.53)	0.47	0.20	-0.09	1.78(0.44)	-0.01	0.12	-0.07	2.33(0.71)	0.41	-0.04	-0.38
IS10	2.11(0.60)	0.04	0.40	-0.01	1.33(0.50)	-0.36	-0.22	-0.20	1.22(0.44)	-0.57	-0.07	-0.18
IS11	2.00(0.71)	-0.11	0.45	0.09	2.22(0.44)	0.35	0.06	0.09	2.11(0.60)	0.23	-0.11	0.06
IS12	2.11(0.60)	0.21	-0.17	-0.07	1.22(0.44)	-0.47	0.22	0.03	2.11(0.33)	0.23	-0.06	-0.04
IS13	2.22(0.44)	0.19	0.19	-0.07	1.56(0.53)	-0.23	0.24	0.00	2.00(0.50)	0.07	-0.04	0.03
IS14	2.67(0.50)	0.56	0.03	0.22	2.44(0.53)	0.53	-0.05	-0.28	2.00(0.50)	0.12	0.08	0.03
IS15	2.11(0.33)	0.10	-0.13	-0.04	2.22(0.67)	0.33	0.09	0.27	2.22(0.44)	0.32	-0.11	0.02
IS16	2.00(0.71)	-0.09	0.19	0.29	2.44(0.53)	0.53	0.20	-0.06	2.22(0.67)	0.37	0.19	-0.30
IS17	2.22(0.67)	0.21	-0.12	0.18	2.44(0.53)	0.57	-0.04	-0.09	2.44(0.73)	0.42	-0.22	-0.08
IS18	1.89(0.33)	-0.05	-0.13	-0.18	1.44(0.73)	-0.30	-0.04	0.31	1.56(0.53)	-0.27	-0.13	0.01
IS19	1.44(0.53)	-0.53	-0.05	-0.01	1.33(0.71)	-0.39	0.13	0.39	1.44(0.53)	-0.29	0.13	0.24
IS20	1.78(0.83)	-0.37	-0.23	0.26	1.11(0.33)	-0.57	-0.11	-0.14	1.56(0.73)	-0.29	0.23	0.22
IS21	1.56(0.53)	-0.37	0.17	-0.23	2.00(0.50)	0.20	0.25	-0.17	1.44(0.53)	-0.37	0.02	0.08
IS22	1.78(0.44)	-0.23	-0.05	-0.15	1.00(0.00)	-0.66	0.03	-0.06	1.22(0.44)	-0.48	0.26	-0.10
IS23	2.11(0.33)	0.16	-0.05	-0.05	1.78(0.44)	-0.02	0.13	-0.06	1.78(0.67)	-0.13	-0.03	0.22
IS24	2.11(0.33)	0.05	0.16	-0.03	2.33(0.50)	0.49	-0.01	-0.02	2.33(0.71)	0.43	0.29	0.24
IS25	2.00(0.71)	0.01	-0.46	0.11	1.78(0.67)	0.00	0.11	0.03	2.11(0.78)	0.27	0.34	-0.23
IS26	2.11(0.60)	0.11	-0.32	0.08	2.00(0.50)	0.20	0.25	-0.17	2.00(0.71)	0.04	-0.26	0.10
IS27	1.67(0.50)	-0.34	-0.01	-0.08	1.22(0.44)	-0.48	0.07	-0.24	1.33(0.50)	-0.46	-0.12	-0.29
IS28	2.11(0.33)	0.18	0.04	-0.25	1.22(0.44)	-0.51	-0.13	0.01	1.78(0.44)	-0.15	-0.20	0.09
IS29	2.33(0.50)	0.36	-0.17	0.08	2.22(0.67)	0.40	-0.18	0.09	2.22(0.44)	0.26	-0.22	-0.11
IS30	2.44(0.53)	0.39	0.25	0.06	2.11(0.78)	0.29	-0.25	0.19	2.22(0.44)	0.32	-0.11	0.02
Overall	2.00(0.61)				1.79(0.69)				1.87(0.69)			
Variance Accounted For		30.51%	17.47%	15.17%		55.36%	10.63%	9.20%		45.81%	12.83%	10.93%
Correlation of Weight with *M* for Raters		-0.40	-0.06	0.80		-0.43	0.40	-0.45		-0.10	-0.20	-0.31
Correlation of Weight with *M* for Profiles		0.97	0.10	0.20		1.00	0.00	0.00		0.99	-0.10	0.03

Note. Weights for raters are normed weights; weights for profiles are projected weights.

reported in Table 1 confirm that Dimension 1 corresponds to the average ratings of the profiles (r = .97). Similarly, Dimension 3 is highly correlated (r = .80) with the stringency of the raters, as evidenced by the average ratings they assigned across profiles. Dimension 2, however, has very low correlations with both average profile rating (r = .10) and rater stringency (r = .06).

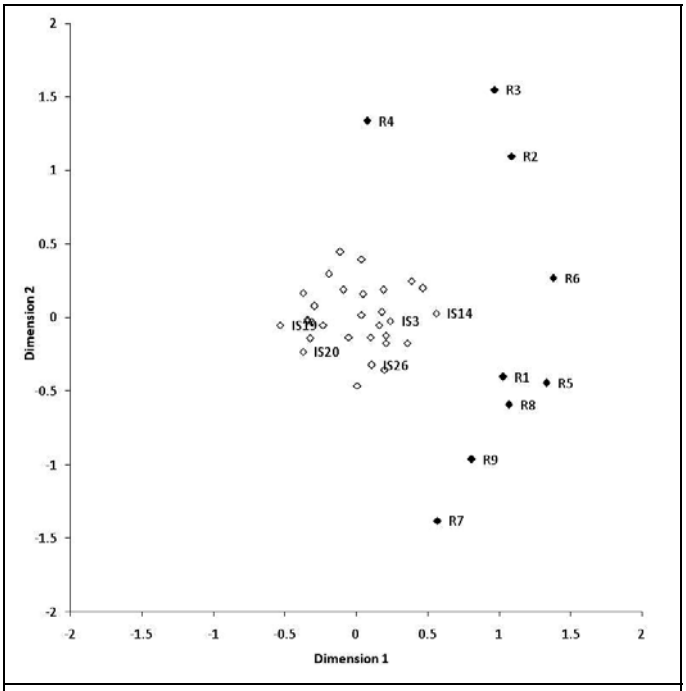

Figure 1a. Part 1, Dimensions 1 and 2.

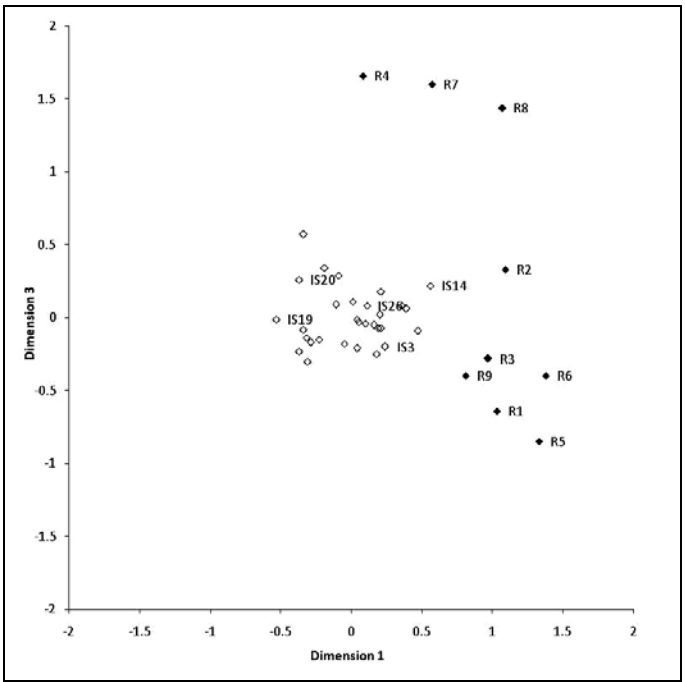

Figure 1b. Part 1, Dimensions 1 and 3.

To interpret Dimension 2, we ordered the raters by their placement on Dimension 2 and studied their interview responses for patterns. The ordering of raters on Dimension 2 corresponds best with the raters' beliefs about their role in the rating process. As Figure 1a shows, Raters R3, R4, R2, and R6 provided ratings in the upper half of the distribution. Of these raters, we interviewed Raters R3, R4, and R6, who shared in common a focus on whether the applicants answered the questions as intended. As Rater R6 put it, "If they can't read a question, they shouldn't be a teacher." In a sense, these raters viewed the profile as a test of the applicant's ability to understand and respond to questions about teaching and learning.

At the lower end of Dimension 2 are Raters R7 and R9. These raters have in common a belief that their role as raters is not to determine whether the applicant has followed the instructions and answered the question as intended, but to judge the applicant's suitability for teaching against the raters' own criteria. For example, Rater R7 was willing to give an answer a rating of PASS if he judged that the response showed critical thinking and good insights, even if it did not meet all the criteria for PASS outlined in the rubric. Of the seven raters we interviewed, Rater 9 read the most critically, comparing the content of the responses to his own extensive experience as a teacher. For example, when reviewing profile IS3, he alone noted that the activities the applicant reported doing (e.g., disciplining students) were probably inappropriate for a classroom volunteer, leading him to question the quality – and possibly the veracity – of the experience the applicant described.

Part 2. Diversity

On Part 2, Rater R4 gave the highest ratings and Rater R5 again gave the lowest ratings. Profile IS22 received the lowest rating, with all nine raters agreeing that the response showed INSUFFICIENT EVIDENCE of readiness for the program, while Profile IS8 received the highest average rating.

For Part 2, Dimension 1 accounts for 55.4% of the variance among profiles and raters, Dimension 2 for 10.6%, and Dimension 3 for 9.2%, for a total of 75.2%. The distribution of variance across the dimensions in Part 2 is less equal than for Part 1. In Part 1, Dimensions 2 and 3 together accounted for as much variance as Dimension 1; in Part 2, Dimensions 2 and 3 together account for less than half as much variance as Dimension 1.

For the profiles, the average ratings and the weights on Dimension 1 are perfectly correlated ($r = 1.00$). Both Dimensions 2 and 3 are moderately correlated with raters' stringency ($r = .40$ and -.45, respectively). A careful examination of the interview responses did not suggest any further interpretations of Dimensions 2 and 3.

Part 3. Equity and Social Justice

On Part 3, Rater R8 gave the highest ratings, followed closely by Rater R4; Rater R5 gave the lowest ratings. Profile IS6 received the lowest average rating and IS17 the highest average rating.

For Part 3, Dimension 1 accounts for 45.8% of the variance, Dimension 2 accounts for 12.8%, and Dimension 3 of 10.9%, for total of 69.5%. Again, Dimension 1 weights correspond almost perfectly to the average ratings of the profiles ($r = .99$). However, neither Dimension 2 nor Dimension 3 is clearly related to raters' stringency ($r = -.20$ and -.31, respectively). The interview responses of the raters suggest that Dimension 2 is related to the raters' interpretation of the question. Specifically, Part 3 required applicants to describe "a time when you or someone you know was advantaged or disadvantaged" and the impact of the experience, and then asked, "What did you learn from this experience that has prepared you to work with students and families who have experienced advantage or disadvantage?" The raters who were highest on this dimension were expecting applicants to make a clear connection to education. Both Raters 6 and 9, for example, expected applicants to describe what they would do as a teacher to address disadvantage. At the other extreme, Rater 7 described a response he felt was particularly strong as follows: "She really got the experience and what it felt like ... [the answer was] deeply reflective and honest." Rater 6 described this same response as "generic." We were not able to find a clear interpretation of Dimension 3.

DISCUSSION AND CONCLUSION

This study has several limitations. First, we selected the profiles for discussion in the interviews before we had completed the dual scaling analysis. Selecting profiles with high or low values on Dimensions 2 or 3 might have prompted the raters to provide more insights into the meaning of those dimensions. Second, we were able to interview only eight of the nine raters in the study. The omitted rater rated quite stringently. Interviewing him might have provided additional insights into the dimensions. Finally, we found the interpretation difficult when a dimension accounted for less than 15% of the variance.

In spite of these limitations, this study demonstrates the use of an approach that we believe has great promise for investigations of raters' agreement. The sources of disagreement we found in this study are beyond those researchers usually look for, and they cannot be easily addressed in the typical ways: by clarifying the questions or the rating rubric. Instead, these results point to fundamental disagreements among raters about their roles and about whether and how they should apply the rubrics. These disagreements among raters will be difficult to address; however, based on the information from this analysis, we can begin to address them.

REFERENCES

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995).A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565-579.

Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record, 107*(1), 186-215.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 4*, 64-86.

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*, 403-424.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1994). Interjudge reliability and decision reproducibility. *Educational and Psychological Measurement, 54*, 913-925.

Nishisato, S. (1994). *Elements of dual scaling: An introduction to practical data analysis.* Hillsdale, NJ: Erlbaum.

Nishisato, S., & Nishisato, I. (1998). *The DUAL3 Statistical Software Series for Windows, version 4.1* [computer software]. Toronto: Microstats.

Rudner, L. M. (1992). Reducing errors due to the use of judges. *Practical Assessment, Research & Evaluation, 3*(3). Retrieved April 27, 2009 from http://PAREonline.net/getvn.asp?v=3&n=3 .

Russell, M., & Tao, W. (2004a). Effects of handwriting and computer-print on composition scores: A follow-up to Powers, Fowles, Farnum, & Ramsey. *Practical Assessment, Research & Evaluation, 9*(1). Retrieved April 27, 2009 from http://PAREonline.net/getvn.asp?v=9&n=1 .

Russell, M., & Tao, W. (2004b). The influence of computer-print on rater scores. *Practical Assessment,*

Research & Evaluation, 9(10). Retrieved April 27, 2009 from http://PAREonline.net/getvn.asp?v=9&n=1 .

Saal, R. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413-428.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). Retrieved April 27, 2009 from http://PAREonline.net/getvn.asp?v=9&n=4 .

Appendix A.

Profile Questions

Part 1. Experience

(A) Please list and briefly describe 3 personal experiences that you believe have prepared you for a career in teaching. Consider a wide range of experiences.

You may use point form. There is a 150 word limit (50 words per experience).

(B) Drawing upon one of your selected experiences in Part One (A), explain significant insights that you have gained about teaching and learning. Provide specific examples from the experience to support your insights. You can discuss specific events, teaching strategies, and/or interactions with learners in your response. Identify the experience from those listed above that you are using as a basis of your response.

Use full sentences. There is a 300 word limit.

Part 2. Diversity

Teachers and the students and families with whom they work in schools differ in many ways including, but not limited to gender, race, socio-economic status, sexuality, religion, geographic region, ethnicity, and dis/ability. Please discuss how your own social background and other life experiences either inside or outside of school have prepared you to work with diverse students and families in schools.

Use full sentences. There is a 300 word limit.

Part 3. Equity and Social Justice

The differences that characterize teachers, students and their families (differences that include, but are not limited to gender, race, socio-economic status, sexuality, religion, geographic region, ethnicity, and dis/ability) can be linked to experiences of advantage and disadvantage. Describe a time when you or someone you know was advantaged or disadvantaged because of those differences. What was the impact of the experience? What did you learn from this experience that has prepared you to work with students and families who have experienced advantage or disadvantage?

Use full sentences. There is a 300 word limit.

Appendix B.

Profile Rubric

Part 1. Experience

INSUFFICIENT EVIDENCE: The response does not meet the criteria for PASS.

PASS: The response (1) describes three experiences, (2) provides at least one specific example of interactions with learners from one of the experiences, and (3) describes basic insights about teaching and/or learning they gained from reflecting on the interaction.

HIGH PASS: The response meets the criteria for PASS, plus describes deeper insights about teaching and/or learning they gained from reflecting on the interaction.

Part 2. Diversity

INSUFFICIENT EVIDENCE: The response does not meet the criteria for PASS.

PASS: The response (1) describes the applicant's own background and experiences in terms of gender, race, socio-economic status, sexuality, religion, geographic region, ethnicity, dis/ability and/or other social categories and (2) describes at least one thing they have learned, based on their own experiences, that has prepared them to work with diverse students and families in schools.

HIGH PASS: The response meets the criteria for PASS, plus demonstrates the applicant's commitment to at least one of the following: (1) not making assumptions about others based on cultural stereotypes, (2) learning about diversity from and with others, or (3) applying multiple lenses to understanding diversity.

Part 3. Equity And Social Justice

INSUFFICIENT EVIDENCE: The response does not meet the criteria for PASS.

PASS: The response (1) describes an experience in which the applicant or someone they know felt advantaged or disadvantaged because of their difference(s) and (2) describes at least one thing they have learned, based on this experience, that has prepared them to address equity and social justice through their work with students and families.

HIGH PASS: The response meets the criteria for PASS, plus demonstrates a deeper understanding of the societal or systemic contexts or sources of disadvantage.

Citation

Childs, Ruth A., Ram, Anita & Xu, Yunmei (2009). Combining Dual Scaling with Semi-Structured Interviews to Interpret Rating Differences. *Practical Assessment, Research & Evaluation*, 14(11). Available online: http://pareonline.net/getvn.asp?v=14&n=11.

Corresponding Author

Ruth A. Childs
Department of Human Development & Applied Psychology
Ontario Institute for Studies in Education, University of Toronto
Phone: 416-978-1079
http://fcis.oise.utoronto.ca/~rchilds/

Email: rchilds [at] oise.utoronto.ca