Title
Pairwise comparison versus Likert scale for biomedical image assessment.

Journal
AJR. American journal of roentgenology, 204(1)

ISSN
0361-803X

Authors
Phelps, Andrew S
Naeger, David M
Courtier, Jesse L
et al.

Publication Date
2015

DOI
10.2214/ajr.14.13022

Peer reviewed

Pairwise Comparison Versus Likert Scale for Biomedical Image Assessment

Andrew S. Phelps[1]
David M. Naeger
Jesse L. Courtier
Jack W. Lambert
Peter A. Marcovici
Javier E. Villanueva-Meyer
John D. MacKenzie

OBJECTIVE. Biomedical imaging research relies heavily on the subjective and semiquantitative reader analysis of images. Current methods are limited by interreader variability and fixed upper and lower limits. The purpose of this study was to compare the performance of two assessment methods, pairwise comparison and Likert scale, for improved analysis of biomedical images.

MATERIALS AND METHODS. A set of 10 images with varying degrees of image sharpness was created by digitally blurring a normal clinical chest radiograph. Readers assessed the degree of image sharpness using two different methods: pairwise comparison and a 10-point Likert scale. Reader agreement with actual chest radiograph sharpness was calculated for each method by use of the Lin concordance correlation coefficient (CCC).

RESULTS. Reader accuracy was highest for pairwise comparison (CCC, 1.0) and ranked Likert (CCC, 0.99) scores and lowest for nonranked Likert scores (CCC, 0.83). Accuracy improved slightly when readers repeated their assessments (CCC, 0.87) or had reference images available (CCC, 0.91).

CONCLUSION. Pairwise comparison and ranked Likert scores yield more accurate reader assessments than nonranked Likert scores.

Biomedical imaging research often involves comparing medical image quality by use of subjective human observation. The Likert scale is a commonly used assessment tool. It was developed by the psychologist Rensis Likert [1] in 1932 as a way of measuring attitudes. Use of the scale yields an ordinal dataset with arbitrary numbers that are significant only insofar as they can establish a rank order. Because the data are ordinal, most statisticians agree that it is inappropriate to perform parametric statistics (e.g., mean and SD) with Likert scale data [2–10]. Nonparametric analytic methods have thus been described for Likert scale image assessment [11, 12], yet recent examples of parametric analysis can still be found in the literature [13–15]. The objective of our research was to explain and evaluate a known alternative method of image comparison that does not require a Likert scale. This alternative method is called pairwise comparison.

Before pairwise comparison was introduced into radiology, professional chess players had long been assessed in a pairwise manner at tournaments. In a chess tournament, players are ranked according to the number of matches they have won and lost. The most famous ranking system was introduced in 1960 by Arpad Elo, a physics professor and chess master. The scale of the Elo ranking system has an arbitrary center, and there is no theoretic upper limit. Unlike a Likert scale, the chess rank is not bound by a maximum number of points.

Pairwise comparison is the closest analogue to the chess ranking system and has been well described as an accurate method of image assessment in psychophysics literature [16–20]. Use of the term "pairwise comparison" in our study should not be confused with the use of pairwise comparison for statistical comparison of different readers' results. In our study, pairwise comparison refers to the method whereby an individual reader compares two images side by side and chooses which image is better. Much as in a tournament, each image must be matched once with every other image, and for n number of different images, the number of unique comparisons required is equal to $(n - 1)^2 / 2 + (n - 1) / 2$. This equation excludes identical matchups (e.g., A versus A)

Keywords: image assessment, Likert scale, pairwise comparison

DOI:10.2214/AJR.14.13022

Received April 19, 2014; accepted after revision July 2, 2014.

[1]All authors: Department of Radiology and Biomedical Imaging, University of California, San Francisco, Benioff Children's Hospital, 505 Parnassus Ave, Box 0628, M-396, San Francisco, CA 94143. Address correspondence to A. S. Phelps (Andrew.Phelps@ucsf.edu).

AJR 2015; 204:8–14

0361–803X/15/2041–8

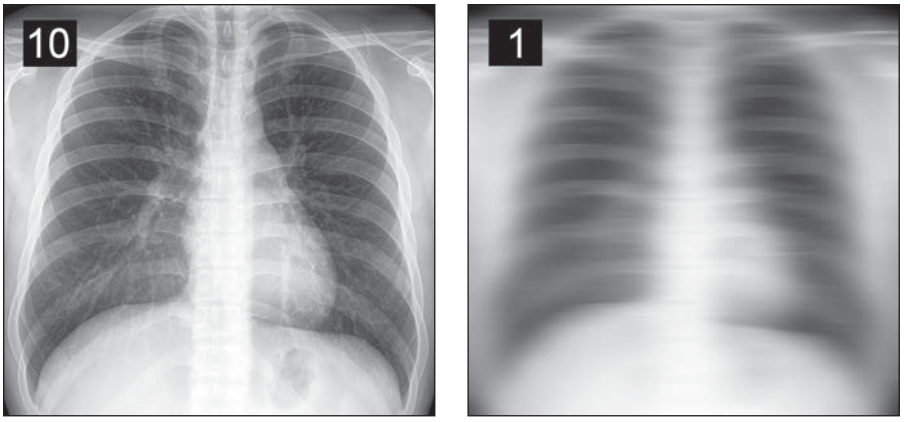

Fig. 1—21-year-old man in normal health.
A, Sharpest (original) radiograph (image score 10).
B, Least sharp radiograph (image score 1) digitally derived from original radiograph by use of motion blur feature of image-editing software.

and switched-order matchups (e.g., A versus B, B versus A). There are radiology literature examples in which pairwise comparison has been used in image assessment, but we could find no study that compared the accuracy of pairwise comparison with the accuracy of Likert scales [21, 22].

Our aim was to compare the accuracy of pairwise comparison with that of a Likert scale in biomedical image assessment. We designed an experiment in which readers would assess chest radiograph sharpness using both pairwise comparison and a Likert scale. Gur et al. [19] used a similar study design, but our study is unique in that we studied pairwise comparison separately from a Likert scale. Our primary hypothesis was that pairwise comparison would be more accurate than a Likert scale. Our secondary hypothesis was that a Likert scale would be more accurate when the Likert scores are converted into ranks, the reader is required to make repeat assessments for each image, and the reader has image references for comparison.

Materials and Methods

This study was granted exempt status from our institution's committee on human research. Informed consent was obtained from all readers.

Images

The first step was to create a standardized set of images that all readers would assess. We decided, for the sake of simplicity, to have the readers assess the sharpness of a single chest radiograph that we digitally blurred. A normal chest radiograph of a young adult was chosen. Using Photoshop CS5 (Adobe Systems), we applied the motion blur effect to the chest radiograph nine times in 10-pixel increments. With the original radiograph, we now had 10 different radiographs, which varied linearly with 10 different degrees of sharpness. We designated each of these radiographs as having actual sharpness scores from 1 to 10, where 1 indicated the least sharp radiograph, and 10 indicated the sharpest (original) radiograph (Fig. 1).

Testing Instrument

The second step was to assemble the different assessment tests. Three tests were created, and all were assembled in a Microsoft PowerPoint presentation. Sample slides from each test are shown in Figure 2. The first test was called the pairwise comparison test (Fig. 2A). This test consisted of 45 side-by-side comparisons of the 10 different radiographs with two radiographs per slide. The unique comparisons were presented randomly, and every unique comparison was included. Each reader was required to determine one of three possibilities: left radiograph is sharper, right radiograph is sharper, or radiograph sharpness cannot be differentiated.

The second test was called the Likert test without references. This test consisted of 10 slides with a single randomly presented radiograph on each slide (Fig. 2B). This test was performed 4 times, with a complete set of 10 different slides shown randomly, for a total of 40 slides. This means that each reader saw each of the 10 radiographs 4 times. Each reader assigned a sharpness score between 1 (least sharp) and 10 (sharpest) for all 40 slides. We chose a 10-point scale (instead of the traditional 5-point Likert scale) so that it would be possible for a reader to achieve 100% agreement with the actual sharpness scores (which also ranged from 1 to 10). Before starting this second test, each reader was shown the actual sharpest and least sharp radiographs (Fig. 1); however, the readers did not have these references after beginning the test.

The third and final test was called the Likert test with references. The only difference from the second test was that this third test also included the reference radiographs scored 1 and 10 flanking the unknown radiograph (Fig. 2C). To maximize and standardize the size of the unknown images across the three tests, the size of the reference images was reduced so that they would fit on one slide.

At the end of the three tests, there were two survey questions. The first question asked which test was easiest. The second question asked which test required the most guessing. The combined presentation was 125-slides long, including all three tests and instruction slides. The readers took the three tests in the order pairwise, Likert without reference images, and Likert with reference images under identical low-lighting conditions and using the same 15-inch (38.1 cm) 2011 Macbook Pro laptop (Apple) with screen resolution of 1440 × 900 pixels. Tests were performed in a dark room that simulated radiology reading room conditions. The readers' assessments and survey responses were read out loud and recorded by the same author, who sat behind them. The time required to complete the three tests was also recorded.

Readers

There were six readers, with different degrees of radiology interpretation expertise: three board-certified pediatric radiologists with a certificate of added qualification (9, 4, and 2 years of experience after residency), one chest radiologist (3 years' experience after residency), one third-year radiology resident, and one radiology postdoctorate researcher. The readers were blinded to the experimental hypothesis and had not seen the radiographs before the test.

Analysis

The investigators were not blinded for analysis of the data. The results were analyzed with a Google Docs spreadsheet (Google) for descriptive statistics. The readers' responses were used to create a sharpness score for each of the 10 radiographs. For the pairwise comparison test, we counted a win each time a radiograph was considered sharper than its paired comparison. If a

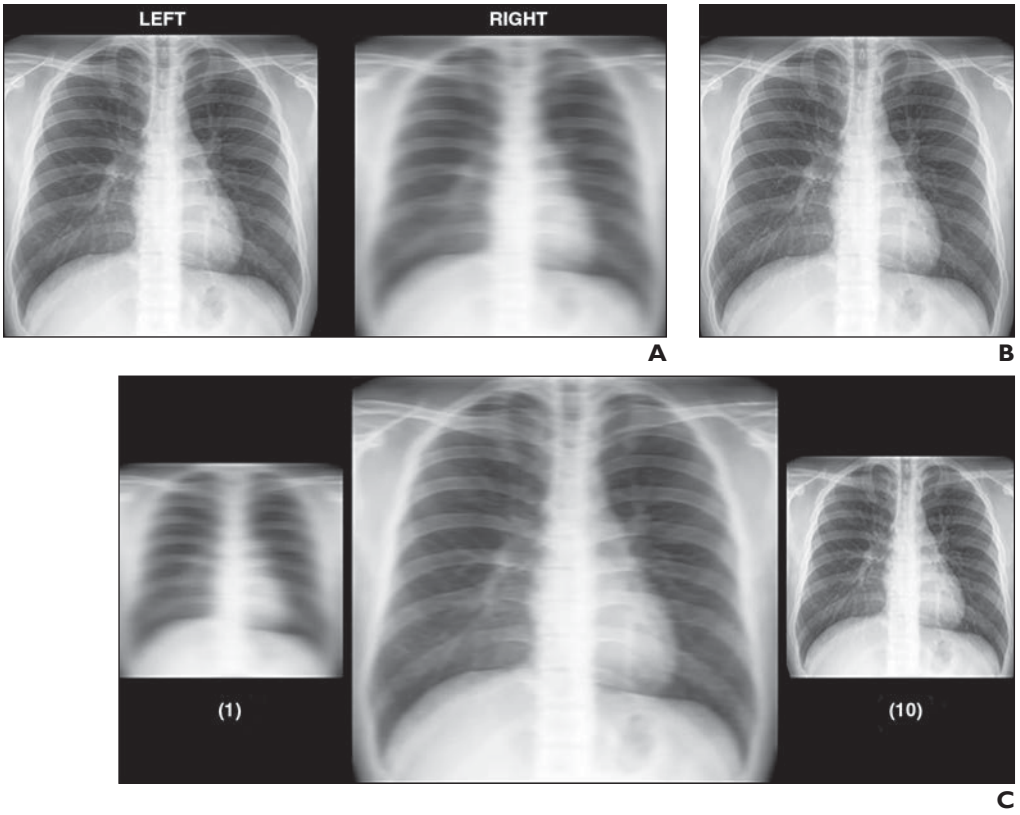

Fig. 2—Sample test slides. Unknown radiographs are all same size in each test, with resultant expense of having smaller reference radiographs.
A, Pairwise comparison test.
B, Likert test without references.
C, Likert test with references. 1 = least sharp, 10 = sharpest.

reader could not distinguish an image pair (i.e., a draw), no points were added or removed for either radiograph. The numbers of wins were tallied and subsequently ranked in descending order from 10 to 1, with 10 representing the most wins (sharpest) and 1 representing the fewest wins (least sharp). For the Likert tests, scores were determined directly by the reader (from 1 to 10), and each radiograph was given an average score with a single decimal place. Subsequently, the individual and cumulative Likert scores were ranked in ascending order from 1 to 10, also on the scale of 1 being least sharp and 10 being sharpest.

TABLE 1: Suggested Significance Cutoffs for Lin Concordance Correlation Coefficient

Concordance Correlation Coefficient	Description
1	Perfect agreement
> 0.99	Almost perfect agreement
> 0.95–0.99	Substantial agreement
> 0.9–0.95	Moderate agreement
> 0–0.9	Poor agreement
0	No agreement
–1	Perfect disagreement

We defined accuracy as the ability of a reader to reproduce the actual sharpness score. To determine accuracy, we calculated agreement between a reader's sharpness scores (and ranks) and the actual sharpness scores. To calculate this agreement, we used the Lin concordance correlation coefficient (CCC) [23], which we calculated with an online calculator [24]. The CCC shows the strength of agreement between two sets of data. Although similar to the Pearson correlation coefficient, which is used to assess linear agreement, the CCC additionally shows departure from the 45° straight line that is expected with perfect agreement (i.e., when $x = y$) [23]. The CCC ranges from 1 (perfect agreement) to –1 (perfect disagreement). We used the recommended cutoffs for statistical significance that are listed in Table 1. A CCC was calculated for each test result for each reader. The second test, Likert test without references, had CCCs calculated for the first 10 slides in addition to the entire 40 slides.

Results

Test Times

The three tests combined took an average of 5:25 minutes (range, 4:19–7:28 minutes) to complete. The individual questions took an average of 3.2 seconds for the pairwise comparison test, 3.6 seconds for the Likert test without references, and 4.0 seconds for the Likert test with references.

Primary Hypothesis: Pairwise Comparison Test Had the Best Accuracy

When pairwise comparison scores were tallied, the resultant ranks had perfect correlation with actual sharpness scores (CCC, 1.0) (Table 2). Moreover, the substantial individual accuracy of each reader's pairwise comparisons was higher than the accuracy of every other Likert score, individual or cumulative (CCC, 0.98–0.99 vs 0.59–0.96) (Table 3).

Secondary Hypothesis: Accuracy Was Substantially Improved by Converting Likert Scores Into Ranks

For the Likert test without references, the readers overassigned low sharpness scores and underassigned high sharpness scores (Fig. 3), resulting in poor cumulative accuracy (CCC, 0.83 for first 10 slides) (Table 2). Having readers repeat their assessments 4 times instead of 1 time slightly improved the cumulative accuracy of their Likert scores (CCC, 0.87 for all 40 slides). Allowing readers access to the reference images also slightly improved the cumulative accuracy of their Likert scores (CCC, 0.91). However, converting the Likert scores into

TABLE 2: Scores and Ranks for Each Radiograph

	Pairwise Comparison (45 Slides)		Likert Without Reference (Slides 1–10)		Likert Without Reference (Slides 1–40)		Likert With Reference (10 Slides)	
Actual Score	No. of Wins	Rank	Score	Rank	Score	Rank	Score	Rank
1 (least sharp)	0	1	1.0	1	1.3	1	1.0	1
2	1	2	1.7	2	1.5	2	1.7	3
3	8	3	2.2	3	2.2	3	1.2	2
4	11	4	3.2	4	2.8	4	2.2	4
5	17	5	4.2	6	3.8	6	3.2	5
6	24	6	3.3	5	3.5	5	5.2	7
7	28	7	4.3	7	5.0	7	5.0	6
8	35	8	5.2	8	5.6	8	6.7	8
9	40	9	8.0	9	8.0	9	8.8	9
10 (sharpest)	43	10	8.8	10	9.3	10	10.0	10
Concordance correlation coefficient	NA	1.0	0.83	0.99	0.87	0.99	0.91	0.98

Note—The combined results for all six readers are presented for each actual chest radiograph score. The pairwise comparison test had perfect agreement with actual sharpness scores with concordance correlation coefficient (CCC) equal to 1.0. Converting the Likert scores into ranks increased the CCC of those tests from 0.83–0.91 to 0.98–0.99. NA = not applicable.

TABLE 3: Accuracy for Each Reader

	Pairwise Comparison (45 Slides)		Likert Without Reference (Slides 1–10)		Likert Without Reference (Slides 1–40)		Likert With Reference (10 Slides)	
Reader	Score	Rank	Score	Rank	Score	Rank	Score	Rank
1	NA	0.99	0.59	0.91	0.63	0.98	0.82	0.97
2	NA	0.99	0.84	0.90	0.96	0.98	0.91	0.95
3	NA	0.98	0.82	0.82	0.90	0.98	0.91	0.97
4	NA	0.99	0.80	0.95	0.79	0.99	0.86	0.97
5	NA	0.98	0.76	0.96	0.83	0.98	0.88	0.96
6	NA	0.99	0.94	0.98	0.94	1.0	0.91	0.96
Cumulative	NA	1.0	0.83	0.99	0.87	0.99	0.91	0.98

Note—Values are concordance correlation coefficients (CCC) for each reader's tests. The accuracy is best with cumulative pairwise comparison assessments (CCC, 1.0), and individual pairwise comparison assessments are better than any nonranked Likert score assessment, individual or cumulative (CCC, 0.98–0.99 vs 0.59–0.96). However, converting Likert scores into ranks significantly improves accuracy (cumulative CCC, 0.98–0.99). NA = not applicable.

ranks led to the most substantial improvement in accuracy for each of the Likert tests (individual CCCs, 0.82–1.0; cumulative CCCs, 0.98–0.99). Improvements in individual accuracies are listed in Table 3 and shown in Figure 4.

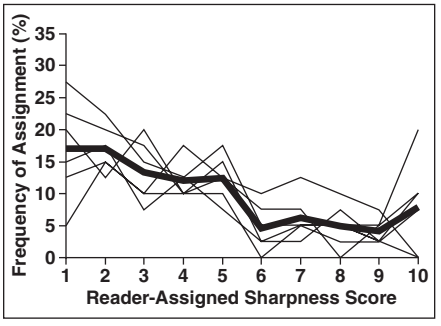

Fig. 3—Graph shows frequency with which 10 possible sharpness scores were assigned during Likert test without reference images, which included 40 separate assessments for each reader. Thin lines represent different readers. Thick line represents average of readers. If scores had been used equally, frequency would be 10% for each score. However, readers overassigned lower scores and underassigned higher scores.

Ranked Likert scores yielded cumulative accuracies that were similar to the cumulative accuracy from the pairwise comparison test, though the pairwise comparison accuracy was still slightly higher (CCC, 1.0 vs 0.98–0.99) (Fig. 5).

Readers Mostly Preferred the Pairwise Comparison Test

All six readers agreed that the Likert test without references required the most guessing. Five of the six readers agreed that the pairwise comparison test was easiest. One of the six thought that the Likert test with references was easiest.

Discussion

The goal of our study was to evaluate an alternative to the Likert scale for evaluating biomedical images. Our primary hypothesis was supported. The pairwise comparison method yielded perfect accuracy for the group and substantial accuracy for individual readers. The individual accuracy with

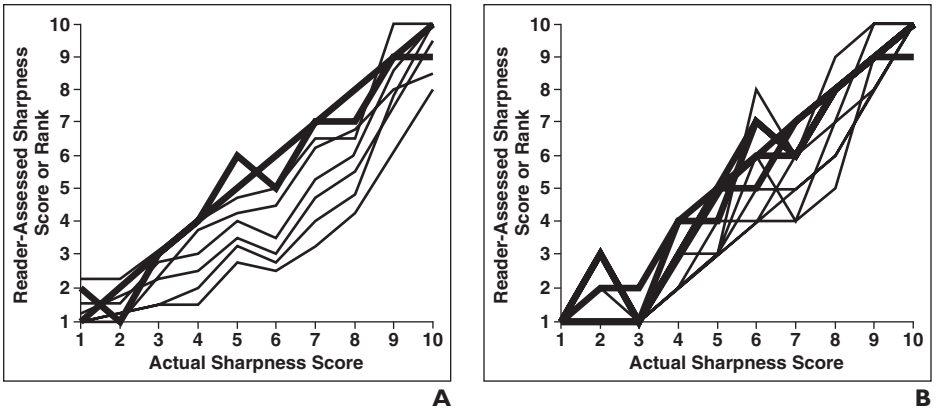

Fig. 4—Graphs show individual reader sharpness scores (*thin lines*) and ranks (*thick lines*) versus actual sharpness scores (all six readers' lines not evident owing to overlap). In both tests, readers' sharpness scores (*thin lines*) tended to underestimate actual sharpness. Converting scores (*thin lines*) into individual ranks (*thick lines*) increases individual agreement with actual sharpness score (CCC, 0.95–1.0 vs 0.63–0.96).
A, Data from Likert test without references, which included 40 separate assessments for each reader.
B, Data from Likert test with references, which included 10 separate assessments for each reader.

pairwise comparison was better than the group accuracy with Likert scores. Moreover, most of the readers thought that the pairwise comparison method was easiest and that the Likert scale assessment without references seemed most like guessing. These findings support previous results in the psychophysics literature [16–20]. We speculate that pairwise comparison is better because it does not require the reader to remember the upper and lower limits of a scale. Because most biomedical imaging research is not going to have a perfect and a worst example to guide readers, the scale independence of pairwise comparison is convenient. The main drawback of the pairwise comparison method is the number of comparisons required. For n images to be compared, the total number of comparisons equals $(n-1)^2/2 + (n-1)/2$. The number of comparisons therefore increases exponentially with the number of images, whereas Likert scale assessments increase linearly with the number of images. The large number of pairwise comparisons increases test design time and reader time.

Our secondary hypothesis was that there are three ways to improve the accuracy of Likert scores without having to use pairwise comparison. The first way is to convert the Likert scores into ranks, and the results supported this. The readers in this study were reluctant to assign high scores, as though they were waiting for better images. This occurred even when the readers were provided with best and worst reference examples on the same slide. Underreporting of the extremes of a Likert scale is known as central tendency bias and has been described in the nonbiomedical literature [25, 26]. However, we corrected for this bias by converting the scores into ranks, which ensured that every reader's extremes were calibrated to be identical. This correction resulted in a substantial increase in accuracy, which was similar

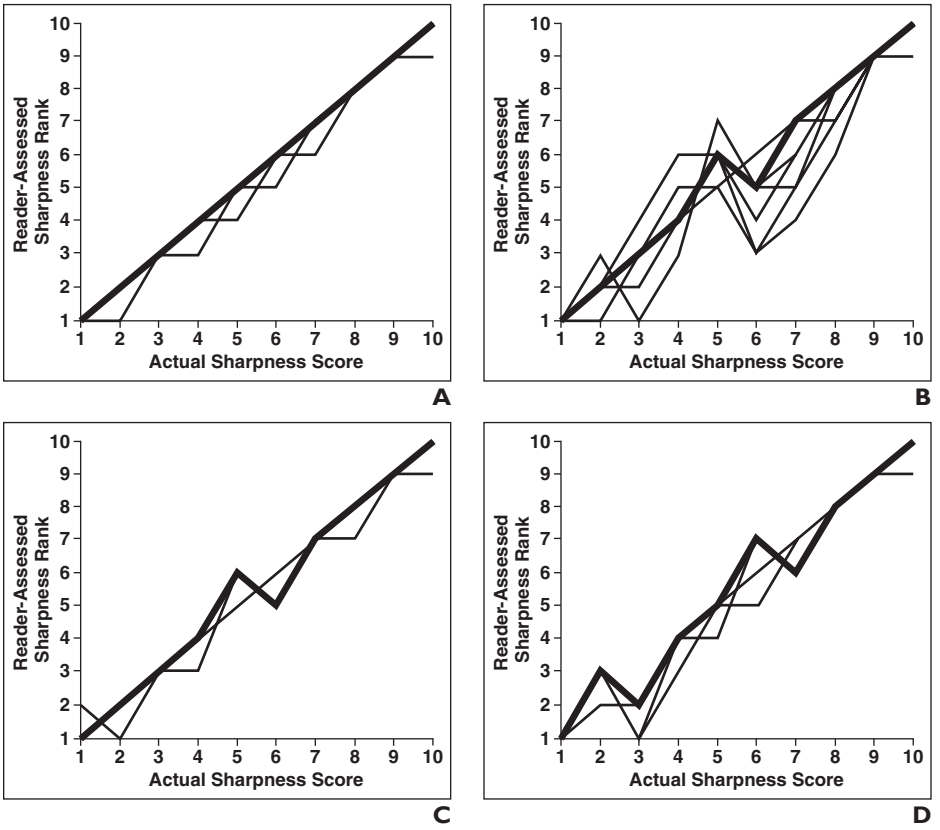

Fig. 5—Graphs show sharpness ranks of all three tests. Thin lines represent individual readers (all six readers' lines not evident owing to overlap). Single thick line in each graph represents cumulative sharpness ranks obtained after combining individual scores. Pairwise comparison method has highest accuracy, and accuracies of ranked Likert methods are similar (CCC, 0.98–1.0).
A, Pairwise comparison test.
B, Likert test without references (first 10 slides).
C, Likert test without references (all 40 slides).
D, Likert test with references.

to the perfect accuracy achieved with pairwise comparison. The two other hypothesized ways to improve Likert scores were not as helpful: Having readers repeat their Likert assessments and providing them with reference best and worst images did not noticeably improve accuracy. Moreover, the minimal improvement seen with the Likert test with references may have occurred simply because it was the last test, and readers had become familiar with the images. Therefore, of the three hypothesized ways to improve Likert score accuracy, the only one that substantially improved accuracy (and reduced interreader variability) was to convert the Likert scores into ranks.

The main limitation of our study relates to whether our results can be generalized to biomedical imaging research looking at more than just image sharpness of normal anatomy. Subjective assessment of image quality is useful, but diagnostic accuracy is far more important, and the results of our study may not be applicable to real clinical images. We also chose to evaluate a single parameter, image sharpness, although more rigorous parameters have been proposed in the literature [11, 12, 27]. We made this choice because image sharpness is an easy parameter to manipulate in a linear manner using fixed pixel increments. We could have chosen to alter tube voltage or current, but that would have required either a phantom or human subjects who provided informed consent, and the imaging results might not have been as predictably homogeneous as those in this study. Because the primary study question was to assess pairwise comparison, it was more important to have tight control over the differences between images. Although pairwise comparison proved beneficial in differentiating uniformly blurred images, perhaps there would be no benefit over a Likert scale for focal changes (e.g., detecting a lung nodule). Further testing is required to assess the performance of pairwise comparison with real clinical images and heterogeneous pathologic changes. We created a hypothetical imaging experiment (Appendix 1) to illustrate how a researcher can use pairwise comparison to assess any subjective imaging parameter and to show that the resultant ranks can be statistically evaluated by use of the Mann-Whitney *U* test [28].

There were other limitations to our study. The radiographs were not reviewed at full diagnostic resolution at a radiology workstation; however, we did not believe that the lower resolution of our radiographs would have any study effect other than possibly nulling the hypothesis. The Likert tests were performed later in the experiment and might therefore have had an unfair advantage, because the readers might have become accustomed to the different radiographs. We chose to perform the pairwise comparison test first, so that any favorable result for the pairwise comparison test could not be assigned to an increase in familiarity with the images. The reference images for the last Likert test were smaller than the unknown images, which may have reduced the reader's benefit of having the references. In addition, the most common Likert scale is a 5-point scale, whereas our scale had 10 points, which might have been more difficult for readers to use consistently. However, results of studies suggest that the number of points on a Likert scale does not affect accuracy [29–31]. We chose a 10-point scale so that with 10 different radiographs 100% accuracy was possible for all tests, not just the pairwise comparison. Last, we had only six readers, and they had varying degrees of clinical radiology experience. With more than six readers and closer to uniform radiology experience, there might not have been as much interreader variability. However, we assumed that most biomedical imaging research is not conducted with a large number of readers. Therefore, our results would still provide practical information for most researchers. Of note, our least experienced reader's accuracy was similar to (and sometimes better than) the accuracy of the more experienced readers, likely owing to the simplicity of assessing image sharpness, which does not require clinical experience.

Conclusion

In keeping with results of previous studies, pairwise comparison is a better method of image assessment than are nonranked Likert scores, having higher accuracy and lower interreader variability. Pairwise comparison is easy for readers, though the number of comparisons required increases exponentially with the number of images studied. Therefore, pairwise comparison should be reserved for smaller biomedical imaging studies that have fewer readers and images. For more extensive studies, converting Likert scores into ranks will yield high accuracy without the need to perform pairwise comparison.

References

1. Likert R. A technique for the measurement of attitudes. *Arch Psychol* 1932; 140:55
2. Knapp T. Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nurs Res* 1990; 39:121–123
3. Brown J. Likert items and scales of measurement? *Shiken: JALT Testing & Evaluation SIG Newsletter* 2011; 15:10–14
4. Kostoulas A. On Likert scales, ordinal data and mean values. achilleaskostoulas.com/2013/02/13/on-likert-scales-ordinal-data-and-mean-values. Updated February 13, 2013. Accessed February 10, 2014
5. Jamieson S. Likert scales: how to (ab)use them. *Med Educ* 2004; 38:1217–1218
6. Blaikie N. *Analysing quantitative data*. London, UK: Sage Publications, 2003
7. Clegg F. *Simple statistics*. Cambridge, UK: Cambridge University Press, 1998
8. Kuzon WM Jr, Urbanchek MG, McCabe S. The seven deadly sins of statistical analysis. *Ann Plast Surg* 1996; 38:265–272
9. Armstrong GD. Parametric statistics and ordinal data: a pervasive misconception. *Nurs Res* 1981; 30:60–62
10. de Winter JC, Dodou D. Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. *Pract Assess Res Eval* 2010; 15:1–18
11. Båth M, Månsson L. Visual grading characteristics (VGC) analysis: a nonparametric rank-invariant statistical method for image quality evaluation. *Br J Radiol* 2007; 80:169–176
12. Smedby O, Fredrikson M. Visual grading regression: analysing data from visual grading experiments with regression models. *Br J Radiol* 2010; 83:767–775
13. Spears JR, Schoepf UJ, Henzler T, et al. Comparison of the effect of iterative reconstruction versus filtered back projection on cardiac CT postprocessing. *Acad Radiol* 2014; 21:318–324
14. Kroft LJ, Veldkamp WJ, Mertens BJ, van Delft JP, Geleijns J. Dose reduction in digital chest radiography and perceived image quality. *Br J Radiol* 2007; 80:984–988
15. Kinner S, Blex S, Maderwald S, Forsting M, Gerken G, Lauenstein T. Addition of diffusion-weighted imaging can improve diagnostic confidence in bowel MRI. *Clin Radiol* 2014; 69:372–377
16. Bradley RA, Terry ME. Rank analysis of incomplete block designs: the method of paired comparisons. *Biometrika* 1952; 39:324–345
17. Good WF, Gur D, Feist J, et al. Subjective and objective assessment of image quality: a comparison. *J Digit Imaging* 1994; 7:77–78
18. Britton CA, Gabriele OF, Chang TS, et al. Subjective quality assessment of computed radiography hand images. *J Digit Imaging* 1996; 9:21–24
19. Gur D, Rubin D, Kart B, et al. Forced choice and ordinal discrete rating assessment of image quality: a comparison. *J Digit Imaging* 1997; 10:103–107
20. Adamic P, Babiy V, Janicki R, Kakiashvili T, Koczkodaj W, Tadeusiewicz R. Pairwise compari-

sons and visual perceptions of equal area polygons. *Percept Mot Skills* 2009; 108:37–42
21. Slone RM, Foos DH, Whiting BR, et al. Assessment of visually lossless irreversible image compression: comparison of three methods by using an image-comparison workstation. *Radiology* 2000; 215:543–553
22. Raslan O, Debnam JM, Ketonen L, Kumar AJ, Schellingerhout D, Wang J. Stereoscopic visualization of diffusion tensor imaging data: a comparative survey of visualization techniques. *Radiol Res Pract* 2013; 2013:780916
23. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; 45:255–268
24. National Institute of Water and Atmospheric Research. Statistical calculators: Lin's concordance. www.niwa.co.nz/node/104318/concordance. Updated 2013. Accessed April 9, 2014
25. Albaum G. The Likert scale revisited: an alternate version. *J Mark Res Soc* 1997; 39:331–348
26. Kostoulas A. Four things you probably didn't know about Likert scales. achilleaskostoulas.com/2013/09/09/four-things-you-probably-didnt-know-about-likert-scales/ Updated September 9, 2013. Accessed February 10, 2014
27. Kohn MM. European guidelines on quality criteria for diagnostic radiographic images in paediatrics. Luxembourg, Luxembourg: Office for Official Publications of the European Communities, 1996
28. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 1947; 18:50–60
29. Matell MS, Jacoby J. Is there an optimal number of alternatives for Likert scale items? *Educ Psychol Meas* 1971; 13:657–674
30. Jacoby J, Matell MS. Three-point Likert scales are good enough. *J Mark Res* 1971; 8:495–500
31. Rockette HE, Gur D, Metz CE. The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. *Invest Radiol* 1992; 27:169–172
32. Social Science Statistics website. Mann-Whitney U-test calculator. www.socscistatistics.com/tests/mannwhitney/Default2.aspx. Updated 2014. Accessed February 10, 2014
33. Lowry R. Mann-Whitney test. VassarStats website. vassarstats.net/utest.html. Updated 2014. Accessed February 10, 2014

APPENDIX 1: Hypothetical Clinical Experiment Showing How to Use Pairwise Comparison and the Mann-Whitney *U* Test

Hypothetical Methods

A hypothetical low-dose head CT protocol is developed. A retrospective comparison with five studies in each group is made between head CT studies performed before and after the new low-dose protocol is implemented. A representative axial image is obtained at the same anatomic level for each of the 10 studies, yielding 10 different images. A single reader performs a randomized pairwise comparison between all images. With 10 images, 45 unique comparisons are required: $(n - 1)^2 / 2 + (n - 1) / 2$. For each comparison, the reader decides which image has better gray-white differentiation, and the number of wins is tallied for each image (maximum number of wins is 9). The number of wins is then ranked, from best (first) to worst (10th).

Hypothetical Results

The hypothetical results of this pairwise comparison are listed in Table 4. The hypothetical reader thinks that image 8 (a high-dose CT image) has better gray-white differentiation than all nine other images (hence, it has 9 wins, ranking first). At the opposite end, the reader thinks that image 2 (a low-dose CT image) has worse gray-white differentiation than all nine other images (hence, it has 0 wins, ranking 10th). Overall, the high-dose CT images appear to have better gray-white differentiation than the low-dose CT images (ranks 1, 2, 3, 5, 6 versus 4, 7, 8, 9, 10), but are these different ranks statistically significant?

TABLE 4: Hypothetical Pairwise Comparison Results

Image No.	CT Dose	No. of Wins	Rank
1	Low	6	4th
2	Low	0	10th (worst gray-white differentiation)
3	Low	2	8th
4	Low	3	7th
5	Low	1	9th
6	High	8	2nd
7	High	5	5th
8	High	9	1st (best gray-white differentiation)
9	High	7	3rd
10	High	4	6th

Note—A hypothetical experiment is performed to show how pairwise comparison can be applied to clinical images. A single reader performs pairwise comparison of 10 different head CT images. A win is counted whenever the reader thinks an image has better gray-white differentiation than the comparison. The reader thinks that image 8 is better than all the other images, hence image 8 has a total of 9 wins, putting it into first place.

Statistical Analysis of Ordinal (Ranked) Data

To determine a statistically significant difference between two groups of ordinal (ranked) data, the Mann-Whitney *U* test should be used [28]. The Mann-Whitney *U* test is the nonparametric equivalent of the Student *t* test. Easy-to-use Mann-Whitney *U* test calculators online [32, 33] even provide a *p* value so that the investigator need not bother with the actual *U* value and significance lookup tables. When our hypothetical data are entered into such a calculator, we obtain $p = 0.04$, indicating that the better gray-white differentiation with the high-dose CT protocol is statistically significant.