Forced Choice and Ordinal Discrete Rating Assessment of Image Quality: A Comparison

David Gur, David A. Rubin, Barry H. Kart, Arleen M. Peterson, Carl R. Fuhrman, Howard E. Rockette, and Jill L. King

This study compared a five-category ordinal scale and a two-alternative forced-choice subjective rating of image quality preferences in a multiabnormality environment. 140 pairs of laser-printed posteroanterior (PA) chest images were evaluated twice by three radiologists who were asked to select during a side-by-side review which image in each pair was the "better" one for the determination of the presence or absence of specific abnormalities. Each pair included one image (the digitized film at 100 μm pixel resolution and laser printed onto film) and a highly compressed (~60:1) and decompressed version of the digitized film that was laser printed onto film. Ratings were performed once with a five-category ordinal scale and once with a two-alternative forced-choice scale. The selection process was significantly affected by the rating scale used. The "comparable" or "equivalent for diagnosis" category was used in 88.5% of the ratings with the ordinal scale. When using the two-alternative forced-choice approach, noncompressed images were selected 66.8% of the time as being the "better" images. This resulted in a significantly lower ability to detect small differences in perceived image quality between the noncompressed and compressed images when the ordinal rating scale is used. Observer behavior can be affected by the type of question asked and the rating scale used. Observers are highly sensitive to small differences in image presentation during a side-by-side review.

KEY WORDS: image quality, observer performance, study methodology, ratings

RECEIVER OPERATING characteristic (ROC) studies have been used in recent years to discriminate between observers' performance when using different display modalities. When the differences are small, these studies require a large number of carefully selected cases, as well as a large number of observers, to be meaningful.[1-4] If displayed images are very similar it can be argued, in principle, that when observers cannot identify which images belong to what group (display mode), observer performance should not be affected. This suggests that a more sensitive subjective non–ROC study designed to detect small differences in perceived quality might be used to assess the need for a comprehensive observer performance study.[5] Although such studies have been used successfully in a variety of applications in recent years, the methodology used and the type of questionnaire implemented in different applications vary significantly from study to study.[5-8] It should be noted that despite their utility, such studies cannot be considered a substitute for an objective observer performance study.

Sensitivity to small differences in either perceived quality or the observers' ability to perform a specific diagnostic task (eg, detection of a nodule) has been enhanced through a study design that includes side-by-side comparisons.[5-8] Both the forced choice between two modalities[6,7] and the ordinal rankings of several modalities[5,8] have been used for this purpose. In some cases the question posed was related to changes of the images' "look and feel" after processing (eg, edge enhancement, data compression) as compared with the original image. The latter was known to the observer (clearly labeled) and was used as a reference image (a gold standard).[5] Although such studies have been shown to be highly sensitive (hence requiring a limited observer effort), the proper type of questions posed to observers and their potential impact on the study results have not been adequately addressed. In addition, these studies have not been used in a multitask environment, based on the assumption that observers are likely to subjectively rate one mode as "better" or "worse" across the board for all diagnostic tasks being investigated. This study attempts to explore the following: first, given a side-by-side review of pairs of images and five different diagnostic tasks, we wanted to compare a five-category ordinal rating of relative quality to a two-alternative forced-choice approach; and second, we attempted to evaluate the validity of implementing a multitask environment in such studies.

From the Department of Radiology, University of Pittsburgh, Pittsburgh, PA.

This work is supported in part by grants from the National Cancer Institute (CA60259, CA66594, and CA67947).

Address reprint requests to David Gur, MD, Allegheny Health Education and Research Foundation, 5th Ave Place, Suite 2900, 120 Fifth Ave, Pittsburgh, PA 15222.

0897-1889/97/1003-0002$3.00/0

METHOD

A total of 140 high-quality posteroanterior (PA) chest radiographs were digitized at 3504 × 4205 × 12 bit matrices with a high-resolution (100 μm pixel size) film digitizer (Lumisys; Sunnyvale, CA) with a measured modulation transfer function (MTF) at the Nyquist frequency (5 line pairs/mm) of 0.41 and 0.39 in the horizontal and vertical directions, respectively. The MTF was computed from digitized image data of a lead test pattern with image segmentation and Fourier transform. A compressed image was then generated from each digitized image by a psychophysical quantization approach.[5] The psychophysical quantization scheme is based on the belief that, for any given viewing condition, the importance of particular frequencies varies according to the contrast sensitivity of the human visual system. The psychophysical quantization factors were calculated based on the methods described in Nill[9] and Ngan et al[10] that used visual system data from Kelly.[11] For the purpose of the psychophysical calculation, we assumed that the images would be displayed on laser-printed film with a printing pixel spacing of 80 μm and viewed initially from a distance equal to the diagonal of the film (56 cm). This means that all spatial frequencies were represented with equal "visual fidelity" at this viewing distance.[9-11] The selection of this particular approach was based on a prior study in which this psychophysical scheme was perceived by experienced observers to produce images that were judged visibly to be the closest (most similar) to the original noncompressed images.[5] The data compression for the whole set of images was 63.6:1, with compression ratios for individual images ranging from 53:1 to 80:1. The rationale for selecting highly compressed images was that such images might be viewed at a picture archiving and communication system (PACS) workstation or by teleradiology.

After adjusting the lookup table to generate images that would match the original conventional radiograph to within 0.07 optical density (OD) in the range of 0.25 to 3.2, the digitized noncompressed and the compressed and decompressed image were laser printed at full size onto film (Kodak Ektascan; Rochester, NY). Each pair of images was labeled with a case number, and each image within a pair was randomly designated and labeled "A" or "B." Case verification protocols for the actual presence or absence of specific abnormalities have been described elsewhere.[12] The set included 57, 56, 26, 17, and 18 images visualizing interstitial disease, nodule, pneumothorax, alveolar infiltrates, and rib fracture. Twenty-five images were actually negative for all five abnormalities.

Three experienced, board-certified radiologists were presented twice with 140 pairs of laser-printed films displayed side by side on adjacent viewboxes. Each mode of the experiment was performed during three to four sessions in which 35 to 47 comparisons were made. During the review of each case (a pair of laser-printed images), the diagnostic truth was displayed on a computer monitor so that observers were aware of the abnormalities actually present (or absent), including the nodule's location (when applicable) and whether visible septal lines were noted when, and if, a diagnosis of interstitial disease had been made. This was done by displaying the scoring form routinely used in our ROC studies, along with the appropriate fields being populated in the form for each case (Fig 1).[12] The radiologists were told that each pair contained two different images. In the first series of sessions they were asked to rank order the relative

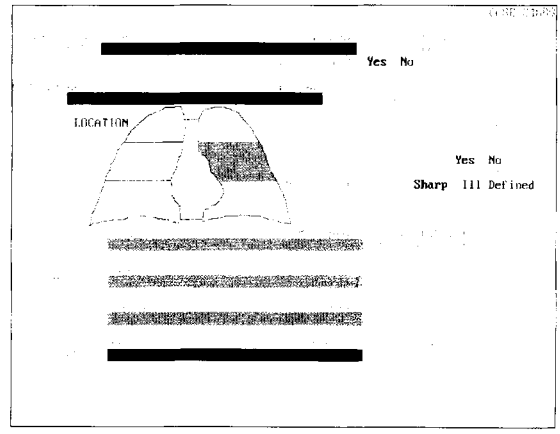

Fig 1. A demonstration of the displayed diagnostic "truth" for a specific case. This template is also used as the scoring form during observer performance (ROC) studies, without knowledge of the truth displayed.

differences between the two images. The five-category ordinal rating scale was defined as follows:

1. Image "A" is *much better* than image "B" for determining the correct diagnosis for the presence or absence of this abnormality (eg, nodule).
2. Image "A" is *somewhat better* than image "B" for determining the correct diagnosis for the presence or absence of this abnormality.
3. There is no difference between image "A" and image "B" for determining for the presence or absence of this abnormality.
4. Image "A" is *somewhat worse* than image "B" for determining for the presence or absence of this abnormality.
5. Image "A" is *much worse* than image "B" for determining for the presence or absence of this abnormality.

Observers were asked to rate each pair of images for each of five specific abnormalities: (1) interstitial disease, (2) nodule, (3) pneumothorax, (4) alveolar infiltrate, and (5) rib fracture. Note that while the diagnostic truth for each case was provided, the images were not identified as to which was the noncompressed and which was the compressed and decompressed. During the second part of the experiment, the observers were forced to choose one image in each pair ("A" or "B") as the better image for the specific diagnostic task of determining the presence or absence of a particular abnormality. In both parts of the experiment, they were asked to physically reorder the films several times before making a final decision to avoid being biased by specific arrangements of the images and viewbox display quality. During the studies, each observer was asked to provide subjective comments concerning task difficulty. After completion of the ratings, the results were decoded with the correct subgrouping of images (ie, noncompressed and compressed), and the data were tabulated by reader, abnormality, and type of questionnaire used (five category ordinal versus a two-alternative forced choice).

A *t*-test was used for the ordinal rating to test whether the average score was significantly different from No. 3 ("no difference"). Although the data are ordinal, because of the

Central Limit Theorem, the sample mean obtained from 140 observations should be approximately normal. Using a binomial distribution, we also compared the frequency of 1 and 2 ratings grouped together with the frequency of the 4 and 5 ratings grouped together. Conclusions were essentially the same as those obtained with the *t*-test. When combining over readers or abnormality, we considered the correlation of scores between readers rating the same abnormality or a single reader rating different abnormalities and used the usual formula for summing normal variables with nonzero correlation. To determine statistical significance for the two-alternative forced choice experiment, we used the normal approximation to the binomial distribution to test whether the probability of selecting the noncompressed image differed significantly from .5. When combining over readers or abnormalities, we fit a bivariate binomial distribution[13,14] to estimate the correlation between scores and then used the sum of the normal random variables approximating the individual binomials as the test statistic.

RESULTS

The results of the five-category ordinal rating experiment by observer and abnormality are provided in Table 1. From this table it is clear that in the majority of instances most observers rated the pairs of images as "comparable" or "equivalent" (1859/2100, 88.5%) for determining the presence or absence of the five abnormalities in question. Noted was the fact that with the exception of three cases (rated by reader 1), all other ratings fell into three categories (no extreme ratings). Although there was slight asymmetry for some readers for a specific abnormality, only interstitial disease as evaluated by reader 3 approached statistical significance ($P < .01$) in regard to a preference for the noncompressed more than for the compressed images. When individual scores were combined over all abnormalities, there was some indication of asymmetry, but the results were not statistically significant for any individual reader. When results were combined over three readers for a specific abnormality, only interstitial disease showed statistically significant asymmetry with the results favoring the selection of the noncompressed image ($P = .03$). Of 420 possible paired comparisons, 273 (or 65%) were rated the same for all five diagnostic questions. By reader, it was 87 (62%), 63 (45%), and 123 (88%). The results were similar for positive and negative cases, namely, with the abnormality actually present or absent.

The results of the two-alternative forced choice experiment are provided in Table 2. From this table it is clear that rated differences between the compressed and noncompressed images increased. Each of the three readers consistently selected more noncompressed than compressed and decompressed images as the "better" to determine the presence or absence of each of the five abnormalities. This was done in 1,403 of 2,100 ratings (or 66.8%). For ten of the 15 possible combinations of a specific abnormality as rated by an individual reader, the results showed a statistically significant preference for the noncompressed images. Combined results over all readers for each of the specific abnormalities are all statistically significant ($P < .001$). When results for each reader are combined over the five abnormalities, each of the three readers showed a preference for noncompressed images ($P < .001$, $P < .01$, and $P < .05$ for readers 1, 2, and 3, respectively). Different cases were rated non-uniformly across all diagnostic questions. However, the number of paired comparisons where a different choice was given as related to a different abnormality was comparable to that in the five-category ordinal rating experiment. From 420 possible comparisons, 281 (or 67%) were rated the same for all abnormalities. By reader, it was 84 (60%), 63 (45%), and 134 (96%). In both experiments, interreader variability was noted; reader 1 selected a larger fraction of the noncompressed images as the "better" ones in the two-alternative forced-choice portion of the experi-

Table 1. Results from the Five-Category Ordinal-Rating Experiment by Reader and Abnormality

	Reader 1					Reader 2					Reader 3				
Abnormality	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Interstitial disease	0	11	119	10	0	0	37	78	25	0	0	5	135	0	0
Nodule	1	14	115	10	0	0	20	106	14	0	0	5	129	6	0
Pneumothorax	2	7	126	5	0	0	25	97	18	0	0	0	139	1	0
Alveolar infiltrates	0	3	133	4	0	0	1	139	0	0	0	0	139	1	0
Rib fracture	0	3	134	3	0	0	4	132	4	0	0	2	138	0	0

Note: (1) Noncompressed image was rated as much better than compressed image, (2) Noncompressed image was rated as somewhat better than the compressed image, (3) Noncompressed and compressed images were rated as equivalent for diagnosis, (4) Compressed image was rated as somewhat better than the noncompressed image, (5) Compressed image was rated as much better than noncompressed image.

Table 2. The Number of Images Selected During the Two-Alternative Forced-Choice Experiment as Being "Better" by Reader and Abnormality

	Reader 1		Reader 2		Reader 3	
Abnormality	NC	CMP	NC	CMP	NC	CMP
Interstitial disease	113	27	89	51	82	58
Nodule	121	19	78	62	81	59
Pneumothorax	119	21	89	51	81	59
Alveolar infiltrates	119	21	78	62	82	58
Rib fracture	112	28	77	63	82	58

Abbreviations: NC, noncompressed; CMP, compressed.

ment, whereas reader 2 was more selective (attempted to rate small perceived differences) using the five-category ordinal rating. The readers also showed differences in their tendencies to prefer different modalities (compressed versus noncompressed) for different abnormalities. The average correlation of the selection of preferred modalities (ie, compressed versus noncompressed images) for different pairs of abnormalities was 0.27 for reader 1, 0.41 for reader 2, and 0.96 for reader 3.

DISCUSSION

Typical intra- and interreader variability in observer performance studies results in low sensitivity to the detection of small differences between display modes to a point where sample-size requirements and hence, the efforts required to demonstrate differences, are often impractical. As a result, many studies performed to date have failed to reject the null hypothesis (ie, no statistically significant differences could be measured). At the other extreme, large differences in performance between modes can be easily demonstrated, but the difference is typically so notable that an observer performance study is not needed. Therefore, it is desirable to develop non–ROC-type studies that are highly sensitive to small changes and at the same time correlate with actual observer performance. Unfortunately, there is only limited experience with study designs of this type, and many related issues are poorly understood.

The results of our study are consistent with our previous non-ROC-type studies in that small intermode differences that are perceived to be "very difficult" to determine by participating observers are actually detected with a high degree of accuracy. Hence, such studies typically require a relatively small sample size. The principle underlying these results may have important implications in the general design of experiments in which an attempt is made to quantify a subjective clinical judgment.

Ordinal rating scales are generally accepted as having greater statistical power when the full range of possible ratings is used. When the rating scale includes an "equivalent for diagnosis" category, however, the ordinal scale may actually result in less discriminating power if this category is frequently used. A forced-choice design provides one method of eliminating the possibility of overutilization of this category.

Although some minimal trend in the increasing selection of noncompressed images as "better for diagnosis" in each pair was identified as the study progressed, the fraction of the noncompressed images selected in the first, second, and third groups of cases were quite similar. Hence, the study conclusions were not significantly affected by the case-reading order. In our study, there may have been mode-order effect because the two-alternative forced-choice mode followed the five-category ordinal scale mode. We believe it cannot account for the large difference in the results.

The fact that for two of the readers a significant fraction of images were not rated the same for all five diagnostic questions is encouraging in that it indicates that using a multi-abnormality setting may provide additional information to that obtained from a single abnormality study. Some readers (as demonstrated by reader 3), however, will tend to select as "preferable" the same modality (image) for all abnormalities. The comments made by all observers during the study, that the task was so difficult for many cases that they felt "like flipping a coin in a large number of the comparisons" are consistent with the lack of extreme ratings, as well as with our previous experiences in similar studies.[6] The difference was that multiple specific diagnostic tasks, rather than image "sharpness," were rated in this study and the diagnostic truth was provided. The latter approach was taken to avoid ratings following the misclassification of cases (eg, a case with a very subtle nodule could be rated as "equivalent" because the rater misclassified it as negative, but actual differences could be detected once the abnormality is noted).

This preliminary study demonstrated that a two-alternative forced-choice methodology better identified (highlighted) small differences in perceived image quality to perform specific diagnostic tasks as compared with the methodology using a five-

category ordinal rating scale. This observation may have other implications in ROC-type studies in that multicategory or continuous rating scales were assumed (and hence used frequently) to be the best rating approach to be used in these types of studies.[15] If in reality this type of rating results in a less decisive response pattern by observers, the information ascertained per unit observer's effort may actually decrease as compared to a two-alternative forced-choice experiment. A related observation was previously noted in an experiment to assess the effect of observer training to distribute answers over a wide range on the study results.[16] Clearly, more work is needed in this regard.

REFERENCES

1. Rosenthal MS, Good WF, Costa-Greco MA, et al: The effect of image processing on chest radiograph interpretations in a PACS environment. Invest Radiol 25:897-901, 1990

2. Metz CE: Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 24:234-245, 1989

3. Rockette HE, Obuchowski NA, Gur D, et al: Effect of experimental design on sample size. Proc SPIE 1446:276-286, 1991

4. Obuchowski NA, Zepp RC: Simple steps for improving multiple-reader studies in radiology. AJR 166:517-521, 1996

5. Holbert JM, Staiger M, Chang TS, et al: Selection of processing algorithms for digital image compression: A rank order study. Acad Radiol 2:273-276, 1995

6. Good WF, Gur D, Feist JH, et al: Subjective and objective assessment of image quality—a comparison. J Digit Imag 7:77-78, 1994

7. Good WF, Maitz GS, Gur D: Joint photographic expert group compatible data compression of mammograms. J Digit Imag 7:123-132, 1994

8. Britton CA, Gabriele OF, Chang TS, et al: Subjective quality assessment of computed radiography hand images. J Digit Imag 9:21-24, 1996

9. Nill NB: A visual model weighted cosine transform for image compression and quality assessment. IEEE Trans Commun COM-33:551-557, 1985

10. Ngan KN, Leong KS, Singh H: Cosine transform coding incorporating human visual system model. Proc SPIE 707:165-171, 1986

11. Kelley DH: Visual contrast sensitivity. Opt Acta 24:107-129, 1977

12. Thaete FL, Fuhrman CR, Oliver JH, et al: Digital radiography and conventional imaging of the chest: A comparison of observer performance. AJR 162:575-581, 1994

13. Hamdan MA, Jensen DR: A bivariate binomial distribution and some applications. Austral J Statist 18:163-169, 1976

14. Hamdan MA, Nasro MO: Maximum likelihood estimation of the parameters of the bivariate binomial distribution. Commun Statist Theor Method 15:747-754, 1986

15. Rockette HE, Gur D, Metz C: The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. Invest Radiol 27:169-172, 1992

16. Gur D, Rockette HE, Good WF, et al: Effect of observer instruction on ROC study of chest images. Invest Radiol 25:230-234, 1990