Hardin, R.A., and Beckermann, C., "A New Standard for Radiographic Acceptance Criteria for Steel Castings: Gage R&R Study," in Proceedings of the 63rd SFSA Technical and Operating Conference, Paper No. 2.6, Steel Founders' Society of America, Chicago, IL, 2009.

A New Standard for Radiographic Acceptance Criteria for Steel Castings: Gage R&R Study

Richard A. Hardin and Christoph Beckermann

Department of Mechanical and Industrial Engineering
The University of Iowa, Iowa City, IA 52242

Abstract

A two stage repeatability and reproducibility study was conducted for the SFSA on a draft radiographic acceptance standard. In the new standard, the reader measures and determines the maximum fractional length of indications on the radiograph along a specified direction. In the first stage of the study SFSA members applied the new standard to the ASTM E186 standard radiographs giving data on "between reader" reproducibility. While in the second "in-house" stage of the study, seven members of our laboratory measured three radiographs four times to generate data on repeatability and reproducibility. The first stage results gave a wide range of confidence intervals for the standard radiographs. The CC type shrinkage had consistently lower error margins and the largest error margins occurred in the CA and CB type shrinkage when the indications were aligned in the direction of interest. The direction of interest greatly effected the measurements of the CA and CB shrinkage, but not the CC type. There were 21 of the 30 radiograph ratings in the first stage of the R&R study that had fairly low reproducibility errors of less than ±1 levels, using the five acceptance levels in the proposed standard. Four of the remaining nine ratings had mean ratings larger than the most severe level so they would fall automatically into that rating. Therefore the standard appeared to work meaningfully in evaluating 24 of the 30 radiographic evaluations. In the stage 2 of the study, much lower overall errors due to both repeatability and reproducibility were found; ±0.25, ±0.62, and ±0.36 levels for the three radiographs. The smaller error of the in-house study is assumed due to the readers being instructed personally on the procedures and careful control of how the measurements were made. It is concluded that the new standard is viable. Nevertheless, it could still be improved by clarifying when indications are to be combined together in the measurement process of the standard. Also, the use of a strip of a prescribed width rather than a line to make the measurements would decrease the sensitivity of the measurements to the position of the line when indications are aligned along the direction of interest.

I. INTRODUCTION

At last years' SFSA Technical and Operating Conference a new quantitative standard for radiographic non-destructive evaluation of steel castings was presented [1]. This new standard is based on measuring the fractional length of indications on the radiograph along a specified direction, the "direction of interest" (DOI). The maximum fractional length found is compared with an acceptance criterion, a maximum allowable fractional length, which is specified by the designer of the casting. The designer determines the acceptance criteria by assigning a maximum allowable indication fraction based on their own assumptions of its effects on performance. It is therefore up to the designer to decide how the indications affect material

properties and relate to loading, stresses and the ultimate performance of the casting in service. By making these design assumptions more or less conservative, the standard can be used to ensure that the indications present on the radiographic film will not limit the component performance to less than the designer's requirements.

The inability of the current ASTM radiographic testing (RT) standards (ASTM E186, E280, E446) to provide any relationship between rating level and part performance arises in large part from their subjectivity. Also, the current RT standards are only relevant to workmanship and cannot be made relevant to performance. In the current standards, a subjective comparison is made between the test radiograph and the standard radiographs. The evaluator (or reader) is further required to prorate the area of interest on the test radiograph to the reference radiographs. Disregarding gray levels on the radiographs, the reader assigns the rating based on the reference radiograph that most closely matches the test radiograph. It has been demonstrated in an earlier repeatability and reproducibility (R&R) study [2] that readers have difficulty distinguishing the rating levels, and image analysis of the ASTM E186 reference radiographs revealed there is significant quantitative similarity between levels [2]. In this study the average confidence interval from the gage R&R study of 128 radiographs was ±1.4 levels. The failure to determine the radiographic quality level of a component with sufficiently small repeatability and reproducibility errors and the inability to relate ASTM levels to performance has led to the development of this new standard.

In this paper, a two stage gage R&R study using the new standard is presented. In the first stage of the study, the reader-to-reader reproducibility is analyzed. Here ten SFSA foundries used the new standard to "rate" the ASTM E186 standard radiographs using the horizontal and vertical directions as the DOIs. Since the radiographs were only rated once in each DOI, repeatability errors could not be determined. In the second part of the study, seven members of the Solidification Laboratory at the University of Iowa rated three radiographs four times with a slight variation in the measurement process one of those four times. In this second part of the study, repeatability and reproducibility errors were determined.

II. PROCEDURES

New Standard Procedure for Radiograph Rating

Although the procedure followed in the new standard is given in great detail elsewhere [1], the key section of the draft standard that describes the rating procedure is [1]: *"The length (l_i) of each indication within the area of interest, along a continuous straight line oriented in the direction of interest, is measured. If the distance between two indication lengths is smaller than the length of the smaller indication, the two indications, together with the space between the indications, are treated as a single indication. The total indication length is obtained as the sum of all indication lengths on the straight line. The maximum total indication length (l_{im}) on any such single straight line is used to assess acceptance of the area of the casting being evaluated. This maximum total indication length (l_{im}) is divided by the specified feature length (L_f) to calculate the maximum indication fraction F ($F = l_{im} / L_f$)."*

In the new standard, no distinction is made between different types of discontinuities (porosity, holes, shrinkage, inclusions, etc.). Only indication lengths longer than 1/16th in (1.6 mm) are considered relevant to the rating. Cracks, defined as an indication on the radiographic film with a length that exceeds 10 times the width, are unacceptable. The area of interest may be just a portion of, or the entire test radiograph. To rephrase and repeat the procedure, the length of the radiographic indications (l_i) are measured in the specified area of interest along straight lines oriented in the DOI. The reader determines the maximum total indication length (l_{im}) summed along a line oriented in the DOI by shifting the line (left-right as shown for example in Figure 1) until the maximum total indication length l_{im} is obtained (also see Figure 1). In practice, the direction of interest could be recorded on the radiographic film by placing an appropriately oriented lead wire on the casting section when the radiograph is made.

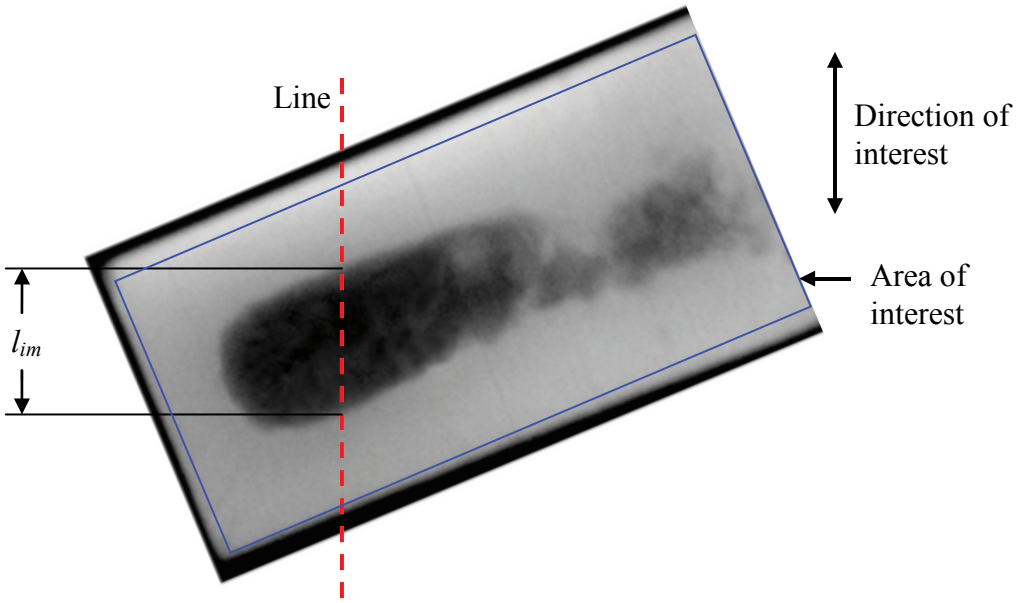


Fig. 1 — Example of the measurement of the maximum total indication length, l_{im}, on a radiograph.

The maximum total indication length (l_{im}) is then divided by some feature length (L_f) to obtain the maximum indication fraction (l_{im}/L_f). The feature length (L_f) is specified by the designer of the casting, or requestor of the RT rating. It (L_f) can be the casting section thickness, a casting feature dimension, or anything the requestor intends to use to relate the RT rating to performance. The maximum indication fraction (l_{im}/L_f) is the basis for the levels of acceptance in the standard. It is currently proposed to have five acceptance levels (1 through 5) corresponding to the 10%, 20%, 30%, 40% and 50% indication percentage *F[%]* levels, respectively, as is shown in Figure 2. In terms of indication fraction, the proposed range of a level is 0.1.

Finally, to repeat for emphasis for the case of multiple indications along lines in the DOI, if the distance between two indication lengths is smaller than the length of the smaller indication, the two indications (together with the space between the indications) are treated as a single indication.

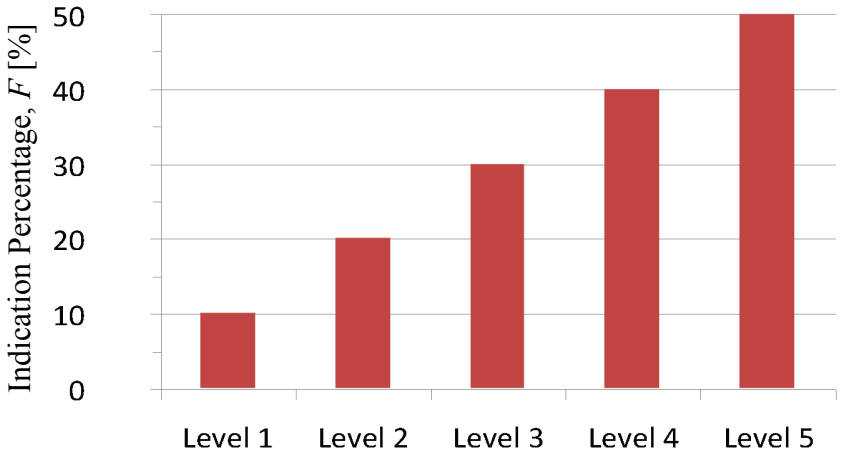


Fig. 2 — Five proposed acceptance levels for indication percentage.

Overview of Gage R&R Procedure

The purpose of the gage R&R was to measure the variability in the measurements and ratings resulting from applying the new standard. While a gage R&R can be done through a variety of methods, here the errors due to reproducibility and repeatability errors are calculated using the ANSI/ASME Power Test Codes 19.1 [3] test uncertainty standard as presented in Figliola and Beasley [4]. In addition, for the in-house portion of the study, analysis of variance (ANOVA) is used to test whether the proportion of the "between readers" variation to "within all readers" variation is statistically significant. If this value (calculated F-statistic) is large relative to the critical F-statistic for the degrees of freedom and probability level selected, then the reader variability is a statistically significant error that should be remedied.

Factors affecting errors in the new standard, our "measurement system", include

- the measurement device (ruler) used: its $1/16^{th}$ inch resolution, how much does it obscure the indications and is it easy to align and keep aligned with the DOI?

- the reader-to-reader variability or "operator error": includes a reader's ability to make accurate measurement, and read, understand and follow the standard procedure.

- how the measurements are made: factors include alignment of measurement device, lighting of the radiographs and room used to make measurements, fixture used for radiographs and measurement devices, how the data is recorded.

- the radiographs themselves: some x-rays are easier to read than others, the reader's rating can also be sensitive to the alignment of the indications relative to the DOI as will be shown.

It is not possible to execute an experimental test matrix to determine separate error contributions from all the factors listed above. The primary goal of this gage R&R was to establish the overall

repeatability and reproducibility errors resulting from typical applications of the standard and compare them. After this, the reasons for these errors can be explored, and recommendations can be made to reduce these errors and improve the standard.

Procedure for Stage 1 of Gage R&R: SFSA Members

Ten SFSA member foundries participated in the reproducibility study of the new standard by evaluating the ASTM E186 standard radiographs for shrinkage indications according to the new standard. These were chosen because many SFSA member foundries have them in hand. Evaluations were made using the horizontal and vertical directions as the DOI on the standard radiographs, and the radiograph dimensions in those respective directions were used as the specified feature lengths L_f. The readers were provided with a handout containing the new standard, their instructions, and a sheet to record their measurements (the maximum total indication length (l_{im}), the feature length (L_f) and the maximum indication fraction (l_{im}/L_f)). Given that there are three radiograph types (CA, CB and CC), five levels of severity of each type, and two DOIs; thirty radiographic ratings were made by each reader. Since the ratings were performed once by each reader, repeatability error could not be determined.

Procedure for Stage 2 of Gage R&R: Solidification Lab Members

The second stage of the Gage R&R study was conducted to investigate repeatability errors and compare them to reproducibility errors. In addition, to examine reader-to-reader differences in evaluating the radiographs in more detail, a procedure was developed to record the indications observed by the readers and the location on the radiograph where the maximum indication length was measured. Three radiographs of 5 inch wide by 1 inch thick plates from feeding distance plate trials were evaluated according to the new standard in this stage of the study. Two of the radiographs appeared to be qualitatively similar, and the third appeared to have noticeably more shrinkage indications. The radiographs were assigned numbers, with Radiograph #1 being the one with the lesser indications, Radiograph #2 the one with the most indications, and Radiograph #3 the other one with less severe indications. These radiographs are shown in Figures 3 through 5. The DOI is taken to be the width direction, as indicated in each of these figures. Seven members of the Solidification Laboratory at the University of Iowa evaluated the radiographs four times. Like the SFSA members in Stage 1, they were given a handout that included the new standard, and some examples of applying it [1]. Unlike stage 1, they were also instructed individually on using the standard in a brief question and answer session. In three of the four evaluations, the readers placed a transparent cover sheet (clear Dura-Lar film, 14” x 17”, 0.005” thick) over the radiograph and outlined the indications observed on the radiograph, as defined according to the new standard, before measuring them. In the other evaluation, the readers measured the radiographic indications without the cover sheet. The four readings were performed over a four week period to allow sufficient time between repeated measurements. The readers were provided with a data sheet on which they indicated all possible maximum indication lengths, with the entries ordered from the top of the radiograph to the bottom. At each indication position along the radiograph length (the position of the line in the DOI), that data was recorded on the sheet by placing a mark on the transparency sheet off to the side of the radiograph. For the case without the cover sheet, they placed a non-marking arrow sticker at the indications measured, so that the locations could be recorded. This way the data could

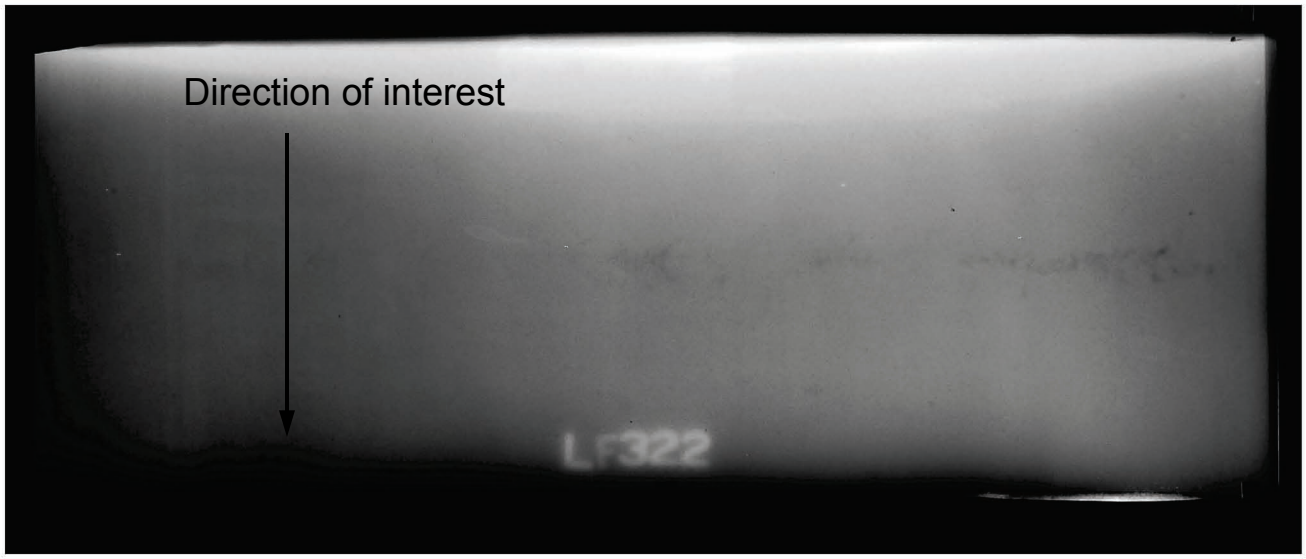


Fig. 3 — Radiograph #1 used in our “in-house” rating study.

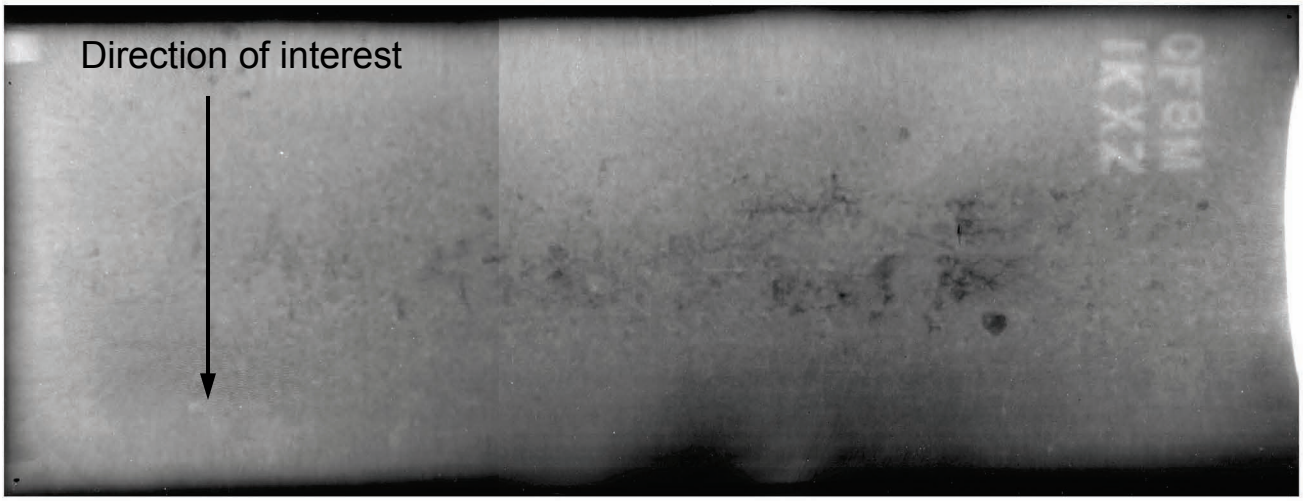


Fig. 4 — Radiograph #2 used in our “in-house” rating study.

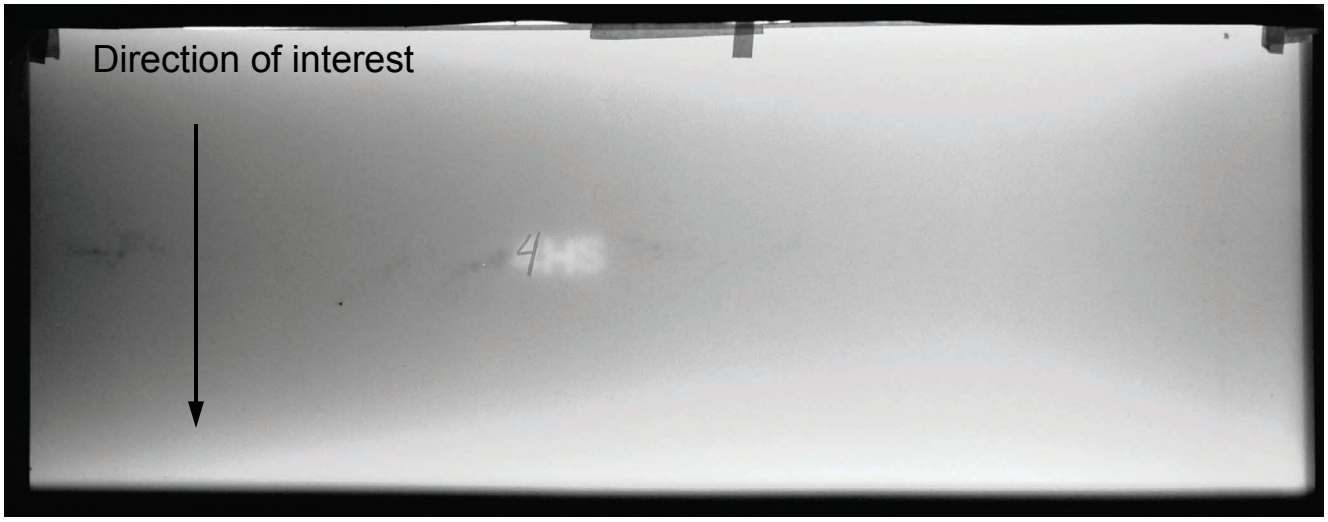


Fig. 5 — Radiograph #3 used in our “in-house” rating study.

be checked later since the indication and its measurement could be tracked. Only after all possible maximum indication lengths were measured and recorded was the determination made as to which was the largest. Finally, the time required for each radiographic rating was recorded for each rating.

III. RESULTS

Results from Stage 1 of Gage R&R by SFSA Members

The maximum indication fractions measured by the SFSA members in the first stage of the gage R&R study are presented for each ASTM indication type and DOI used in Figures 6 through 11. The results are categorized by the five severity levels for the ASTM shrinkage radiograph type and DOI. As indicated in Figure 6 at each severity level, the mean of the ten SFSA member measurements is given by the white bar at the left of the data, and the error bars on the mean are ± one standard deviation of the ten measurements. The ten reader measurements are ordered according the reader as indicated in Figure 6, so a given reader can be tracked.

For the CA and CB radiographs, in Figures 6 through 9, there is considerably more scatter than for the CC radiographs, in Figures 10 and 11. The sponge-type shrinkage indications in Figures 10 and 11 are well defined when compared to the vein-type indications in the CA and CB radiographs. Notice in Figures 10 and 11, that the scatter and standard deviations are more consistent for the CC type shrinkage, and the indication fractions consistently increase with severity level regardless of the DOI (horizontal in Figure 10 and vertical in Figure 11). There are

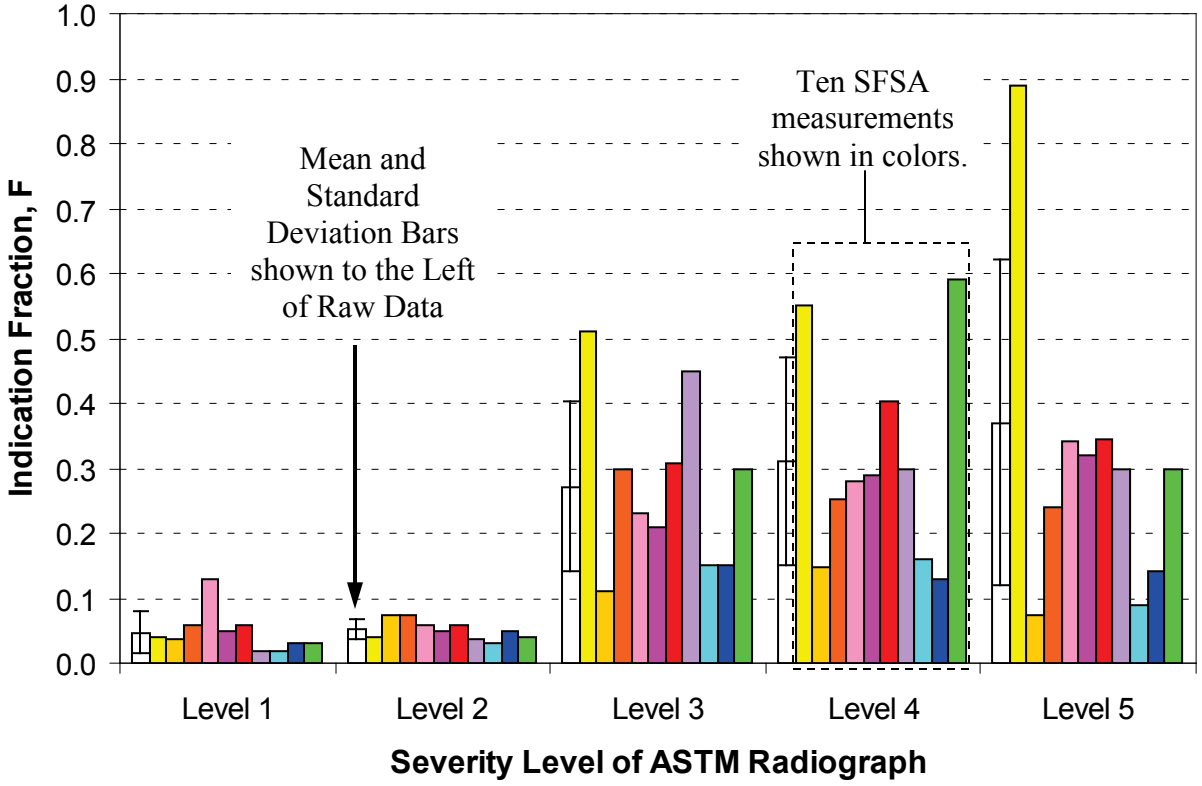


Fig. 6 — Indication fraction for the ten SFSA measurements for shrink type CA, horizontal direction.

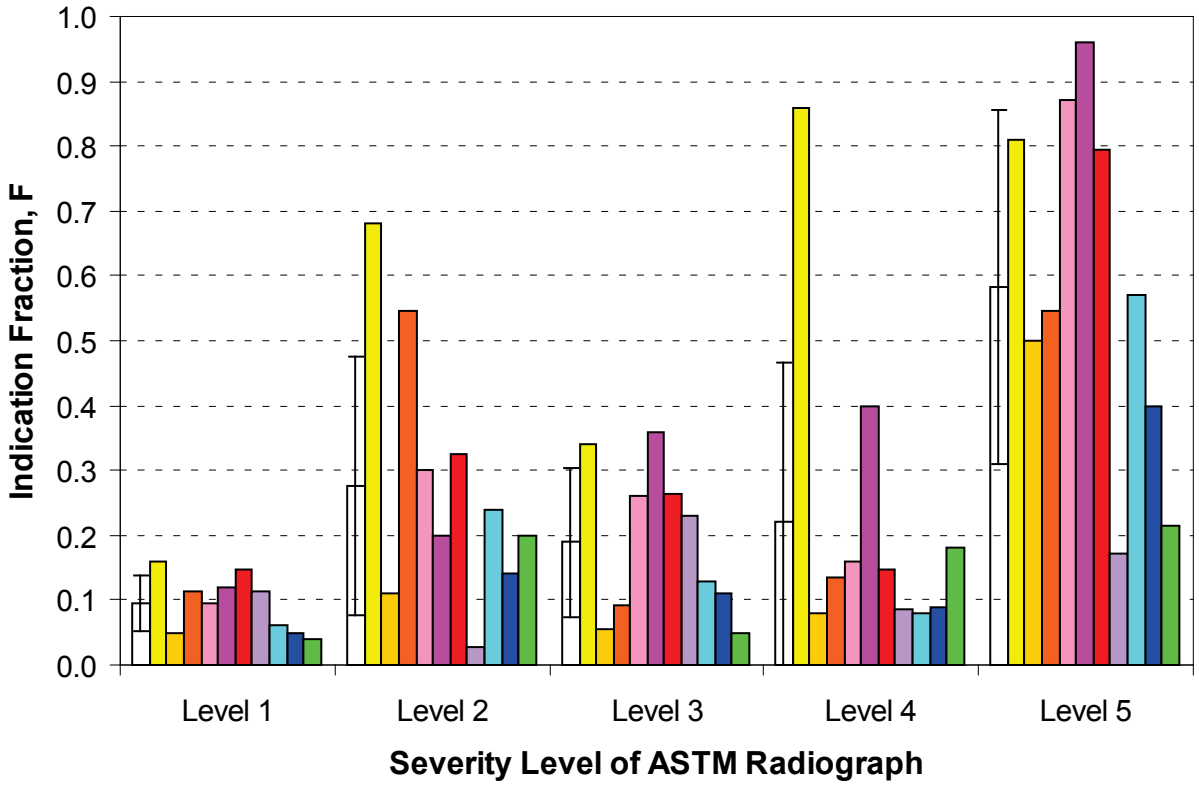


Fig. 7 — Indication fraction for the ten SFSA measurements for shrink type CA, vertical direction.

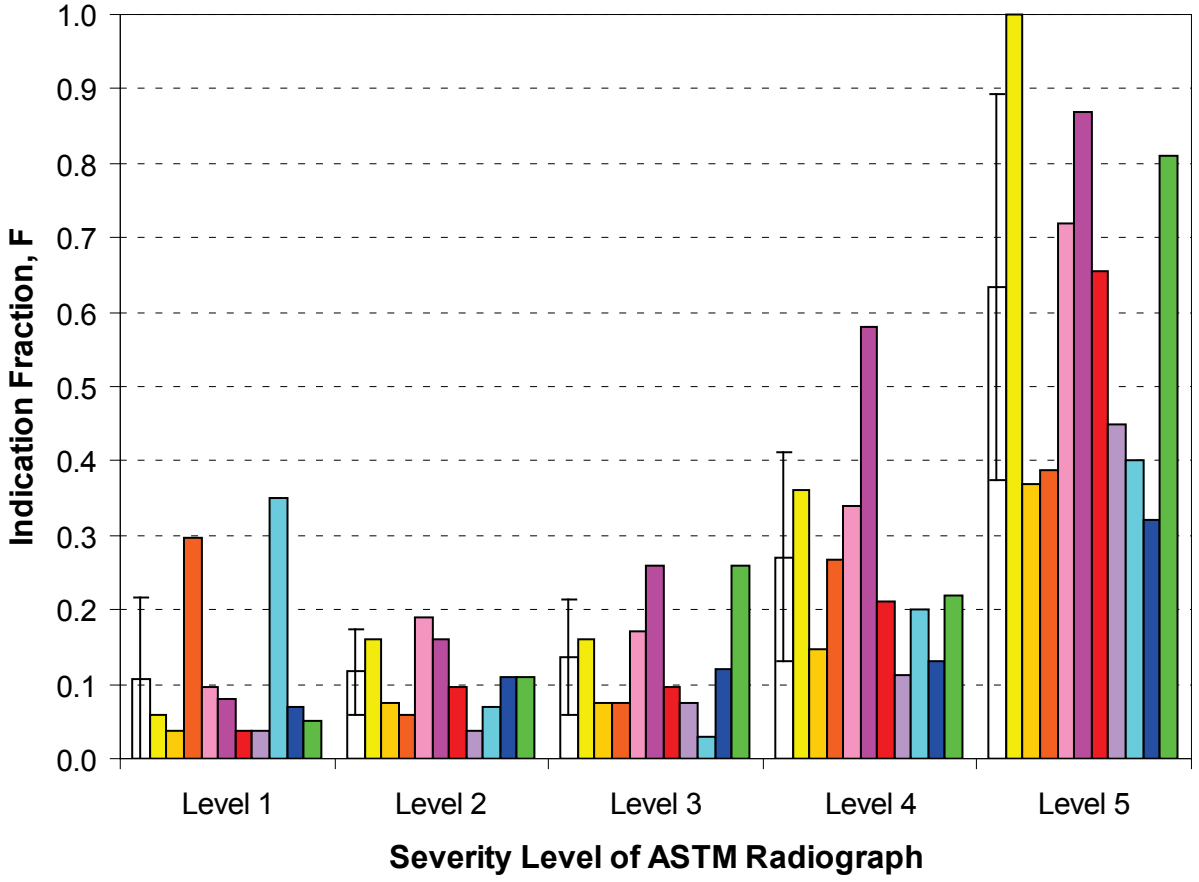


Fig. 8 — Indication fraction for the ten SFSA measurements for shrink type CB, horizontal direction.

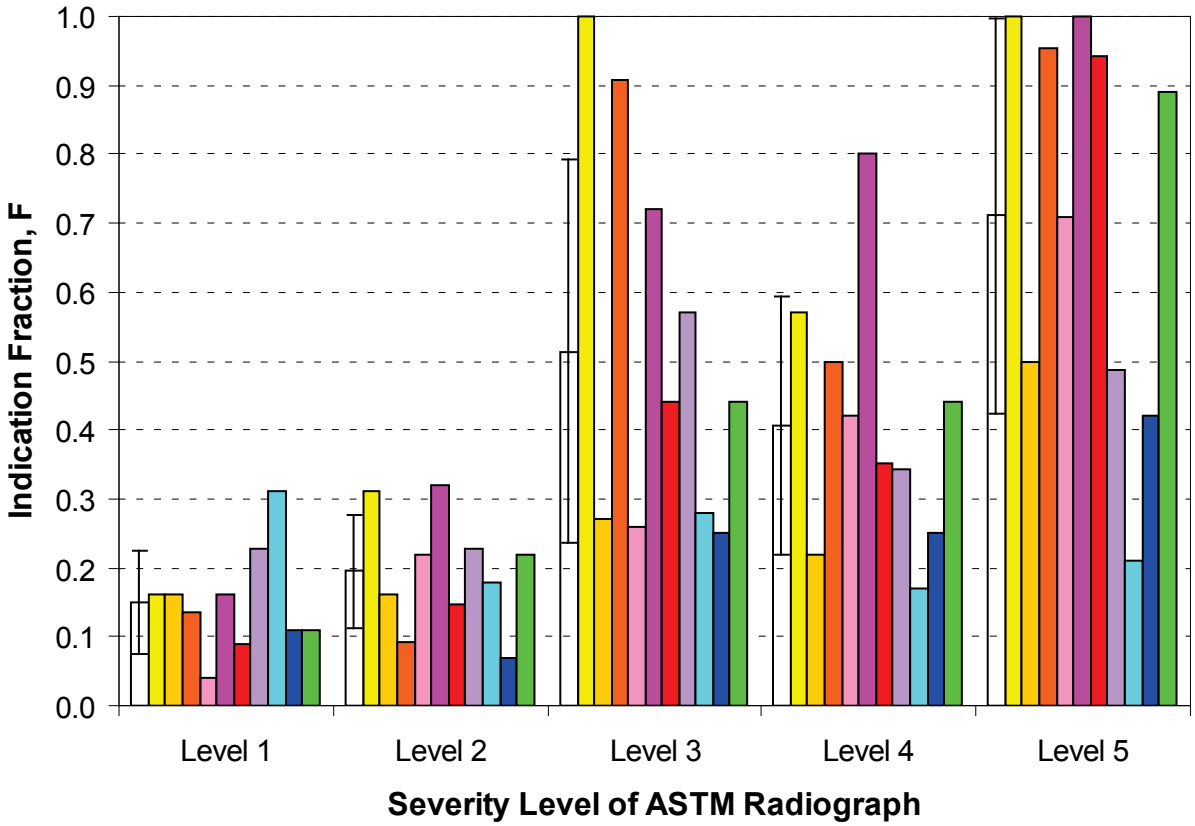


Fig. 9 — Indication fraction for the ten SFSA measurements for shrink type CB, vertical direction.

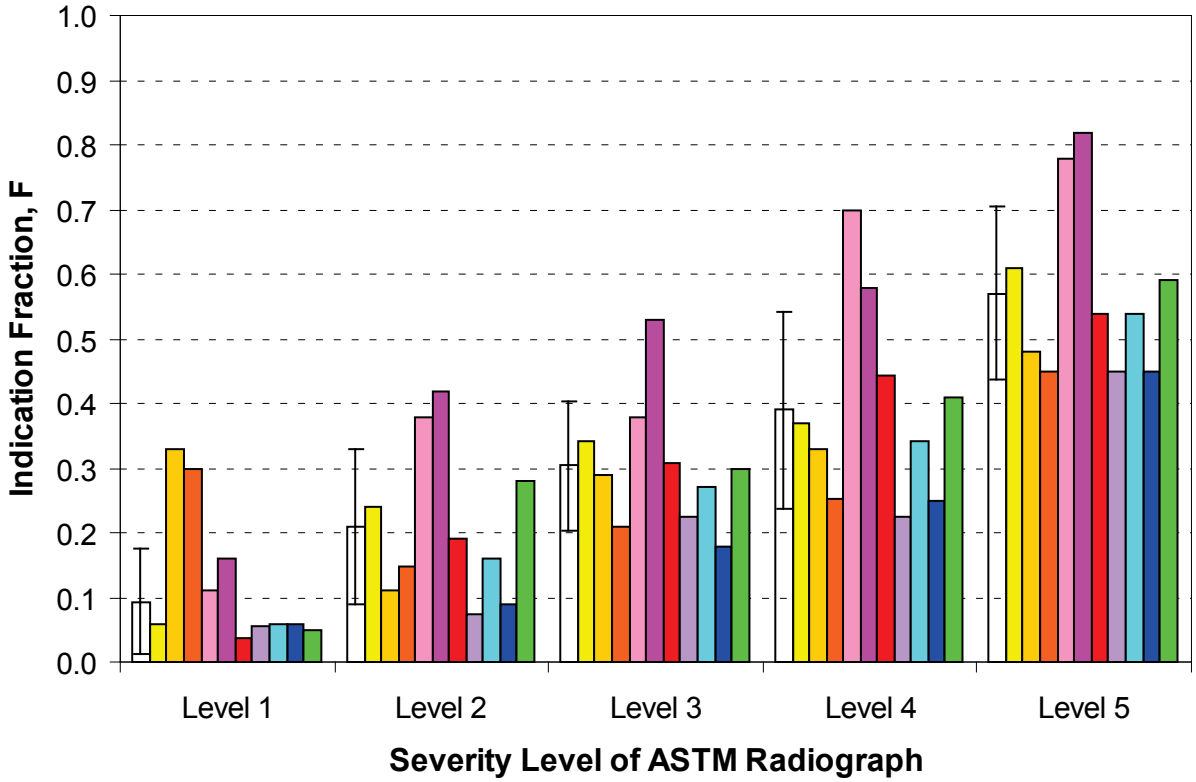


Fig.10 — Indication fraction for the ten SFSA measurements for shrink type CC, horizontal direction.

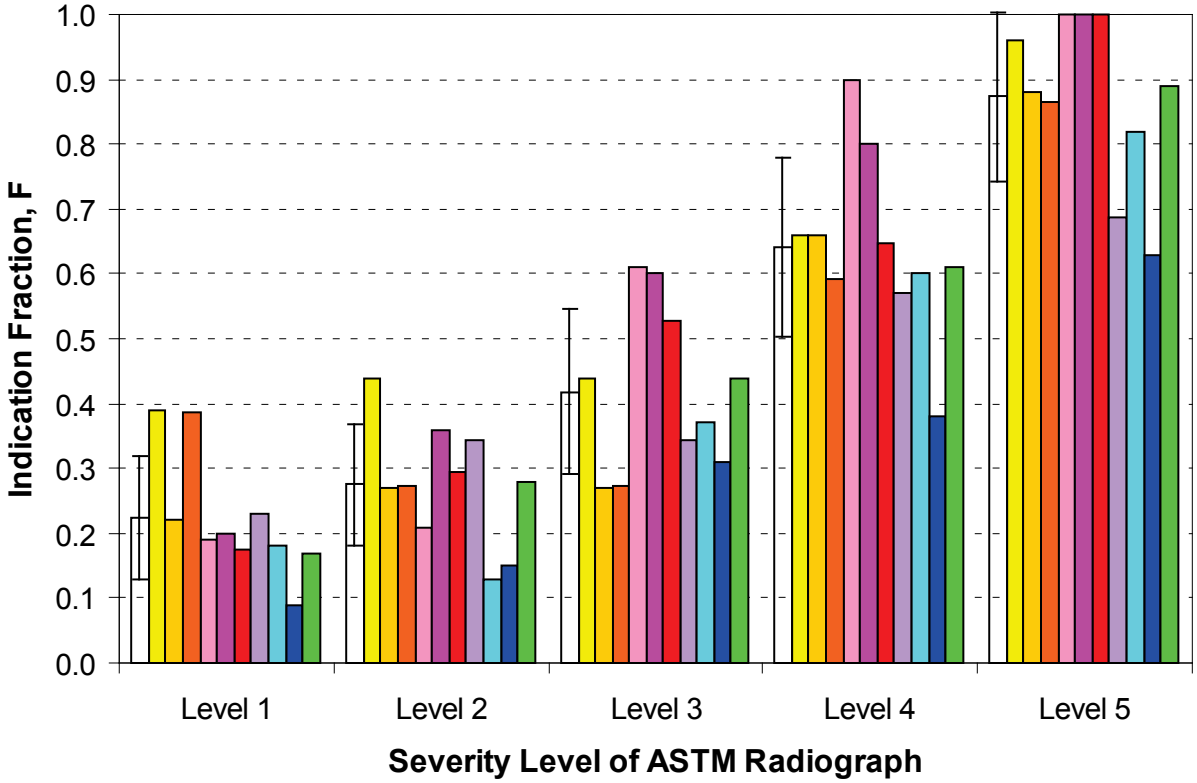


Fig. 11 — Indication fraction for the ten SFSA measurements for shrink type CC, vertical direction.

large discrepancies for a given severity level for the CA and CB type shrinkage depending on the DOI. There is also greater disagreement between the readers for CA and CB shrinkage for a given DOI. In particular, note the difference for type CA shrinkage between the horizontal DOI (figure 6) and the vertical DOI (Figure 7) for severity level 2. In Figure 6 the readers are all in agreement the indication faction is small; the mean is about 0.05 and the standard deviation is very small relative to most other radiographs. In Figure 7, however, the mean rating for the vertical direction is quite large for severity level 2, and the disagreement and standard deviation is also quite large. The reason for this can be readily explained if we examine the standard radiograph for CA level 2 shown in Figure 12. Since the indications are aligned in the vertical orientation, the maximum indication fraction in that direction should be larger. Also, the vertical alignment causes the measurement to be quite sensitive to the position of the line of measurement in the DOI. A small change in the position of the line of measurement can change the indication fraction considerably, and this is reflected in the large scatter seen in the level 2 data in Figure 7. Observe in Figure 13, that the standard radiograph for level 3 type CB shows a similar vertical alignment, and in Figure 8 for the horizontal DOI the mean value and standard deviation for level 3 type CB are both relatively small. In Figure 9, for the vertical DOI, the mean and standard deviation are large, just as was seen for type CA level 2.

In Figures 6 through 11, 21 of the 30 radiograph ratings had fairly low reproducibility errors of less than ±1 levels (indication fraction ±0.1 using the five acceptance levels in the proposed standard). Four of the remaining nine ratings had mean ratings larger than the most severe level, so they would fall automatically into level 5. Therefore the standard meaningfully evaluates 24 of the 30 radiographic evaluations. Additional experience with the standard will improve this.

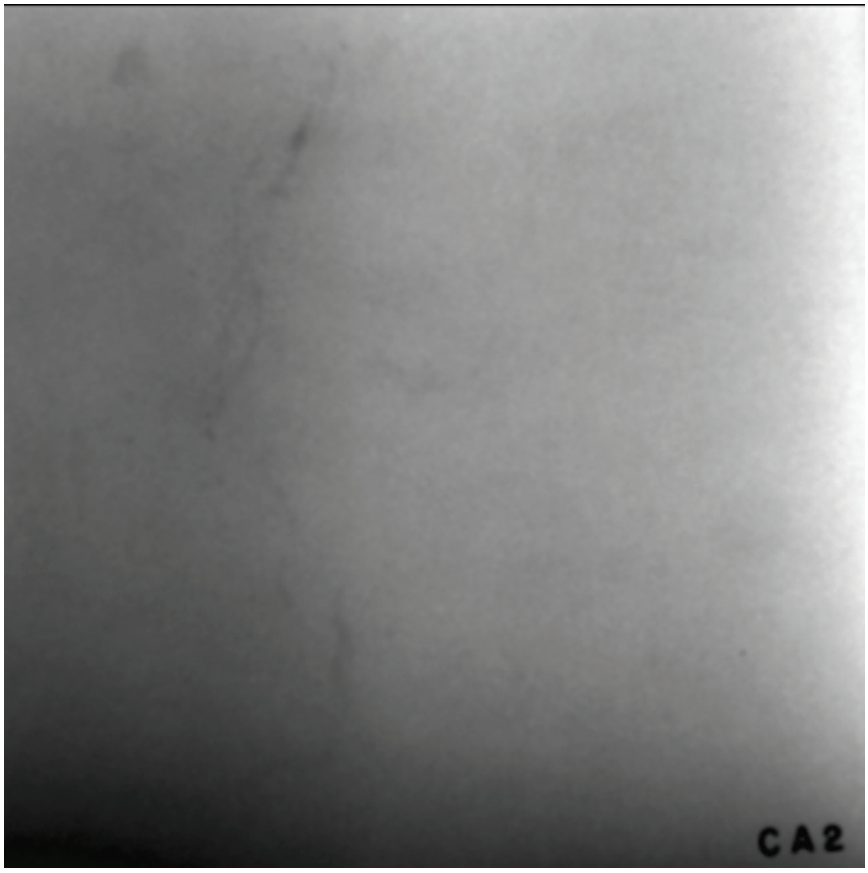


Fig.12 — Radiograph from ASTM Standard E186, type CA level 2. Note vertical alignment of indications.

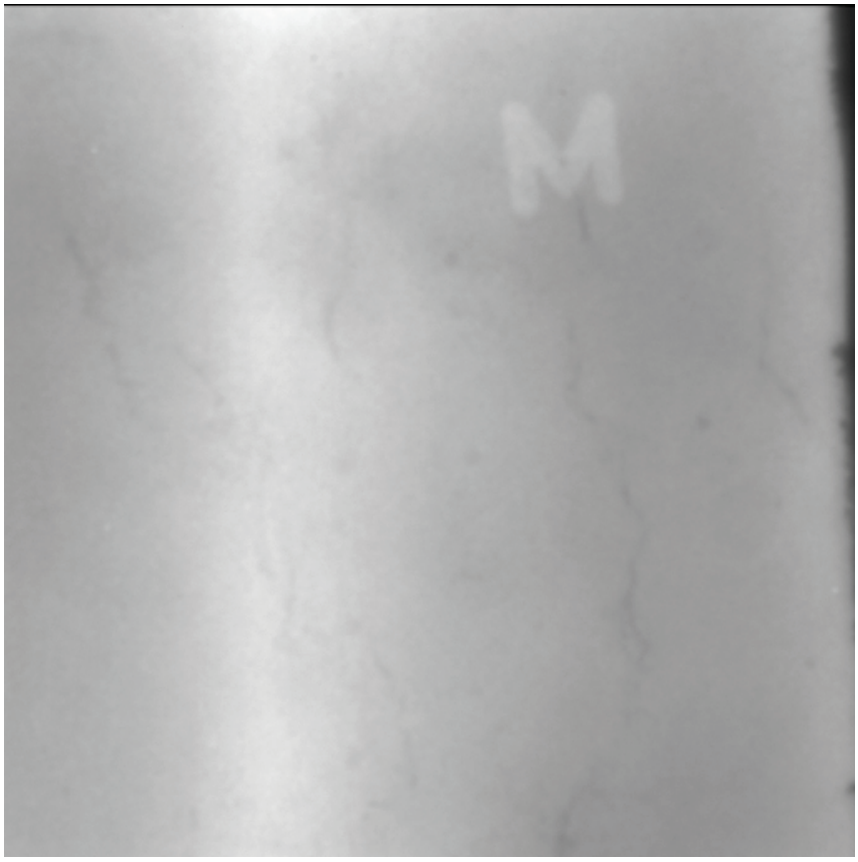


Fig.13 — Radiograph from ASTM Standard E186, type CB level 3. Note vertical alignment of indications.

In this stage of the study, the only error that can be determined is due to the reader-to-reader reproducibility. Using $S_{F,1}$ to denote the standard deviation for the reader ratings, the error or 95% confidence uncertainty interval $U_{F,1}$ for the true value of a given indication fraction rating is given by

$$U_{F,1} = \pm \frac{S_{F,1}}{\sqrt{M}} \cdot t_{\nu,P} \tag{1}$$

where M is the number of readers (10) and $t_{\nu,P}$ is the Student-t statistic for ν degrees of freedom and P % probability. In this case, $\nu = 9$ and $P = 95\%$, so $t_{9,95} = 2.262$. The results for the mean and confidence interval (as error bars) are plotted for all types and levels of indications in Figures 14 and 15, respectively, for the horizontal and vertical DOIs. These results can be viewed from two different perspectives. From one perspective, the results simply represent the ratings and reproducibility errors from thirty radiographic evaluations made using the new standard. From another perspective, one might view this as a quantitative evaluation of the ASTM E186 standard radiographs. From this second perspective there is significant overlap in the levels, given the magnitude of the confidence intervals. Generally, but not always, ratings made for the vertical DOI are larger than for the horizontal DOI, and the indication fraction typically increases with the severity level, but, again, not in every case.

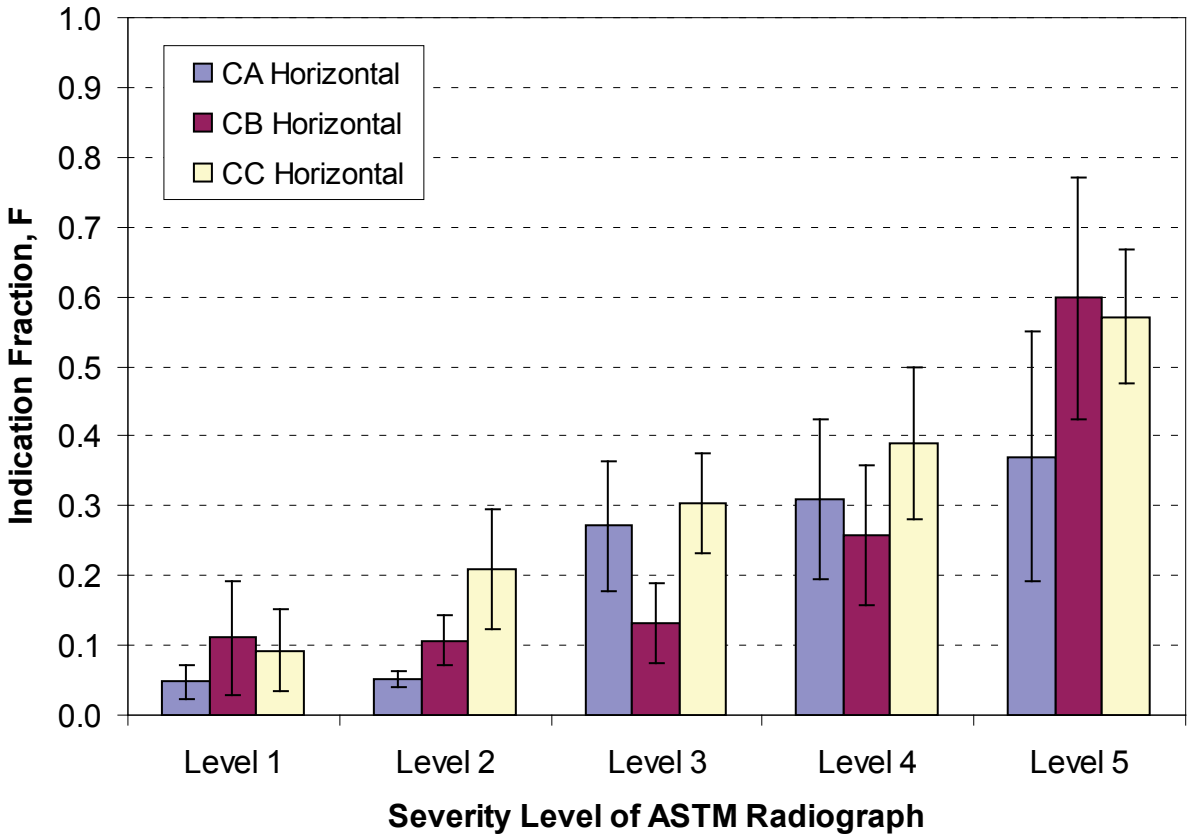


Fig. 14 — Mean indication fractions from SFSA measurements of ASTM radiographs for horizontal direction of interest; all shrink types. Error bars are 95% confidence intervals.

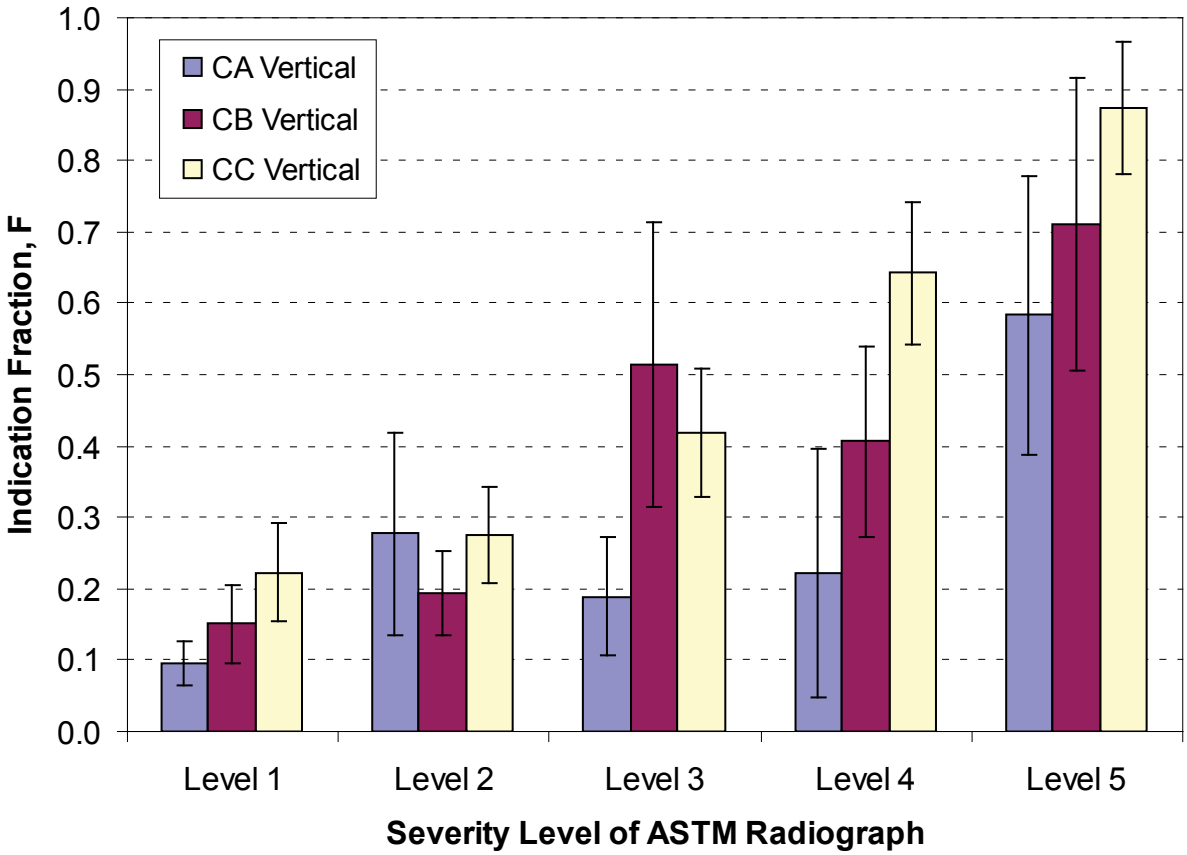


Fig. 15 — Mean indication fractions from SFSA measurements of ASTM radiographs for vertical direction of interest; all shrink types. Error bars are 95% confidence intervals.

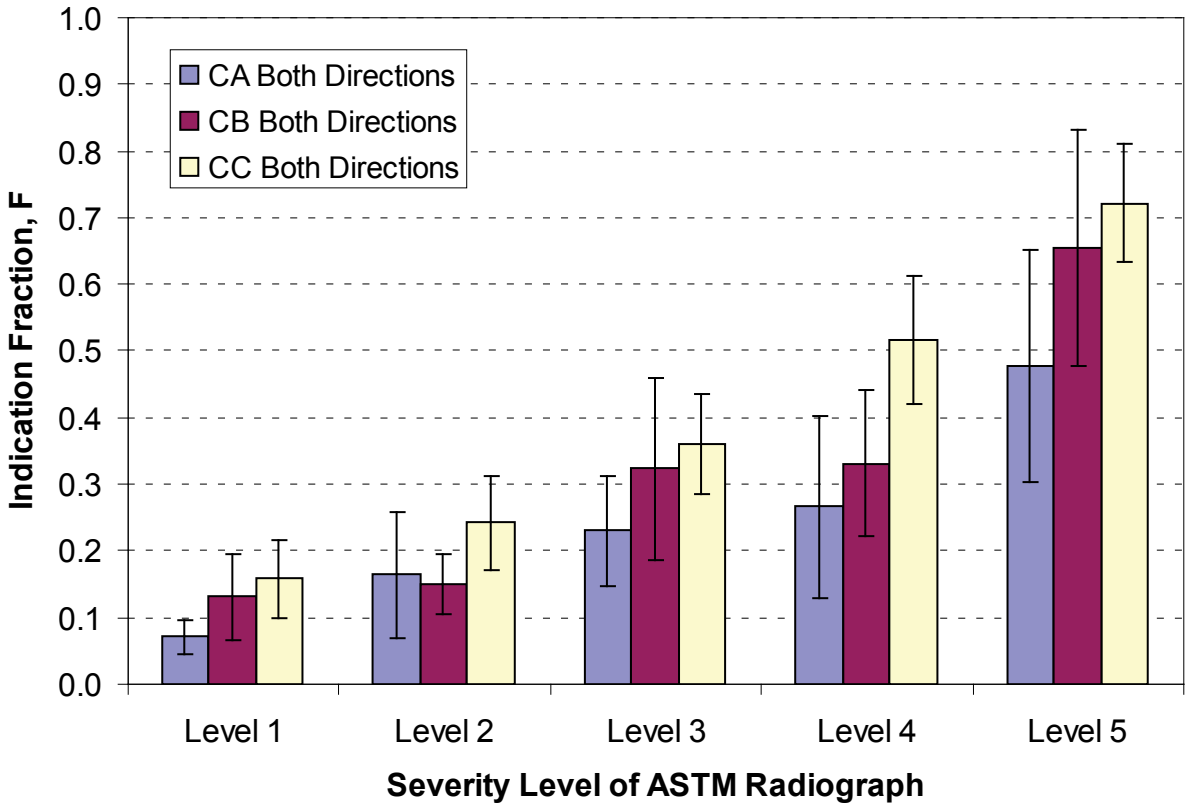


Fig. 16 — Mean indication fractions from SFSA measurements of ASTM radiographs after combining horizontal and vertical direction data for all shrinkage types. Error bars are 95% confidence intervals.

When both horizontal and vertical results are combined using pooled statistics (this combines the two direction results treating them as replicated measurements), the pooled standard deviation for a given radiograph level and type is

$$S_{F,1} = \sqrt{\frac{\sum_{n=1}^{2} S_{Fn}{}^{2}}{N}} \tag{2}$$

where n is the index for the two directions, N is the total number of DOIs ($N = 2$), and S_{Fn} is the standard deviation for the horizontal ($n = 1$) and vertical ($n = 2$) directions. The pooled uncertainty then becomes

$$U_{F,1} = \pm \frac{S_{F,1}}{\sqrt{MN}} \cdot t_{\nu,P} \tag{3}$$

where now $\nu = (N)(M\text{-}1) = (2)(10\text{-}1) = 18$, so that $t_{18,95} = 2.101$ and $S_{F,1}$ comes from equation (2). The combined mean for both DOIs and the error bar confidence interval from equation (3) is plotted in Figure 16 for all levels and types of shrinkage. By pooling the data in this way, the data is closer to an overall quantitative evaluation of the radiographs. Here, the indication fraction for type CA is less than for types CB and CC for all levels except level 2, where CA and CB are essentially equal. Type CC shrink has consistently the highest indication fraction. These findings mirror the results of a detailed image analysis of the same ASTM reference radiographs performed previously [2].

Pooling the data one more time, and combining all shrinkage type results (CA, CB and CC), the levels of the ASTM radiographs can be compared using the indication fraction. The calculations follow from equations (2) and (3) with $N = 6$ corresponding to the three types and two DOIs, all combined together. The t-statistic is $t_{54,95} = 2.005$. The results of the pooled mean and confidence interval from the SFSA members' measurements for both directions and all types of shrinkage are given in Figure 17. There are three key observations to note about this figure: 1) the means of the five levels correspond well to the five proposed indication fraction levels shown in Figure 2, 2) the is significant overlap between severity levels 2, 3 and 4, meaning it is difficult to distinguish between them, and 3) severity levels 1, 3 and 5 are clearly distinguishable. It is interesting to note that observations 2) and 3) above were important conclusions of the earlier image analysis study [2].

Results from Stage 2 of Gage R&R by Solidification Laboratory Members

In Table I the mean and standard deviations of the three maximum indication fraction measurements made **with** the transparency cover sheet, and the statistics for the single rating made **without** the cover sheet, are given for the three radiographs used in the study. The data show that radiographs #1 and #3 are quantitatively similar in their severity, and that radiograph #2 is much worse. The raw data for these maximum indication lengths are plotted in Figures 18, 19 and 20 for radiographs #1, #2, and #3, respectively, arranged by the reader number. In these figures, the red solid square symbols give the data measured without using the cover sheet, the red solid line is the mean of that data (value in Table I), and the red dashed line is the standard deviation of that data (value in Table I). For the data taken using the cover sheet, black hollow square symbols are used to plot the median of the three measurements, while the error bars plot

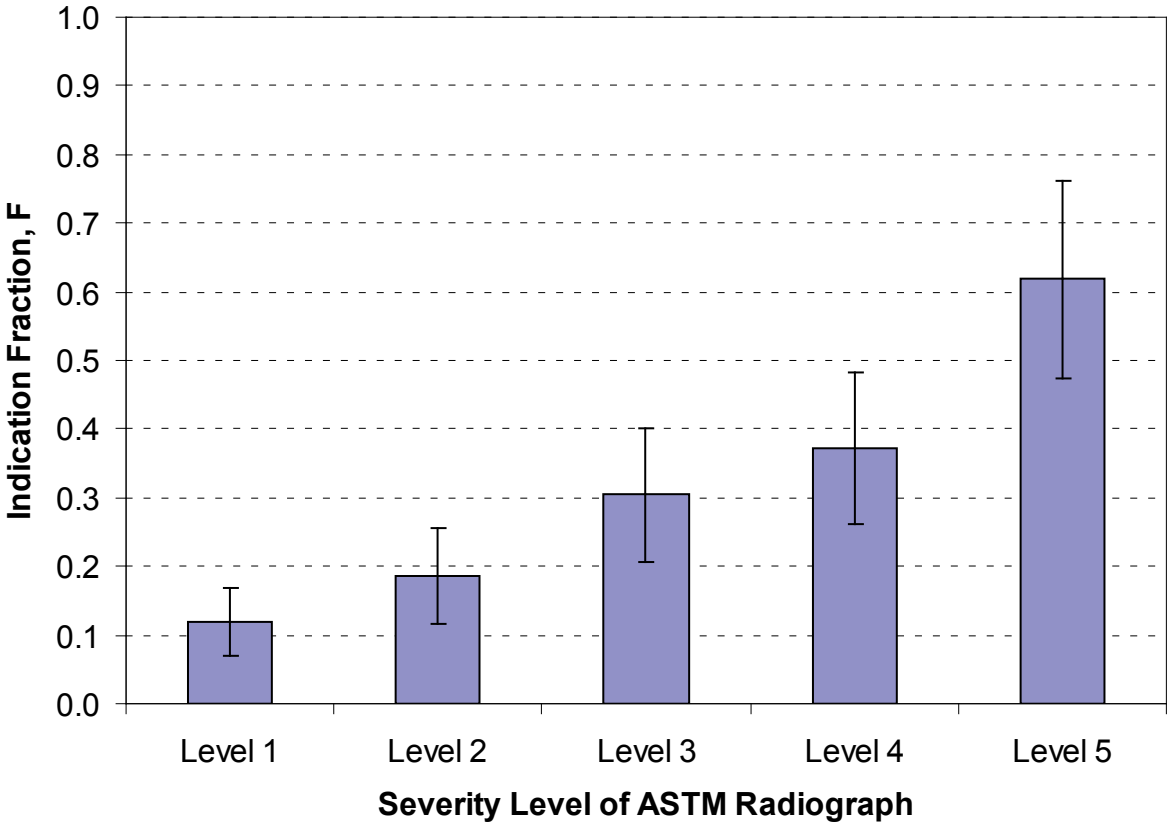


Fig. 17 — Mean indication fractions from SFSA measurements of ASTM radiographs after combining all direction and all shrinkage type data for a given level. Error bars are 95% confidence intervals.

the maximum and minimum values measured. The black solid and dashed lines give the mean and standard deviation, respectively, of the cover sheet data. In some cases, the error bars and median points coincide, which implies that repeated measurements for some readers gave identical results.

The raw data show that it does not matter whether or not the cover sheet is used since an ANOVA test between both methods show that they are not significantly different (P-values > 0.4 for all three radiographs). Based on this, all four readings can be combined to determine the

Table I. Mean and standard deviation of maximum indication lengths across width direction with and without marking indictions on a transparency sheet from in-house gage R&R.

	Measurements Made with Transparency Sheet		Measurements Made without Transparency Sheet	
Radiograph ID	Mean Length (in)	Standard Deviation of Length (in)	Mean Length (in)	Standard Deviation of Length (in)
X-Ray #1	0.58	0.14	0.53	0.11
X-Ray #2	1.77	0.50	1.62	0.31
X-Ray #3	0.50	0.22	0.59	0.23

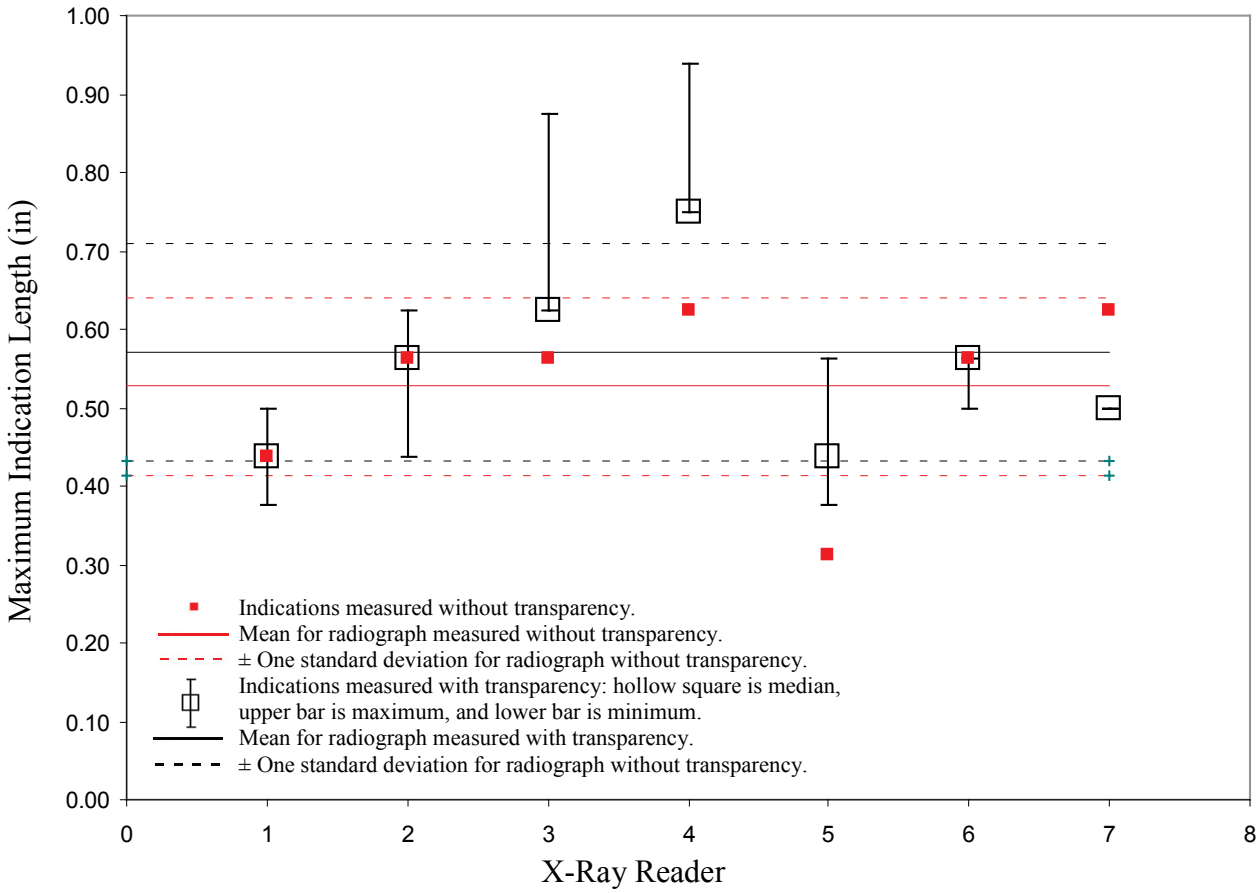


Fig. 18 — Maximum indication lengths measured for radiograph #1 in the in-house study.

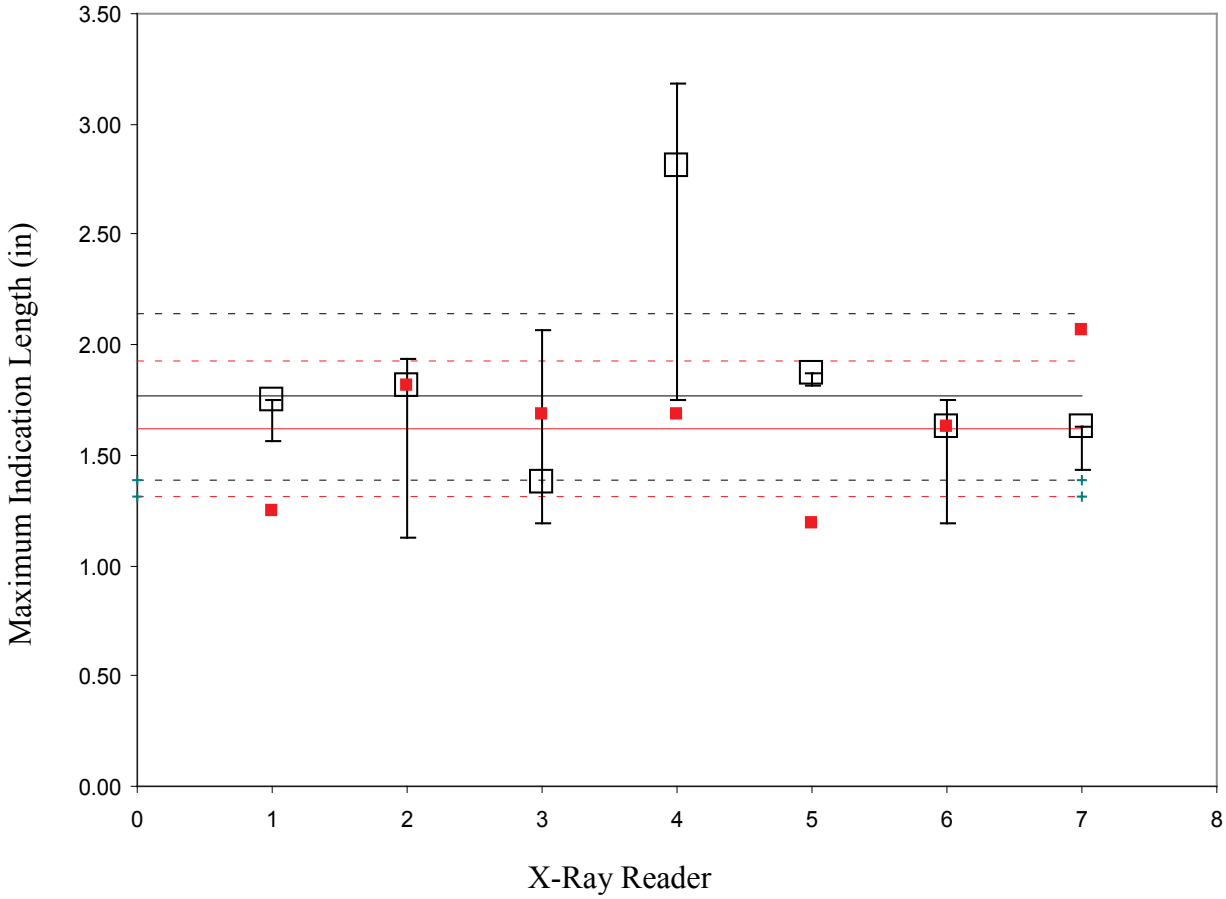


Fig. 19 — Maximum indication lengths measured for radiograph #2 in the in-house study.

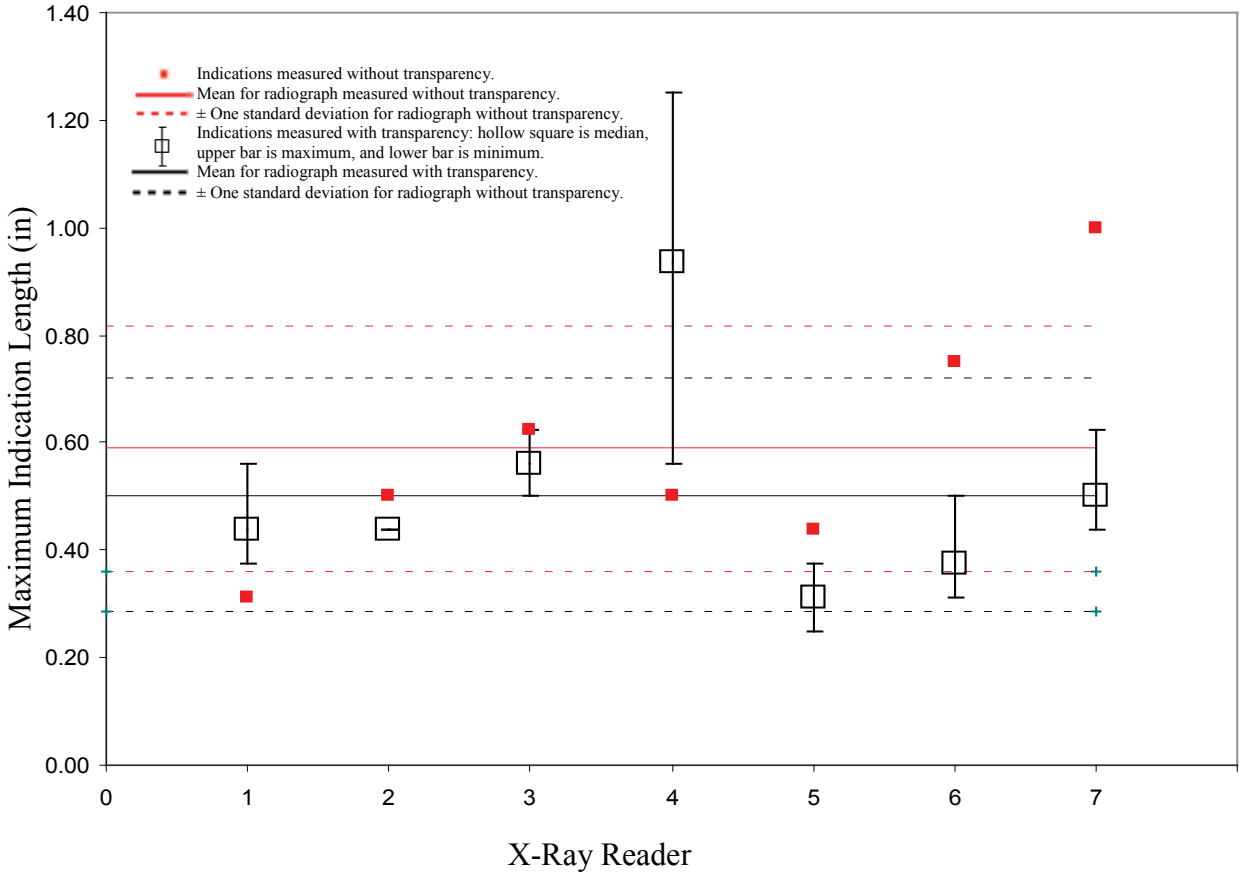


Fig. 20 — Maximum indication lengths measured for radiograph #3 in the in-house study.

repeatability error. In terms of the data of different readers, note that readers #1 and #2 are consistent and usually measure close to the mean. Reader #4 is in many measurements above the mean when using the transparency, but not when using the cover sheet.

Using the radiograph width as the feature length, L_f, the maximum indication fractions were calculated, and these are plotted in Figures 21, 22 and 23 for the three radiographs. Note that in these figures the same scale is used so that the scatter in the data can be readily compared. Symbols and line types used have the same meaning as in Figures 18 to 20.

The results for the time it took the readers to measure and rate the radiographs is given in Table II. The mean evaluation time for a single radiograph is typically around 5 minutes, and always below 10 minutes. Using a cover sheet required an additional time investment, and also caused much more variability. Without the cover sheet, the time to evaluate a radiograph was 30% to 50% shorter. However, tracking this time as the readers progressed from their first to third readings using the cover sheet, the evaluation times decreased dramatically, as seen in Table III. By the third reading with the cover sheet, the evaluation times were about the same as when no cover sheet was used.

In this stage of the gage R&R, three sources of error can be identified contributing to the overall uncertainty in the ratings. The first error is due to the reader-to-reader reproducibility, and as in the first part of the study it will be denoted as $U_{F,1}$. The second error is the individual reader repeatability, which will be designated as $U_{F,2}$. There is a third error $U_{F,3}$ due to the resolution of the measurement instrument used, a ruler with 1/16" resolution. This error is half the resolution, so that with L_f = 5" it is given by

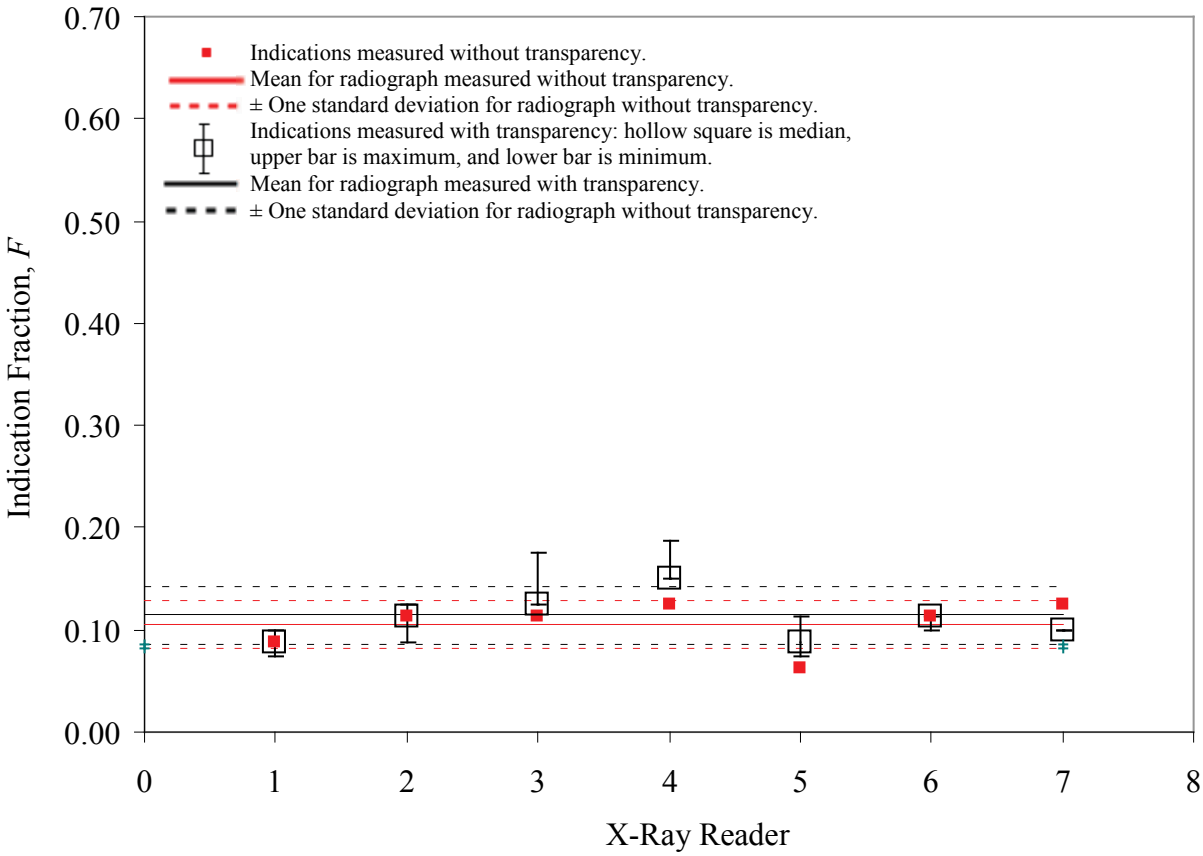


Fig. 21 — Maximum indication fractions measured for radiograph #1 in the in-house study.

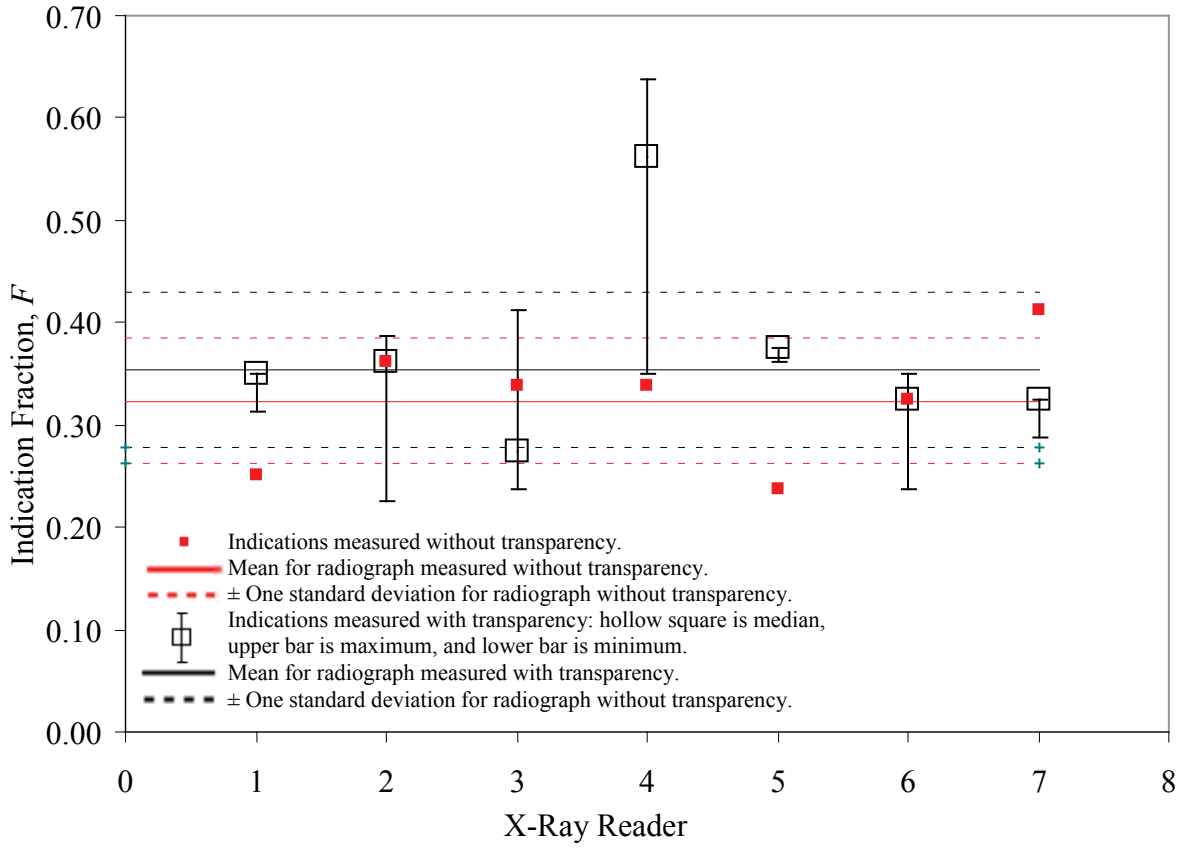


Fig. 22 — Maximum indication fractions measured for radiograph #2 in the in-house study.

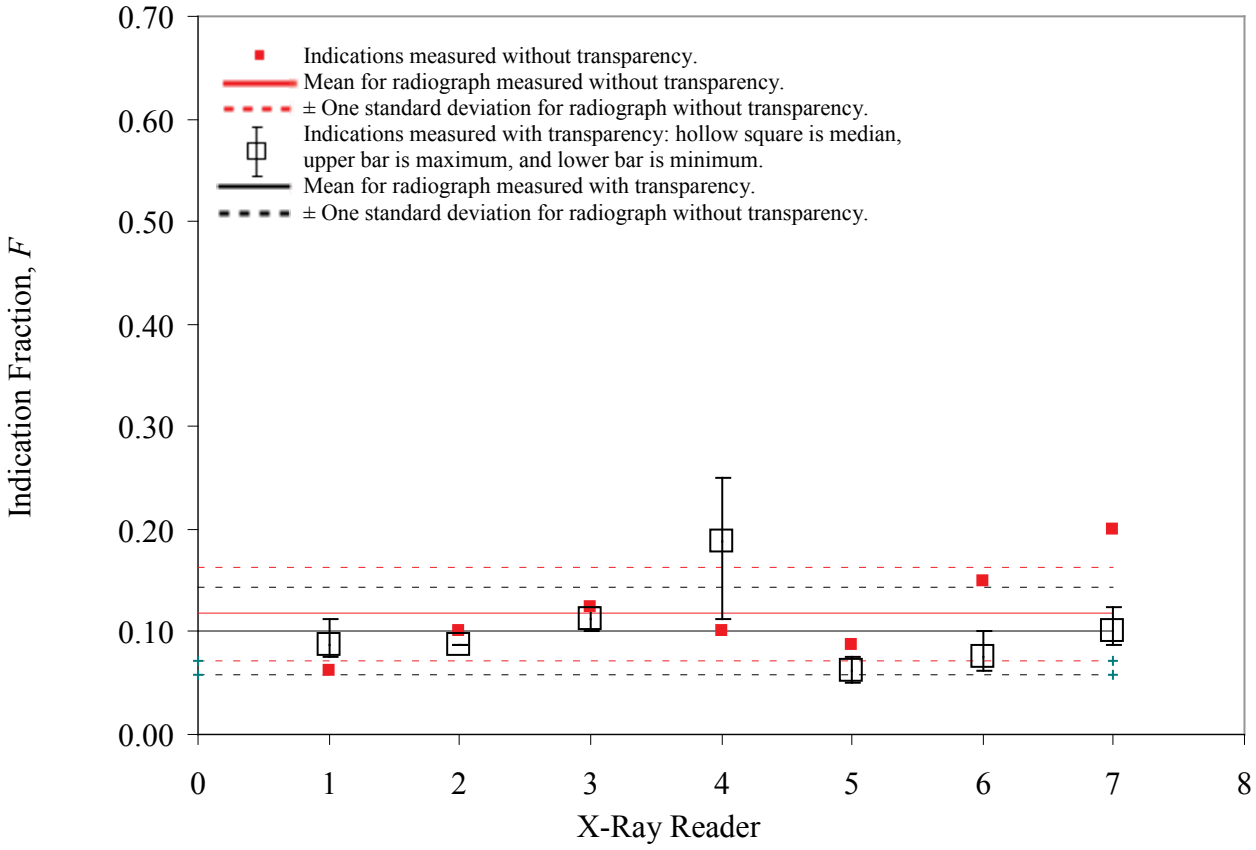


Fig. 23 — Maximum indication fractions measured for radiograph #3 in the in-house study.

Table II. Mean and standard deviation data for time required to measure and rate radiographs with and without marking indictions on a transparency sheet from in-house gage R&R.

	Measurements Made with Transparency Sheet		Measurements Made without Transparency Sheet	
Radiograph ID	Mean Time to Rate (min)	Standard Deviation of Time to Rate (min)	Mean Time to Rate (min)	Standard Deviation of Time to Rate (min)
X-Ray #1	6.86	4.92	3.71	1.38
X-Ray #2	9.62	4.74	6.71	2.14
X-Ray #3	4.71	1.95	4.00	1.15

Table III. Progression of time required from first to third ratings of the radiographs made using the transparent cover sheet

	Measurements with Transparency, First Reading		Measurements with Transparency, Second Reading		Measurements with Transparency, Third Reading	
Radiograph ID	Mean Time to Rate (min)	Standard Deviation of Time to Rate (min)	Mean Time to Rate (min)	Standard Deviation of Time to Rate (min)	Mean Time to Rate (min)	Standard Deviation of Time to Rate (min)
X-Ray #1	10.9	9.3	5.4	2.8	4.3	3.1
X-Ray #2	13.7	7.8	7.6	2.9	7.6	4.8
X-Ray #3	6.3	4.0	4.1	1.9	3.7	2.1

$$U_{F,3} = \pm \frac{U_0}{2 \cdot L_f} = \pm \frac{0.0625"}{2 \cdot 5"} = 0.00625 \tag{4}$$

The standard deviation arising from the reader-to-reader reproducibility is $S_{F,1}$ which is determined from the deviation of a reader's mean indication fraction ($\overline{F}_m$, where the *m* subscript denotes the *m*-th reader) and the overall mean indication fraction for all readers, $< \overline{F} >$, is

$$S_{F,1} = \sqrt{\frac{\sum_{m=1}^{7} (\overline{F}_m - < \overline{F} >)^2}{M - 1}} \tag{5}$$

where *M* is the number of readers, 7. The uncertainty for reader-to-reader reproducibility is determined from

$$U_{F,1} = \pm \frac{S_{F,1}}{\sqrt{M}} \cdot t_{\nu,P} \tag{6}$$

where $\nu = M\text{-}1 = 6$ so $t_{\nu,P} = t_{6,95} = 2.447$.

The repeatability error for all readers $U_{F,2}$ is determined using pooled statistics on the four evaluations of each radiograph. The variable *n* will denote the *n*-th reading with the total number of repeated readings being $N = 4$. The standard deviation for repeatability for the *m*-th reader is S_{Fm} which is given by the deviation between that reader's *n*-th indication fraction measurement F_{mn} and their average from the four measurements $\overline{F}_m$. This is

$$S_{Fm} = \sqrt{\frac{\sum_{n=1}^{4} (F_{mn} - \overline{F}_m)^2}{N - 1}} \tag{7}$$

The pooled standard deviation for the repeatability within all readers is then

$$S_{F,2} = \sqrt{\frac{\sum_{m=1}^{7} S_{Fm}{}^2}{M}} \tag{8}$$

and the repeatability error within all readers $U_{F,2}$ is

$$U_{F,2} = \pm \frac{S_{F,2}}{\sqrt{M \cdot N}} \cdot t_{\nu,P} \tag{9}$$

where $\nu = (M)(N\text{-}1) = (7)(4\text{-}1) = 21$, so for this confidence interval $t_{\nu,P} = t_{21,95} = 2.080$. The total error in the indication fraction $<U_F>$ is determined from the root sum of the squares (RSS) of the individual errors

$$< U_F > = \sqrt{U^2{}_{F,1} + U^2{}_{F,2} + U^2{}_{F,3}} \tag{10}$$

The overall mean indication fraction, errors in indication fraction due to reproducibility and repeatability, and the total error in the mean are summarized in Table IV. This data is plotted in Figure 24 where the overall mean indication fractions and 95% confidence intervals for the three radiographs are given as the white bar to the left of the individual reader data. The mean indication fractions and confidence intervals for each of the seven readers are given in Figure 24 by colored bars, as determined from their four ratings of each radiograph. For all three radiographs, the reader-to-reader reproducibility error $U_{F,1}$ is the largest source of error. In the case of X-Ray #1 and #3 it is twice the repeatability error. Note that the repeatability $U_{F,2}$ for X-Ray #1 is 0.0073 and is only slightly larger than the resolution error 0.00625 which indicates good precision for that radiograph. The readers showed poor repeatability $U_{F,2}$ in evaluating radiograph #2 relative to #1 and #3. We will examine this in more detail shortly. When looking at the plot of data for readers in Figure 24, note that reader #4 has consistently higher measurements than the others, and that the #4 mean measurements are outside the total confidence intervals for each radiograph. It can be argued that the data for reader 4 should be excluded from the analysis, and this would reduce the errors considerably. In terms of the 0.1 indication fraction defining a severity level, note the total errors were found to be ±0.25, ±0.62, and ±0.36 levels for the three radiographs in Table 4.

Table IV Overall mean indication fractions for three radiographs evaluated using new RT standard, errors in F due to reader variability, repeatability and the total error in F based on 95% confidence

Radiograph ID	Overall Mean Indication Fraction, $< \overline{F} >$	Error in F Due to Reader-to-Reader Reproducibility, $U_{F,1}$	Error in F Due to Reader Repeatability, $U_{F,2}$	Total Error, RSS of $U_{F,1}$, $U_{F,2}$ and Ruler Resolution
X-Ray #1	0.1121	0.0227	0.0073	0.0246
X-Ray #2	0.3460	0.0525	0.0319	0.0617
X-Ray #3	0.1049	0.0314	0.0155	0.0355

One-way ANOVA analysis was performed comparing the variability within all readers to the variability between the readers. In this analysis, the null hypothesis is that there is no difference between the measurements of the seven readers other than what one would expect to result from random variations following the normal probability distribution. The results of this analysis will be presented in the typical ANOVA table commonly found in textbooks [5] and as produced by the Excel spreadsheet "Data Analysis" tool. The entries in the table are the sources of variation, the sum of the squares of the indication fraction variations for each source (*SS*), the degrees of freedom for each source (*df*), the mean square of the variation of each source (*MS*), the calculated F-statistic (*F-stat*) and the P-value, and the F-critical value for the significance level chosen and degrees of freedom. The significance chosen for this analysis is $\alpha = 0.05$. If the P-value resulting from the analysis is less than α, than the probability is small relative to α that the differences in variations between the readers and within the readers is random. If this occurs, the null hypothesis (that there is no systematic difference in the readers' measurements) is rejected, and then the differences between readers are significant. An alternate, but equivalent, test is to compare the calculated F-statistic to the F-critical value. If greater than

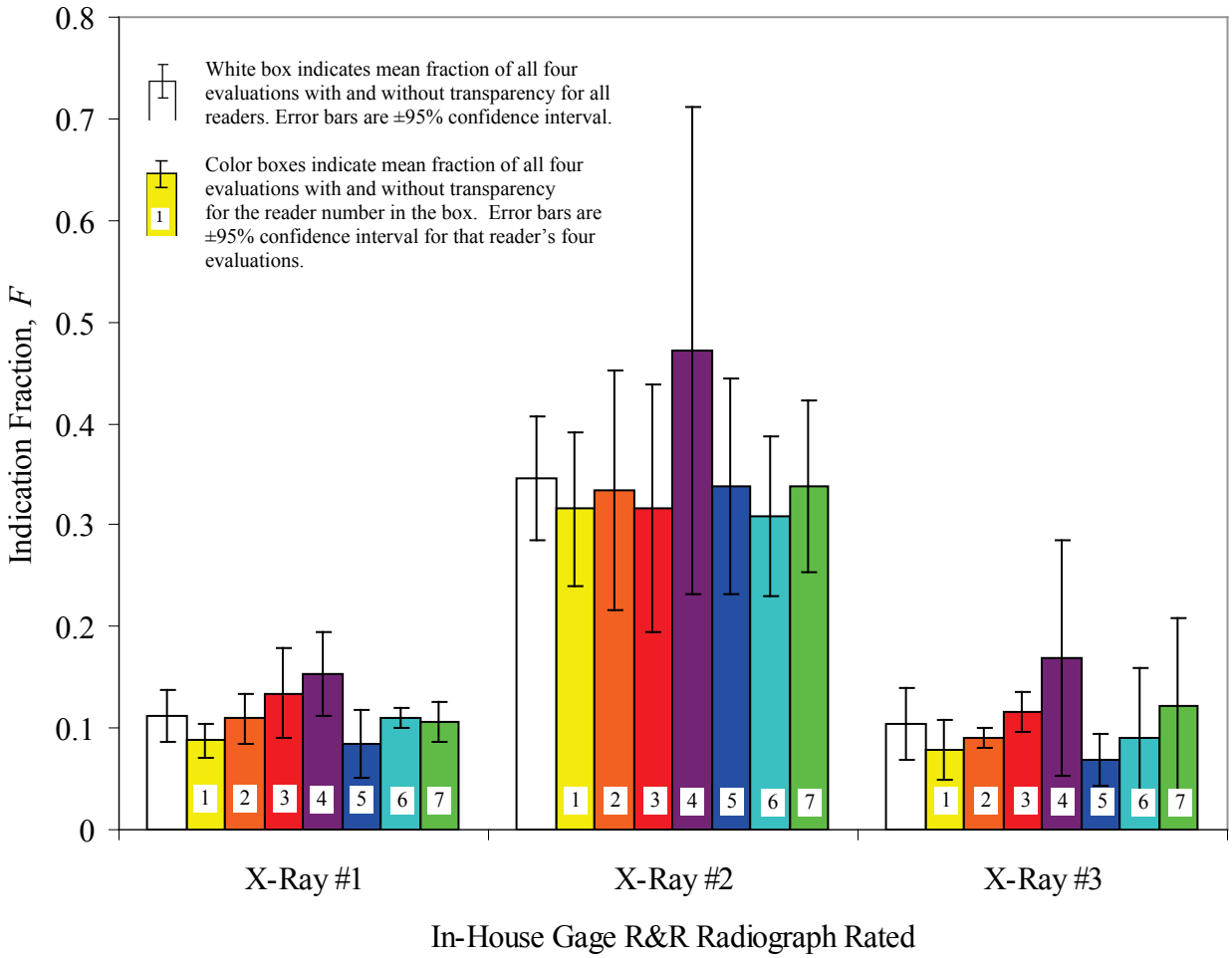


Fig. 24 — Overall mean indication fractions and 95% confidence intervals, and the means and confidence intervals for all readers and all four ratings of each radiograph in the in-house study.

the critical F-statistic for the significance level and degrees of freedom, the differences between the readers is systematic (not random), and also leads us to reject the null hypothesis.

The ANOVA results are given in Tables V, VI and VII for radiographs #1, #2 and #3, respectively, made using the data for all seven readers. In Table V note that the analysis shows the differences between the readers is not random, and there is systematic reader-to-reader variability. For radiograph #2 in Table VI, again using all reader data, it cannot be said that there are systematic differences between the readers. While in Table VII for radiograph #3, the P-value is about 0.03, and just less than our $\alpha = 0.05$ significance level indicating the differences between readers is not random, but not to the degree seen in the radiograph #1 data. Earlier it was mentioned that the errors reported in applying the standard for this portion of the gage R&R could be reduced by throwing out the data for reader #4, since their mean indication fractions were consistently outside the bands of the overall error confidence interval. Doing this in an ANOVA analysis for each radiograph, the results come out as shown in Tables VIII through X. Now the only radiograph that is not meeting the $\alpha = 0.05$ significance level is radiograph #1.

Table V One-way ANOVA table for Radiograph #1 using data from all seven readers

Source of Variation	*SS*	*df*	*MS*	*F-stat*	*P-value*	*F-critical*
Between Readers	0.01441	6	0.00240	**6.90374**	**0.00037**	**2.57271**
Within All Readers	0.00730	21	0.00035			
Total	0.02171	27				

Table VI One-way ANOVA table for Radiograph #2 using data from all seven readers

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F-critical*
Between Readers	0.07724	6	0.01287	**1.95343**	**0.11881**	**2.57271**
Within All Readers	0.13840	21	0.00659			
Total	0.21564	27				

Table VII One-way ANOVA table for Radiograph #3 using data from all seven readers

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F-critical*
Between Readers	0.02765	6	0.00461	**2.98075**	**0.02885**	**2.57271**
Within All Readers	0.03246	21	0.00155			
Total	0.06011	27				

Table VIII One-way ANOVA table for Radiograph #1 using data from all readers, except reader #4

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F-critical*
Between Readers	0.00654	5	0.00131	**4.42941**	**0.00833**	**2.77285**
Within All Readers	0.00531	18	0.00030			
Total	0.01185	23				

Table IX One-way ANOVA table for Radiograph #2 using data from all readers, except reader #4

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F-critical*
Between Readers	0.00328	5	0.00066	**0.16913**	**0.97073**	**2.77285**
Within All Readers	0.06984	18	0.00388			
Total	0.07312	23				

Table X One-way ANOVA table for Radiograph #3 using data from all readers, except reader #4

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F-critical*
Between Readers	0.00863	5	0.00173	**1.86183**	**0.15144**	**2.77285**
Within All Readers	0.01668	18	0.00093			
Total	0.02531	23				

However, its P-value is now about 0.008, which is almost up to the $\alpha = 0.01$ significance level, another often used, but less stringent, test hypothesis level. If the standard were to be followed consistently, and with additional training and experience, these reader-to-reader variations can be reduced to the level of the repeatability variations among all the readers. It is also encouraging that radiograph #2, which had the most indications and required the most complex measurement tasks, showed little reader-to-reader variability in this ANOVA test.

Additional insight was gained into the between reader variability by digitizing and overlaying the transparent cover sheets used by the readers. Only the results from the first reading are presented here, but these results are representative of the other two readings made with the cover sheets. Presented in Figures 25, 26 and 27 are the overlay maps of regions where readers outlined indications for radiographs #1, #2, and #3, respectively. These maps show regions ranging from those where all seven readers were in agreement (black areas) to those where just one reader outlined indications (yellow areas). Also shown in the figures are the locations of the maximum indication lengths and the corresponding reader numbers. Note that in several cases, multiple maximum lengths are given for some readers, because a "tie" occurred in their measurements. For instance, in Figure 25 there are three points along the length of the radiograph (the position of the line in the DOI) where reader #7 found equal maximum indication lengths. For radiographs #1 and #3 there is nearly unanimous agreement (six out of seven readers) on the location of the maximum indication length given by the clusters of red arrows. Examining the data of the two instances where the readers did not agree, the two readers have the location where the other 6 were in agreement in their list of measurements within the resolution of the ruler of being the largest indication measured. For radiograph #2 in Figure 26, note there is more scatter where readers found the maximum indication length. The color map in Figure 26 shows many more regions where only one or two readers detected indications. This reflects the contribution of the radiograph itself to the error in the measurement. Radiographs #1 and #3

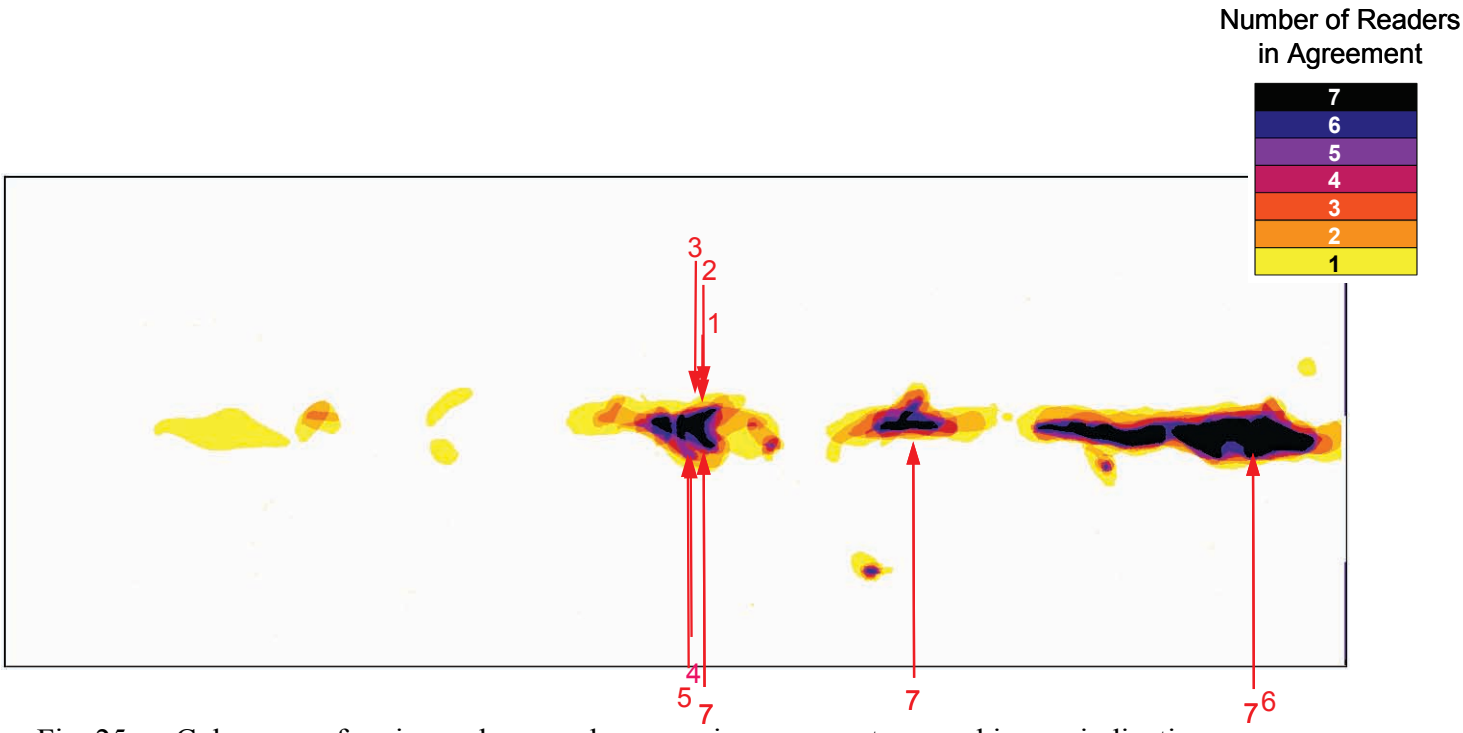


Fig. 25 — Color map of regions where readers were in agreement on marking an indication area on radiograph #1. Locations of maximum indications and corresponding reader numbers are also given.

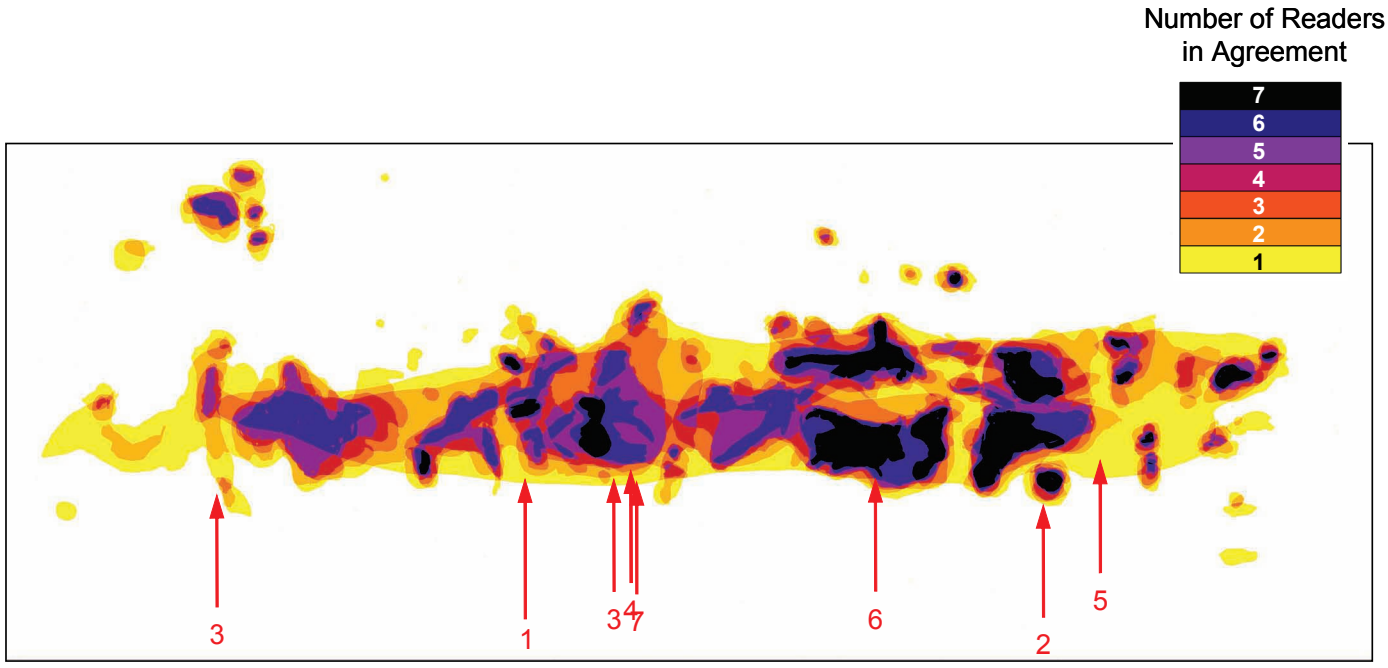


Fig. 26 — Color map of regions where readers were in agreement on marking an indication area on radiograph #2. Locations of maximum indications and corresponding reader numbers are also given.

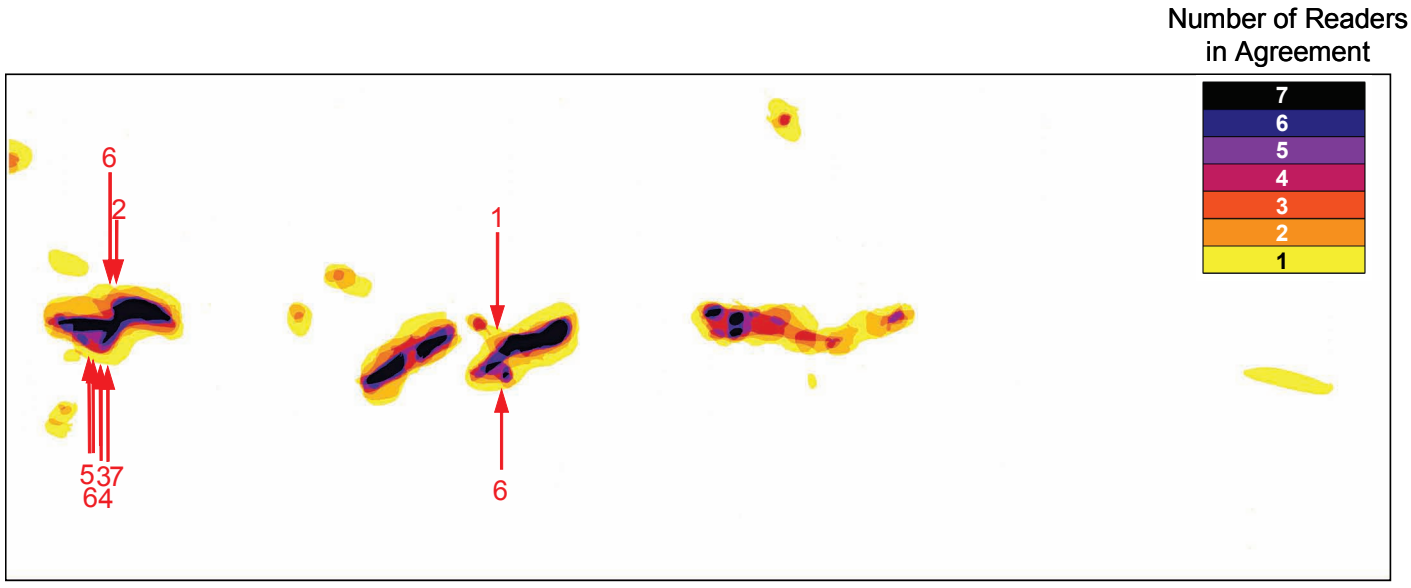


Fig. 27 — Color map of regions where readers were in agreement on marking an indication area on radiograph #3. Locations of maximum indications and corresponding reader numbers are also given.

Fig. 28 — Map of indications outlined on transparency by reader #2 for X-ray #2. Arrow indicates point where maximum indication length in width direction was measured.

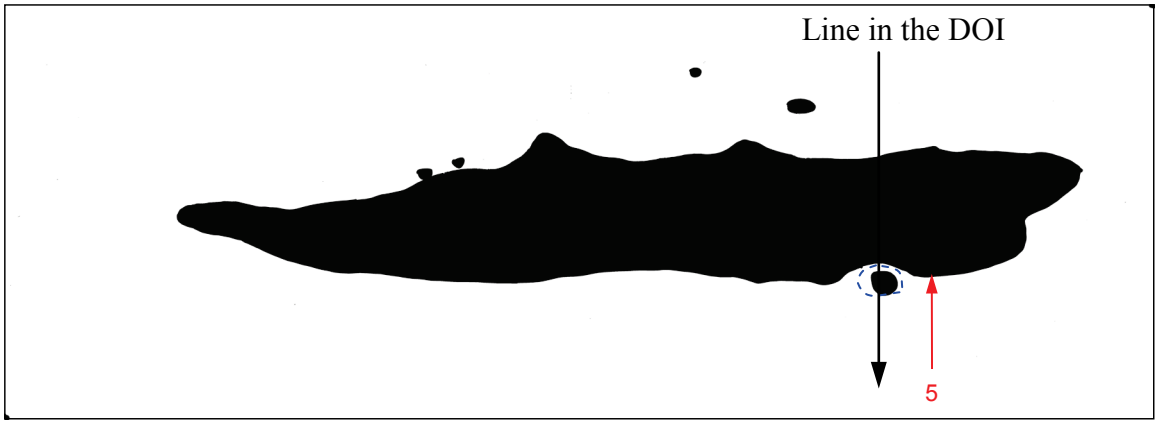


Fig. 29 — Map of indications outlined on transparency by reader #5 for X-ray #2. Arrow indicates point where maximum indication length in width direction was measured.

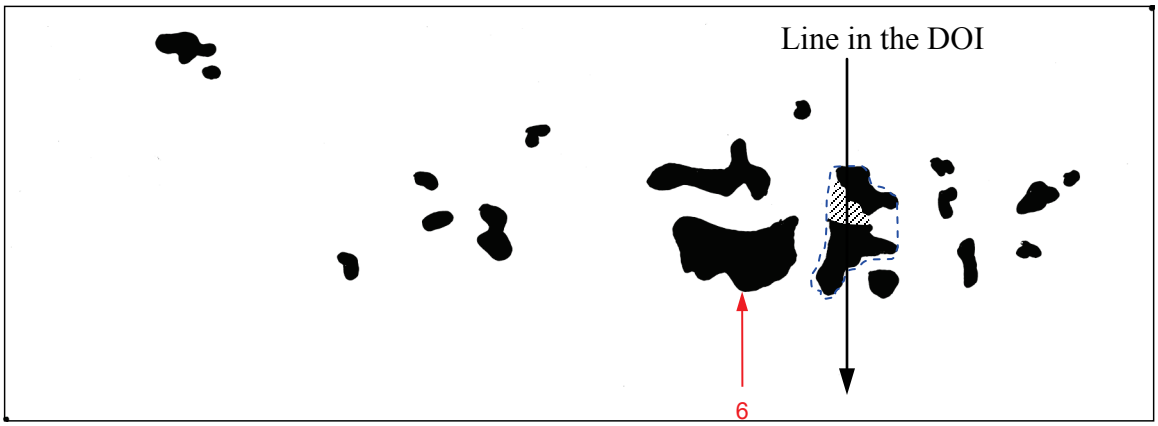


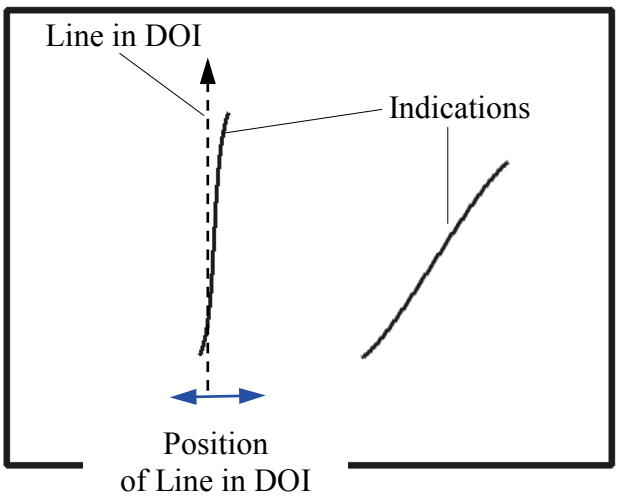
Fig. 30 — Map of indications outlined on transparency by reader #6 for X-ray #2.. Arrow indicates point where maximum indication length in width direction was measured.

have by comparison tighter envelopes where nearly all readers are in agreement and are no doubt easier to rate than radiograph #2 given the time data shown in Tables II and III. Transparency cover sheets of three individual readers' indications for radiograph #2 are given in Figures 28, 29 and 30. Qualitatively, it is interesting to observe the differences in the readers' acuity and interpretation of what they observed. It is also interesting to note whether or not they have followed the standard. In each of these figures indications have been outlined in dashed blue lines where the criteria for combining indications (according to the standard) was not followed. For these indications there are the positions of the line in the DOI where distances between the lengths of indications in the outlined areas are smaller than the lengths of the smaller indications. Where this occurs, these indications should have been combined. For example, the hatched region in Figure 30 was added show a region that should have been combined and included as an indication area.

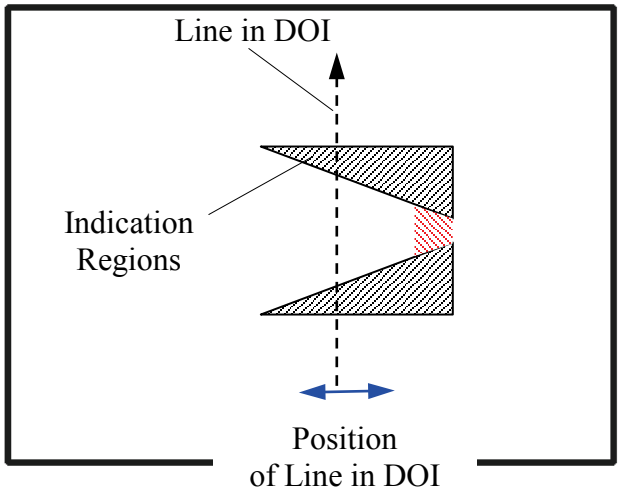
IV. CONCLUSIONS AND RECOMMENDATIONS

In the first stage of the study it was found that 21 of the 30 radiograph ratings had reproducibility errors of less than ±1 levels, using the proposed 0.1 indication fraction range proposed for acceptance levels. Four of the remaining nine ratings had mean ratings larger than the most severe level so they would fall automatically into that rating. The new standard is performing better in evaluating 24 of the 30 radiographic evaluations than the current standard performed in a past study where the error was ±1.4 levels. Still this large level of error was disappointing; it would be desirable to achieve an error lower than ±0.5 levels. In the in-house stage 2 of the study, much lower overall errors due to both repeatability and reproducibility were found; ±0.25, ±0.62, and ±0.36 levels for the three radiographs. The smaller error of the in-house study is very encouraging. It is assumed to be low due to personal instruction given the readers on the standard, procedure to follow, and careful control over how the measurements were made. It is concluded that the new standard is viable.

Nevertheless, it is concluded that the new standard could still be improved in two important areas identified in these studies. As shown in the figure at the right, when an indication is aligned in the direction of interest (DOI), the measurement of the length of the indication is very sensitive to the position of the line in the DOI. The measurement of the indication on the right side of the figure is less sensitive to the line position. Replacing the line with a strip of a standard width will make the measurements less sensitive, easier and more repeatable. Any indications in the strip would be summed to determine the indication length. A strip 0.25 inches wide is recommended.



Next, the method of determining when indication lengths are to be combined in the measurement process must be clarified and simplified. Consider the case of the two triangular indication regions in the figure shown at the right. As the position of the line in the DOI moves to the right, it will eventually be in a position where the distance between the two indication lengths is smaller than the indication lengths. At this point the indication lengths combine to form the region approximated in the figure by the red hatched area. This can be difficult to apply in practice as was shown in the gage R&R study. Additional clarifying text and examples of this process will be added to the standard to remedy this.



ACKNOWLEDGEMENTS

This research was undertaken through the American Metalcasting Consortium (AMC). AMC is sponsored by Defense Supply Center Philadelphia (DSC, Philadelphia, PA) and the Defense Logistics Agency (DLA, Ft. Belvoir, VA). This work was conducted under the auspices of the Steel Founders' Society of America (SFSA) through substantial in-kind support and guidance from SFSA member foundries. In particular, the authors gratefully appreciate the participation of the following foundries in the gage R&R study: Bradken-A. G. Anderson, Bradken-Amite, Bradken-Atchison, Bradken-Atlas, Harrison Steel Castings, Keokuk Steel Castings, Metaltek-Carondelet, Metaltek-Waukesha, Pacific Steel, Sivyer Steel and Spokane Industries. Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors and do not necessarily reflect the views of DSC, DLA, or the SFSA and any of its members.

REFERENCES

1. Blair, M., Monroe, R., Hardin, R.A., and Beckermann, C., "A New Standard for Radiographic Acceptance Criteria for Steel Castings," in *Proceedings of the 62nd Technical and Operating Conference,* SFSA, Chicago, IL, 2008.

2. Carlson, K., Ou, S., Hardin, R., and Beckermann, C., "Analysis of ASTM X-Ray Shrinkage Rating for Steel Castings," *Int. J. Cast Metals Research*, Vol. 14, pp. 169-183, 2001.

3. ANSI/ASME Power Test Codes-PTC 19.1, *Test Uncertainty*, American Society of Mechanical Engineers, New York, 2005.

4. Figliola, R., and Beasley, D., *Theory and Design fir Mechanical Measurements*, 4th Ed., Wiley, 2005.

5. Montgomery, D.C., *Introduction to Statistical Quality Control*, 5th Ed., Wiley, 2005, p. 135.