e-Journal of Nondestructive Testing - ISSN 1435-4934 - www.ndt.net

Stage 1 Round-Robin Report: On the Use of Total Focusing Method for Sizing Flaws in Assessing Probability of Detection

M. Burrowes[1], E. Ginzel [2], G. R. Pereira[3], L. M. Tavares[3]

[1] LNDC / Federal University of Rio de Janeiro, Brazil
e-mail: mariana@metalmat.ufrj.br
[2]University of Waterloo, Waterloo, Ontario, Canada
e-mail: eginzel@mri.on.ca
[3] Federal University of Rio de Janeiro, Brazil
e-mail: tavares@metalmat.ufrj.br
e-mail: gpereira@metalmat.ufrj.br

2023.07.26

Abstract

Reliability of non-destructive (NDT) procedures is usually assessed using probability of detection (POD). However, the reliability of the POD is rarely questioned. A POD usually produces a curve that relates the size of a defect to the likelihood that it will be detected by the NDT procedure. PODs are often used in conjunction with fracture mechanics that can relate the severity of a flaw to the service life of a component (Fitness for Service or FFS). To this end, it is the flaw vertical extent relative to the component thickness that is important. Closely related to the concern for flaw height is how close the flaw is relative to the surface of the component. Proximity to a surface can increase the stresses raised by the flaw and make it more critical to the fitness for service. Use of full matrix capture (FMC) techniques in ultrasonic testing, in conjunction with the post-processing use of Total Focussing Method (TFM), has shown good potential to provide accurate flaw sizing. This paper examines the results of a round-robin trial using CIVA simulated data that were configured to assess flaw height and ligament in simulated welds. Results suggest that FMC with TFM post-processing can provide sizing estimates as good or better than manufacture estimates and can be used as the actual measured values for POD assessments.

Keywords: NDT, POD, TFM, sizing, reliability, flaws

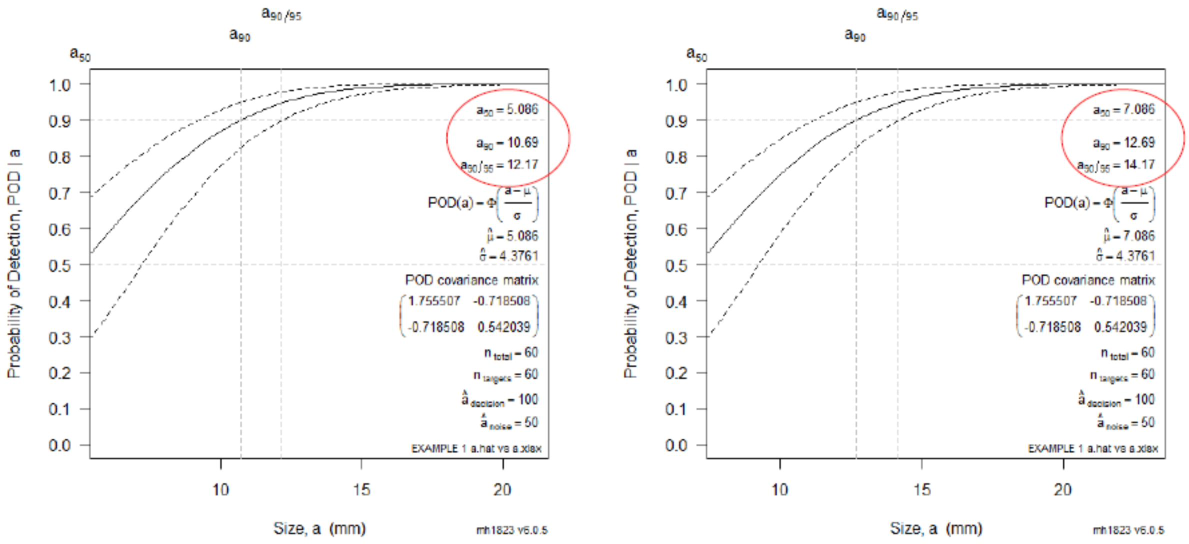
https://doi.org/10.58286/28550

1. Background

In a recent paper [1] the authors described how several standards [2-7] dealt with determining the true flaw sizes when developing a POD. With the exception of DNVGL ST-F101 [5], it is normally accepted that the dimensions provided by the manufacturer of flawed specimens are the true flaw sizes. Yet even the manufacturer provides a tolerance on the stated dimensions. The standards provide no guidance on how to deal with manufacturer tolerances. Perhaps it is assumed that the tolerance from the true values will be incorporated in the "confidence bounds" generated in the POD curves. However, confidence bounds are generally considered to be a result of the variation in the test procedure, not the test specimen.

PODs used in NDT provide a value of a flaw size that can be detected with a high probability and at a high level of confidence. The preferred POD value is 90% and the preferred confidence bound is 95%. Use of the 95% confidence bound suggests that a significant degree of conservatism is merited. It would seem to follow that if POD is to provide a conservative approach, then the tolerance of the flaw size indicated by the manufacturer should also be conservatively used. This could have significant effects on the POD curves.

Figure 1 compares a POD curve that was generated using the nominal flaw size to the same data where all the flaws have a 2mm tolerance added to them. In the plot on the left, where the nominal value is used, a 90|95 value of 12.17mm is identified. In the plot on the right, where 2mm tolerance for possible sizing error is added to the flaws, the 90|95 increases to 14.17mm.



a vs â

a vs â (defects + 2 mm)

Figure 1 - The effect of sizing tolerance on POD

The obvious effect is that the POD curves are shifted to the right by the value of the tolerance. When the tolerance is large, this could make a significant difference to the apparent reliability of the NDT procedure.

It follows that, for PODs to be reliable, the values used for the control variable (flaw size) must be accurate. DNVGL-ST-F101 accepts only destructive test results using macro-photos from cut and polished sections of test specimens as the source of the true size of flaws used in POD development. Even then, there can be some degree of uncertainty. For example, when the location of the flaw is marked on the specimen to be sectioned, it is usually made at the point where the NDT method has obtained the maximum response. This assumes that the maximum response location corresponds to the maximum flaw size. This is not always the case and in some POD programmes, salami cuts in 2mm increments from the marked location can sometimes reveal that the flaw maximum dimension can be larger than 2mm to 4mm away from the maximum response location[1].

In spite of the exceptions, destructive testing probably has the greatest degree of accuracy when determining flaw sizes for POD development. However, destructive testing is not always desirable and it is usually costly. In addition, destructive testing renders the specimen useless for later studies.

One of the recommendations of the referenced paper [1] was to carry out future work to assess the potential accuracy of TFM sizing using a round-robin test on simulated data with a variety of flaw types and geometries. To this end, a collection of simulated FMC data was generated using CIVA simulation software. With the assistance of EXTENDE (the distributor of CIVA) it was possible to remove the flaw images from the files and participants in the round-robin were provided with scan data, similar to what they would see had the data been collected through a physical hardware setup. This paper reports the results of the analyses of the FMC data generated using CIVA simulation software.

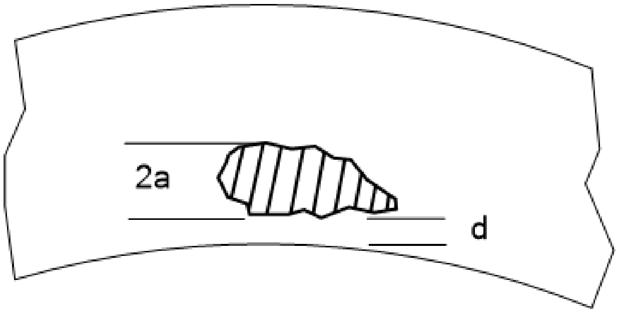
2. The Programme

The round-robin programme was deployed on the importance that flaw size has when developing a POD for an NDT procedure qualification. The underlying assumption when one plots probability versus "*a*" is that "*a*" (the flaw size) is an absolute value that is known. However, in reality, most programmes are based on samples obtained from flaw manufacturers that state tolerances in accuracy on the sizes of flaws they manufacture (e.g., samples are often provided with +/-2 mm tolerance). When the flaws are made naturally, by varying the welding parameters, one has absolutely no idea what size the flaw is, so these usually require destructive testing to estimate the true size. The idea for this work is to demonstrate that, when a true size is required by a qualification process, it could be possible to use TFM algorithms to obtain an accurate flaw size that could serve as basis for the POD. This could avoid destructive testing and loss of the samples for future studies.

Participants in this round-robin were requested to provide values for only two parameters; flaw vertical extent (height) and ligament to the nearest free boundary. Height and ligament values that were to be determined by participants would be typical of the information required for fracture mechanics calculations when "Fitness-For-Service" acceptance criteria are being used. Codes use similar terminology to indicate the flaw vertical extent relative to the thickness. "a" is what they call a surface flaw height and "2a" is used for subsurface flaw height (see Figure 2).

[1] Personal correspondence from Dave Stewart (offshore AUT qualification supervisor) have shown that over 1mm variation may be seen where the macro images demonstrate larger flaw heights 2mm to 4mm away from the marked maximum response.

The ligament (d in Figure 2) is also critical. The shortest ligament to a free boundary has the potential to cause the flaw to interact with the surface due to stress concentration. If the dimension “d” is smaller than some critical value as determined in the calculations, interaction is deemed to consider the flaw as a surface flaw (more critical).



If d<a then the flaw is considered interacting with the surface and the total height would be 2a+d

Figure 2 - Typical Flaw size and ligament for surface interaction

In a pipe, with internal pressure, circumferential stresses form tangential to the radius. They are usually termed “Hoop Stresses” and are a function of the diameter and thickness. It can be demonstrated that the hoop stresses are higher at the outside surface than at the inside surface, that is

Hoop Stress $\sigma_H = \frac{PD_i}{2t} < \frac{PD_o}{2t}$

Where σ_H is hoop stress
P is pressure
D_i is inside diameter of pipe
D_o is outside diameter of pipe
t is wall thickness

From this equation we see that not all ligaments are equal even when they are equal. Figure 3 illustrates a flaw of the same size with a ligament of the same size. The flaw closer to the outside surface would be deemed more critical based on hoop stress considerations.

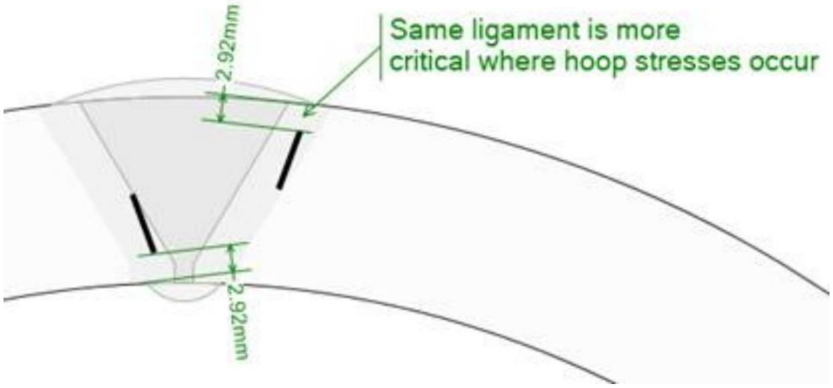


Figure 3 - Significance of flaw size and ligament for surface interaction

The round-robin programme was intended to simulate preparation for a procedure qualification using the principles of probability of detection. In a real qualification programme, it would be normal for the organisers of the programme to have knowledge of the flaws being used. For example, the organisation running the qualification might purchase a series of samples from a flaw manufacturer. They could specify the flaw type and approximate locations and sizes; however, the exact sizes of the flaws would have the manufacturer's tolerance associated with them. A good collection of flawed samples would be useful for evaluating a variety of NDT technologies and the organisers would want to preserve them to evaluate the reliability of other candidate companies or methodologies. To this end, it would be highly desirable to obtain accurate details of the flaws in the flawed samples via non-destructive test methods.

The British Standard BS-7910-2019 has added Annex T. It gives guidance on the use of NDT when deriving an Engineering Critical Assessment (ECA). Tables T1, T2 and T3 summarise capabilities of: Conventional UT, Focussed phased array, Zonal AUT, TOFD and RT. Of these options only the ultrasonic methods have any flaw height and ligament assessment capabilities. The estimates for sizing accuracy are based on full penetration fine-grained isotropic welds (typical of ferritic carbon steel) in plate or pipe joints of wall thickness 10 mm to 25 mm.

These values are summarised here in Table 1.

Table 1 - Typical Ultrasonic Methods' Sizing Capabilities

Capability	Conventional UT	Focussed phased-array	Zonal AUT	TOFD
Through-thickness sizing accuracy	4 mm undersizing, 1 mm oversizing	±1.5	±1.5	±1.5
Ligament sizing accuracy	±3	±1.5	±1.5	±1

Allowing for uncertainties based on where the macro-section is made for destructive testing (as mentioned earlier), it would be reasonable to estimate that destructive testing can provide statistical sizing accuracy on the order of ±0.2mm.

Therefore, if TFM is to be considered a replacement for sizing by destructive testing it should have a statistical sizing capability that is better than the traditional ultrasonic methods and approaching the capabilities of destructive testing.

In order to assess the sizing capability of TFM, the participants in the round-robin were provided 80 files generated by CIVA FMC data acquisition simulation. The 80 files consisted of a single position acquisition on each side of 40 flaws, that were designed using CIVA drawing tools. These flaws were made to be realistic representations of flaws seen in macro photographs.

The data provided had the flaw images removed so only the A-scan data was available to the participants. A description of each flaw was provided in an Excel table (see Appendix 1) similar to what a flaw manufacturer would provide as documentation with fabricated samples. Since the location of the flaw and its nature would be known, the only other aid to the participants was a box outlining the Region of Interest (ROI).

3. Models

FMC and TFM have been used in a wide variety of applications including corrosion assessment; detection and characterisation of high temperature hydrogen attack (HTHA), hydrogen induced cracking (HIC) [8] and sizing of surface breaking fatigue cracks [9].

The study carried out in this round-robin was limited to butt-fusion welds in carbon steel in both plate and pipe form. Thicknesses ranged from 10mm to 32mm with a variety of weld bevel shapes. 40 flaws were simulated and data acquisition used FMC from each side of the weld with the probe positioned to be aligned with the maximum vertical extent of the flaw. This generated 80 files that were identified as being from Side A and Side B for each flaw.

Each flaw simulated was positioned in a weld configuration (CIVA specimen parameters allow the user to specify details of weld geometry that are then incorporated into the 3D image). A flaw was then simulated using the various tools in CIVA, including the ability to import 3D CAD shapes, and the flaws were positioned in the weld in locations appropriate to the nature of the flaw.

After the FMC acquisition had been run, the data was processed to remove the flaw from the image. When the participant opened the file, they would see the test layout without any evidence of the flaw as shown in Figure 4.

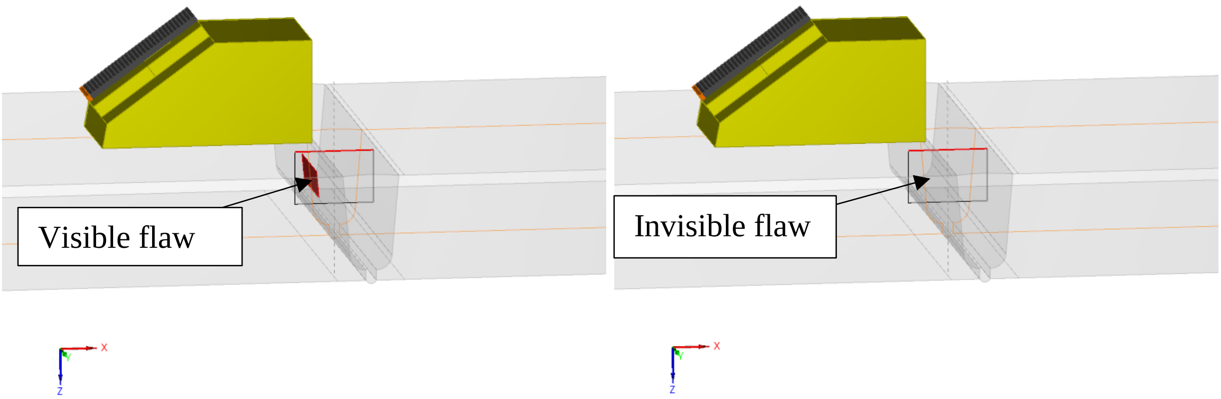
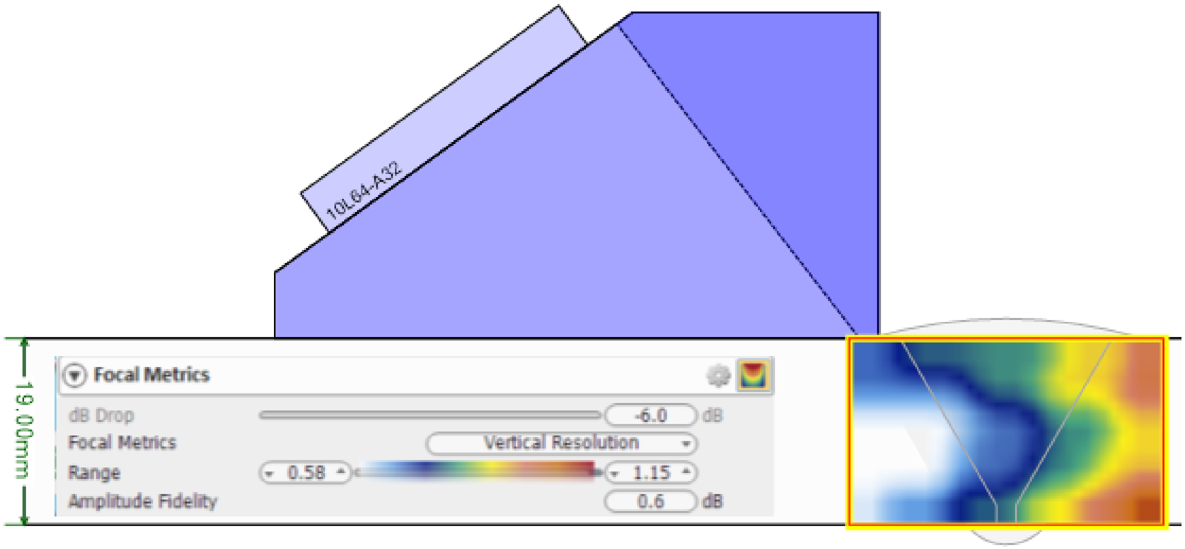


Figure 4 - CIVA display in preparation for FMC acquisition (left) and CIVA display after flaw is removed (right)

Selection of the inspection parameters was based on a tool in ESBeamtool software that provides a map of Focal Metrics in the ROI. Options exist to display;

- Point Spread Function (PSF)
- Focal area
- Sensitivity
- Horizontal resolution
- Vertical resolution

For the purposes of flaw height sizing, the map showing the vertical resolution provided the most useful information. High frequency probes with larger apertures tend to provide better vertical resolution results. Figure 5 illustrates the use of a 64-element linear array 10MHz probe on a refracting wedge placed close to the weld cap on a butt weld in a 19mm thick plate. Using a T-T-T-T mode, the resolution is indicated to be better than about 0.75mm for the entire weld volume.



Focal Metrics

dB Drop	-6.0 dB
Focal Metrics	Vertical Resolution
Range	0.58 – 1.15
Amplitude Fidelity	0.6 dB

Figure 5 - Selecting probe parameters for vertical resolution in FMC acquisition

From a practical perspective, 10MHz probes could experience excessive noise and attenuation in thicker samples so for simulations in welds greater than 25mm a 7.5MHz linear array was selected with 60 elements and 1mm pitch.

After the FMC data acquisition was completed, the A-scan data was preserved in a new CIVA file in which the flaw details had been removed. This left all details of acquisition including the probe parameters and modes used, available to the participants. Because all of the welds simulated were butt joints, the setups on each side of the weld were identical except for skew.

Once all 40 flaws had been scanned from both sides and the flaw details removed from the data acquisition, a call was put out to the NDT community looking for volunteers to carry out sizing analysis. In December of 2022, a request for participants was announced on NDT.net. Shortly thereafter, volunteers from 8 countries and 3 continents requested the files. Background information on the data was supplied to the participants via emails and an Excel spread sheet was provided as a recording table (see Appendix 1 Round-Robin Flaw List).

4. Sizing Analysis

At the time of this round-robin, there were no guidelines for how flaw sizing should be carried out when using TFM algorithms. This meant that each participant was left to develop their own sizing procedure. The FMC acquisition files were made using CIVA 2021. Although several filters are available in the CIVA 2021 for TFM processing (e.g., Mean Subtraction and Scattering Cone), only the classic TFM algorithm is available. During the preparation of the round-robin sample files, CIVA 2023 was introduced. For those participants that could access CIVA 2023, new algorithms were available. These included CF (coherence factor), SCF (sign coherence factor), PCF (phase coherence factor) and PCI (phase coherence imaging).

Since some of these advanced features were not available to all participants, it was expected that there could be a wider scatter of accuracies. A consideration when doing analysis with TFM is the importance of gathering sufficient information. This can be achieved by processing multiple modes from multiple directions. Holloway [11] demonstrated how details of facets in TFM reconstruction of flaw images is improved when multiple modes are used. Simply using the maximum responses from a TFM image could result in a completely incorrect assessment of the flaw. For example, a single mode approach with a flaw having a combination of horizontal and vertical components on each side of a weld bevel would produce a series of "spots" on the TFM image. Merging left and right data and summing four different modes can increase the detail used to identify the shape and extent of the flaws. However, even then, there are conditions that could limit the effectiveness of TFM to get an accurately detailed outline of the flaws. Figure 6 shows a simulation of the typical "stop-start" flaw associated with the automatic welding process (GMAW). The flaw simulates the welding operator stopping the welding machine at the same spot (usually the top of the circumferential weld) after each pass. When the welding process re-starts for the next pass, without having cleaned the flux from the previous pass, a region of lack of fusion can occur. Because it is at the top of the weld, this location can also have a horizontal component. Using a single-sided FMC acquisition results in poor or no backscattered signals and the horizontal component is poorly imaged. Figure 6 illustrates the advantage of multiple modes and assessing the weld from both sides, but also shows the limitation of the single-sided FMC to detect laminar oriented flaws.

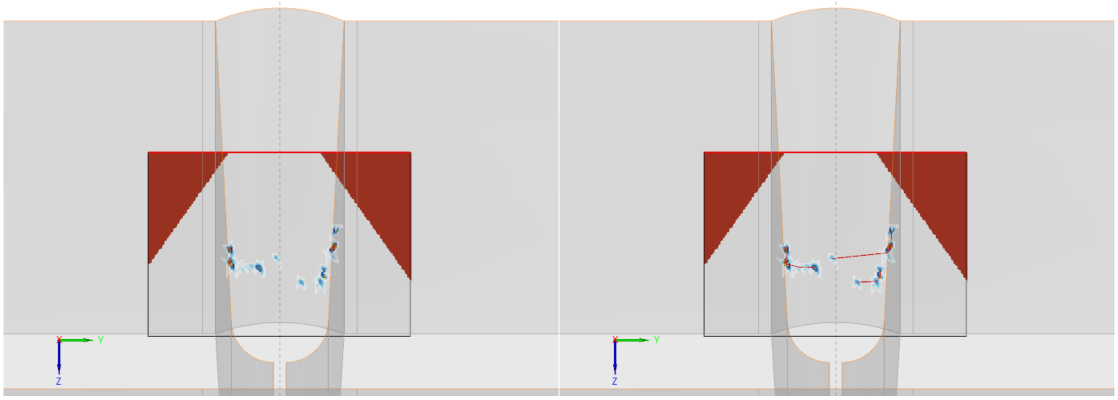


Figure 6 - Merged multiple modes to enhance a stacked welding defect (3-layer stop-start) without flaws seen on left and with flaws seen on right

In the case of a flaw like that is seen in Figure 6, the clue to the nature of the flaw would be in the description of the flaw from the manufacturer. Also, if the flaw had been scanned along the length of the weld and the data displayed as a side-view in addition to the end view, as in Figure 7, the true nature of the flaw as a stop-start would be easier to discern.

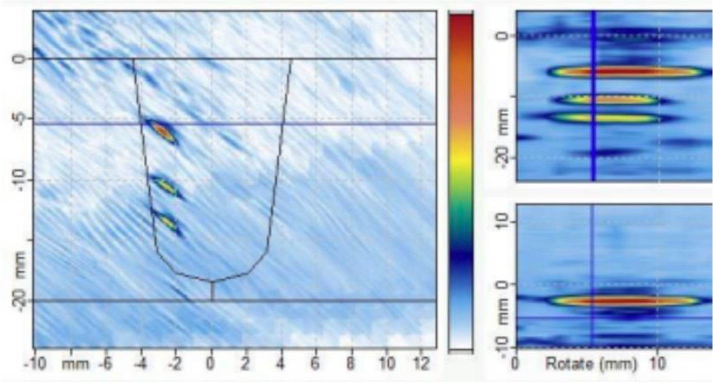


Figure 7 - End and side view of a real stop-start defect (image courtesy Eclipse Scientific)

With no other guidance, participants progressed through the analysis of the scans' FMC data from both sides of the weld and derived the flaw height and ligament as seen in Figure 8 where the participant identified the dimensions for each condition prior to recording it in the reporting table.

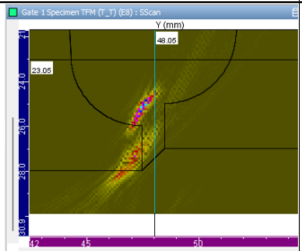
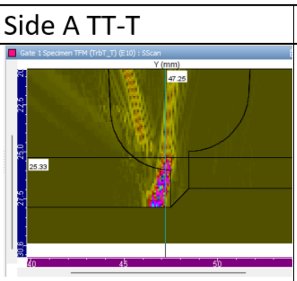
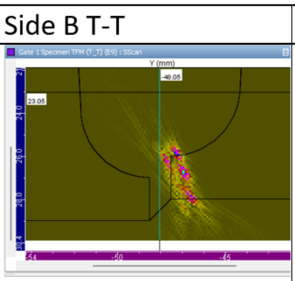
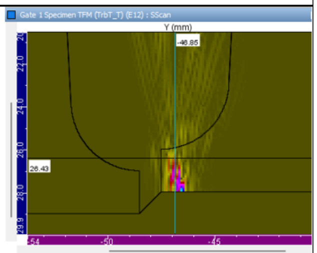
Side A T-T	Side A TT-T	Side B T-T	Side B TT-T
			
Height_V: 2.58 mm Ligament: far 0.00	Height_V: 2.67 mm Ligament: far 0.00	Height_V: 2.08 mm Ligament: far 0.00	Height_V: 1.57 mm Ligament: far 0.00

Figure 8 - Analysis of a simulated incomplete penetration associated with a mismatch

5. Results

Completed reporting sheets were returned to the authors and the results for the flaw size and ligament were compared to the "true" values. The true values of the flaw could be easily determined from the CIVA model parameters. Figure 9 illustrates how, by using the orthographic projection views, the vertical extent and ligaments are measured directly from the Model in the 3D views.

To avoid the complications that could result if the flaw was to have a height and ligament that varies along the weld length, flaws were made to have a relatively uniform dimension in the region of interest. The probe was positioned at the location of the maximum vertical extent of the flaw. In a fracture mechanics application, a flaw as illustrated in Figure 10 would have its height and length defined by the box that bounds it. For the purposes of this project, flaws were made so that a local maximum height and local minimum ligament could be had at a single position.

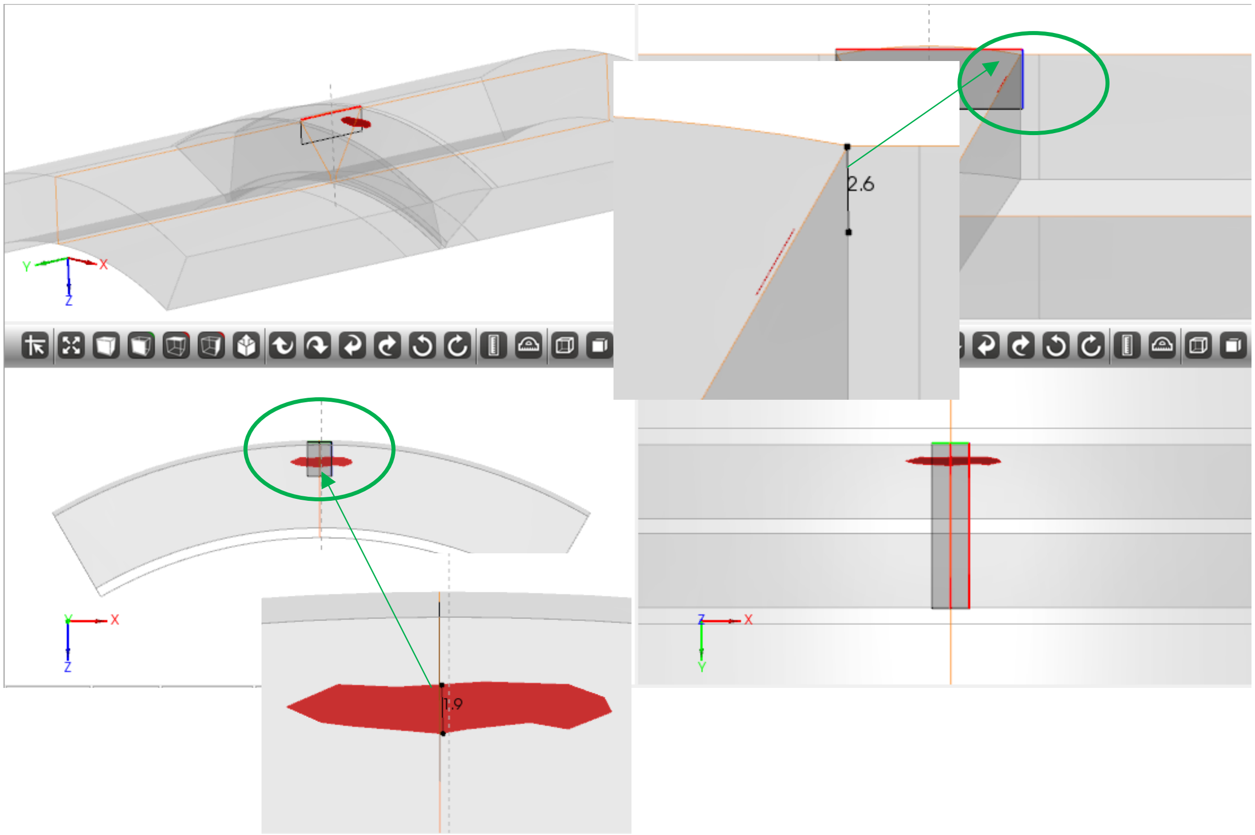


Figure 9 - Determining the true height and ligament using CIVA measurement tools

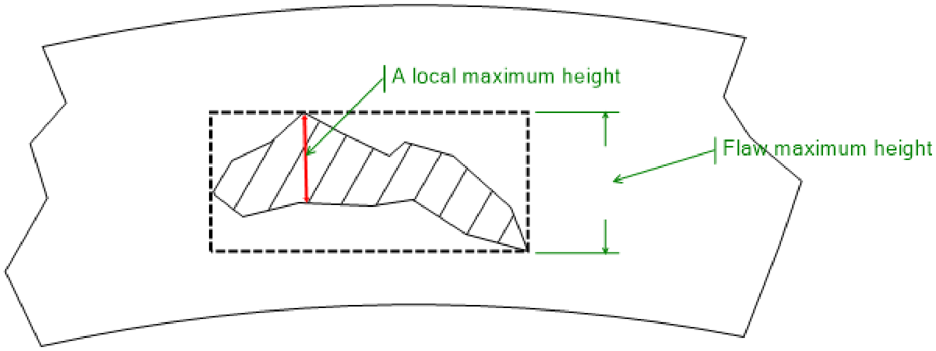


Figure 10 - Flaw height using bounding box would be different from local maximum

It should be noted that all flaws were designed with a length along the weld axis. Initially, it was thought that the FMC data acquisition would be made in 1mm increments along the axis of the weld so that participants would also have to estimate flaw length. However, when the length of time required for a single position (some acquisitions required several days) it was decided to limit the analysis to just a single slice at the flaw maximum vertical extent. The table in Appendix 1 provided participants the length but this was just for information.

Statistical analysis of the returned sizing involved determining the difference between the true flaw height and ligament and the participant analysed height and ligament. This would define the sizing error.

Each participant was provided a summary of their results using the mean error and standard deviation of error. It is important to mention that all analysis presented from here on will refer to the error value and in millimetres. As mentioned, error is here characterized as the difference between what each participant sized regarding height and ligament and the true value, meaning that the closer the measurement error gets to zero the more accurate it is.

An overall summary of these values was then derived using all of the sizing estimates submitted and is shown in Table 2. In order to simplify further analysis, the 10 participants will be called P_i and *i* varies from 1 to 10. The identity of each participant will remain undisclosed, with each one designated a fixed P_i.

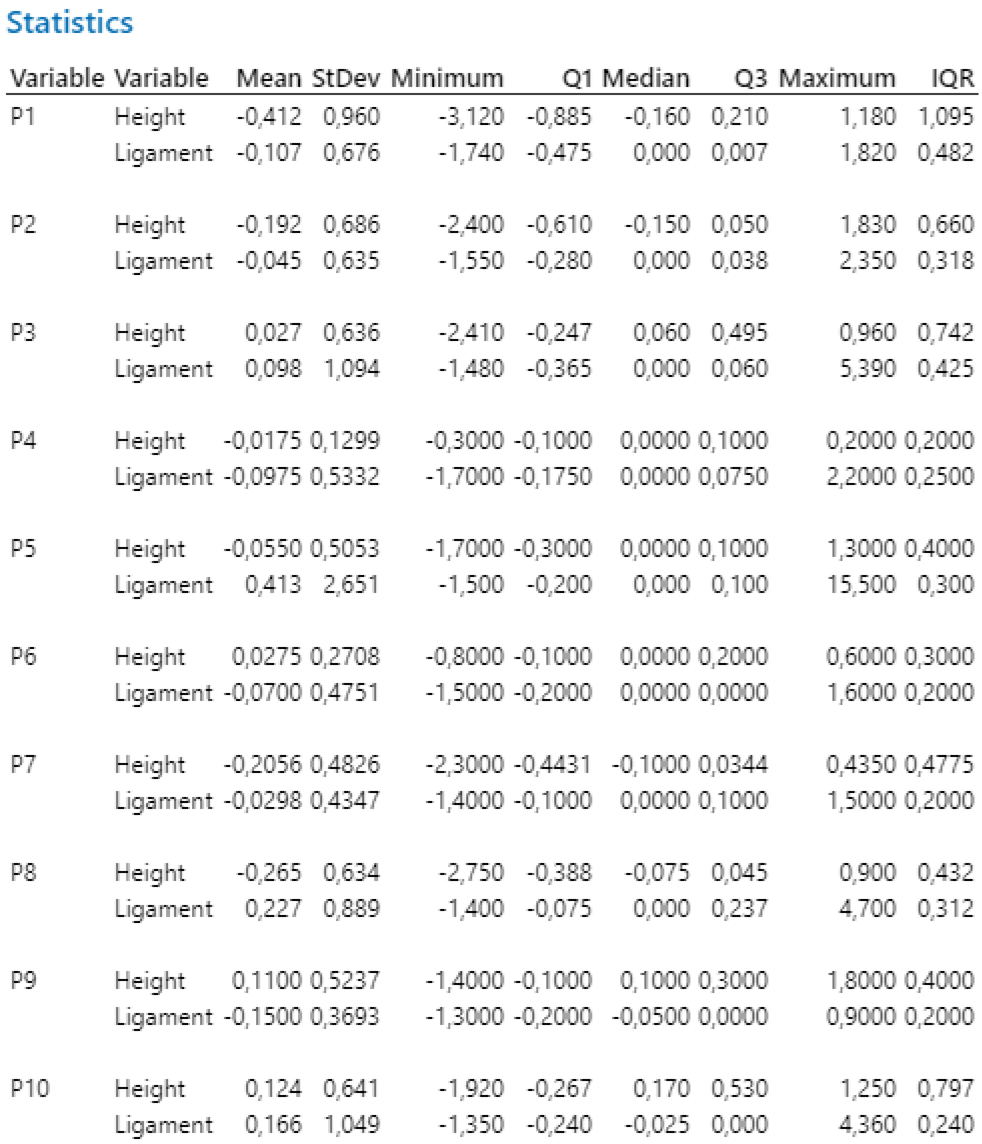
Table 2 brings mean and standard deviation columns for each participant regarding each variable assessed – height and ligament. Every mean value that is positive means that the participant tended to undersized the flaws while when the mean error is negative, means that the participant mostly oversized the flaw. But the standard deviation (StDev) in this case offers an additional information that StDev values close to zero mean that the participant was more accurate. It is worthy to remark that the lowest StDev value regarding height measurements was 0.13 mm while the higher value was only 0.96 mm.

Accuracy does not necessarily mean that the data is also precise. Accuracy tells about the distance of a certain value from a standard or true value. Precision tells how spread those data are. Table 2 brings an interesting factor called IQR, which is Interquartile Range. This value represents the difference between the first quartile and the third one. First quartile can be understood as the middle value between the minimum value of the data set (including outliers) and the median; 25% of all data is below the first quartile, while the third quartile middle value between the median (second quartile) and the highest value (maximum including outliers); 75% of the data is below the third quartile. Hence, the interquartile range represents 50% of the data set values.

Participants that have a close to zero IQR showed to be more precise as the spread of their measurements were narrow. At the same time, participants that showed a close to zero median, were more accurate as the true value is zero. These conclusions are illustrated in Figure 11 regarding both variables.

The boxplot graph in Figure 11 shows the distribution of media values and IQR for all participants regarding height and ligament. There is a particularity that is an extreme outlier value of ligament measurement performed by participant 5 (P5). During all analyses, there was a concern to not eliminate any outlier; except for this one for being clearly not a tendency between participants. In that way, Figure 12 shows the exact same data as Figure 11 but without the value of 15.5 mm of ligament error for P5 for better visualization. The star marks are the outlier's values.

Table 2 - Basic descriptive statistics for all participants regarding height and ligament errors



Statistics

Variable	Variable	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
P1	Height	-0,412	0,960	-3,120	-0,885	-0,160	0,210	1,180	1,095
	Ligament	-0,107	0,676	-1,740	-0,475	0,000	0,007	1,820	0,482
P2	Height	-0,192	0,686	-2,400	-0,610	-0,150	0,050	1,830	0,660
	Ligament	-0,045	0,635	-1,550	-0,280	0,000	0,038	2,350	0,318
P3	Height	0,027	0,636	-2,410	-0,247	0,060	0,495	0,960	0,742
	Ligament	0,098	1,094	-1,480	-0,365	0,000	0,060	5,390	0,425
P4	Height	-0,0175	0,1299	-0,3000	-0,1000	0,0000	0,1000	0,2000	0,2000
	Ligament	-0,0975	0,5332	-1,7000	-0,1750	0,0000	0,0750	2,2000	0,2500
P5	Height	-0,0550	0,5053	-1,7000	-0,3000	0,0000	0,1000	1,3000	0,4000
	Ligament	0,413	2,651	-1,500	-0,200	0,000	0,100	15,500	0,300
P6	Height	0,0275	0,2708	-0,8000	-0,1000	0,0000	0,2000	0,6000	0,3000
	Ligament	-0,0700	0,4751	-1,5000	-0,2000	0,0000	0,0000	1,6000	0,2000
P7	Height	-0,2056	0,4826	-2,3000	-0,4431	-0,1000	0,0344	0,4350	0,4775
	Ligament	-0,0298	0,4347	-1,4000	-0,1000	0,0000	0,1000	1,5000	0,2000
P8	Height	-0,265	0,634	-2,750	-0,388	-0,075	0,045	0,900	0,432
	Ligament	0,227	0,889	-1,400	-0,075	0,000	0,237	4,700	0,312
P9	Height	0,1100	0,5237	-1,4000	-0,1000	0,1000	0,3000	1,8000	0,4000
	Ligament	-0,1500	0,3693	-1,3000	-0,2000	-0,0500	0,0000	0,9000	0,2000
P10	Height	0,124	0,641	-1,920	-0,267	0,170	0,530	1,250	0,797
	Ligament	0,166	1,049	-1,350	-0,240	-0,025	0,000	4,360	0,240

It can be seen that even thought there were outliers, 50% of the results for every participant for both variables (length of each box) were mostly between the standard deviation value of all measurements, which is ± 0.56 mm for height measurements.

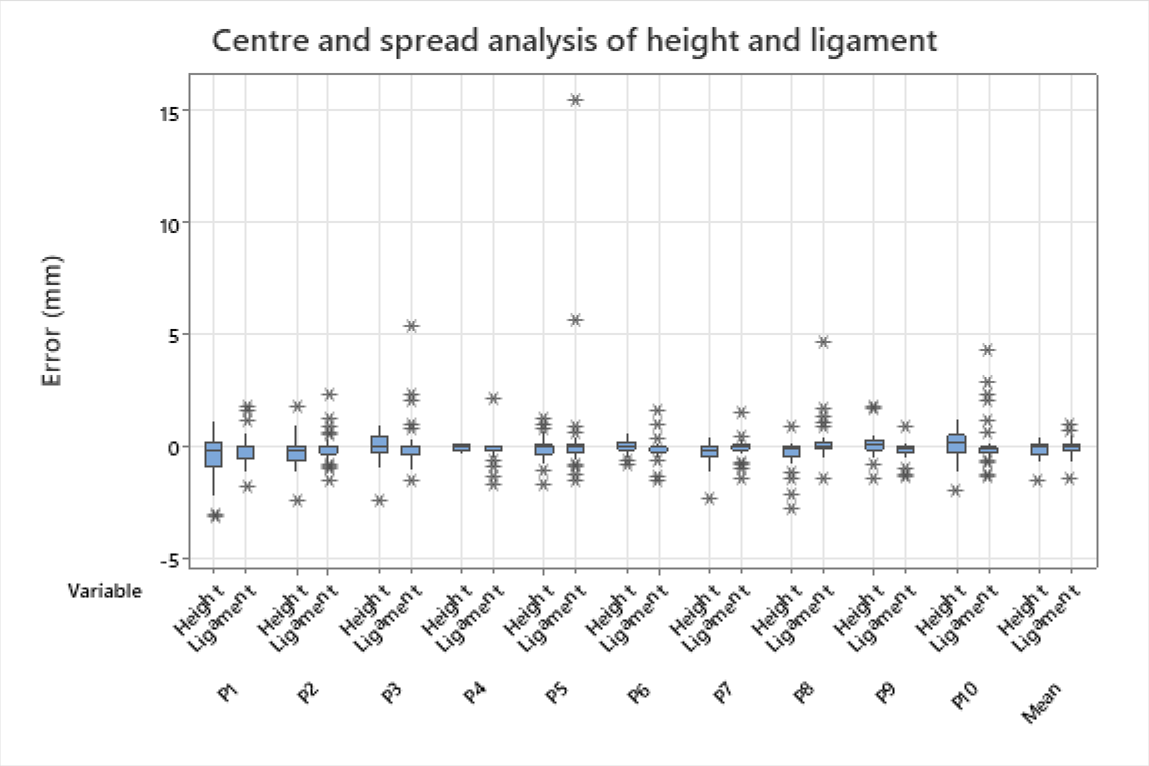


Figure 11 - Boxplot graph for all participants regarding both variables: height and ligament

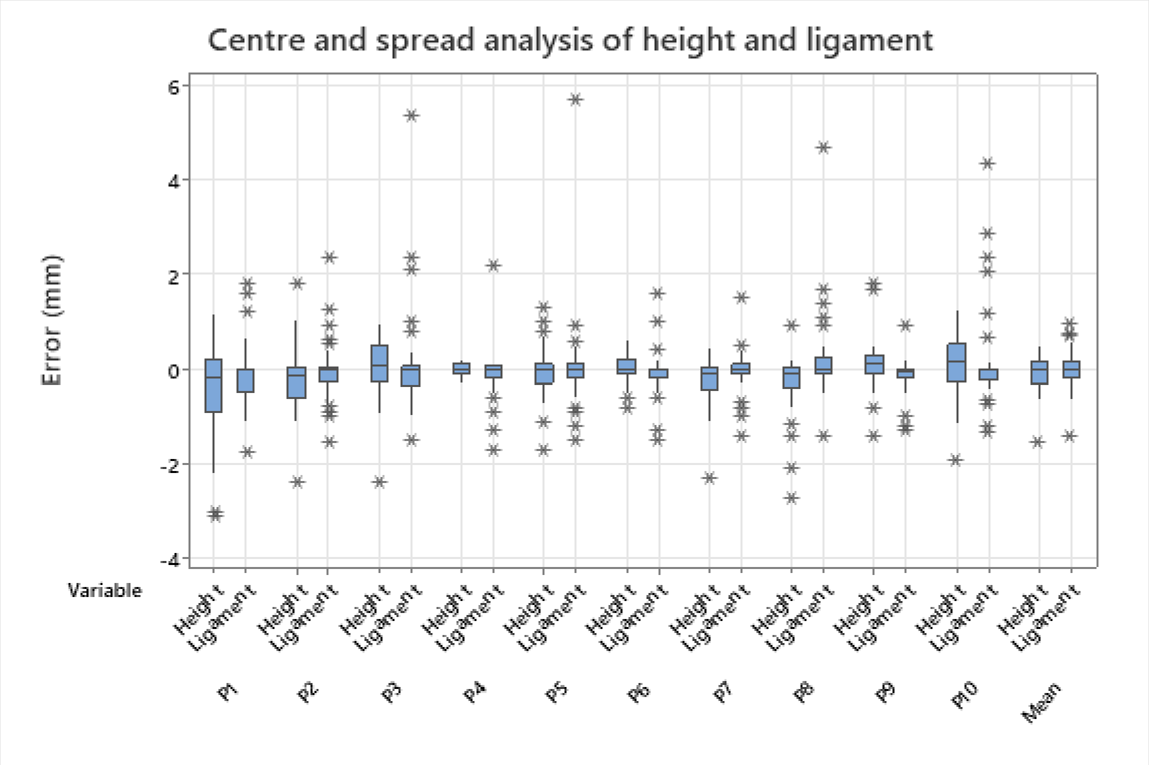


Figure 12 - Boxplot graph for all participants regarding both variables: height and ligament except outlier of 15.5 mm of ligament for participant 5

The value of standard deviation used to represent a precision indicator was calculated through the mean of the standard deviations of each participant. Figure 13 illustrates the boxplots regarding the standard deviation limits and Figure 14 shows the boxplot regarding 2 mm limit.

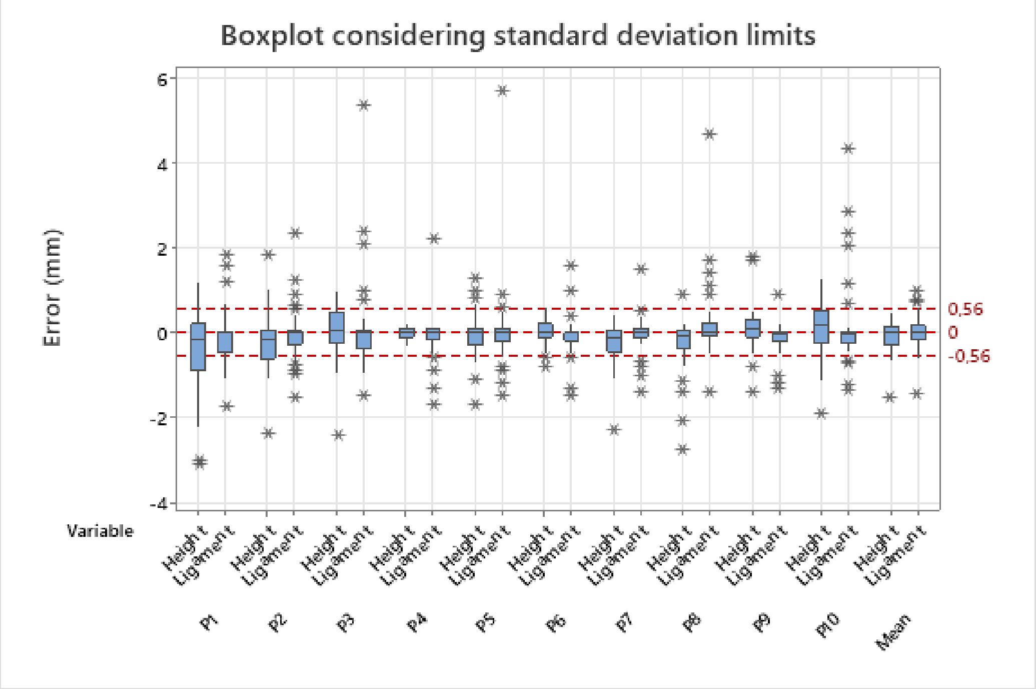


Figure 13 - Boxplot considering the standard deviation limit of 0.56 mm.

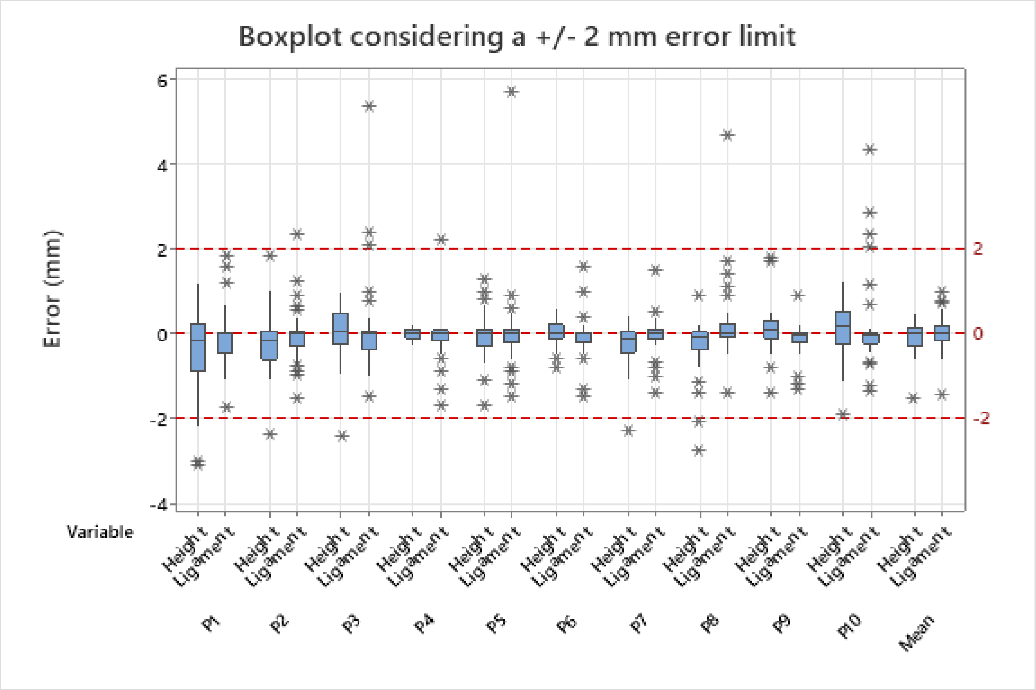


Figure 14 - Boxplot considering a 2 mm error limit

Looking closer at each variable, Figures 15 and 16 show the same boxplots but separated by variables and with mean connected.

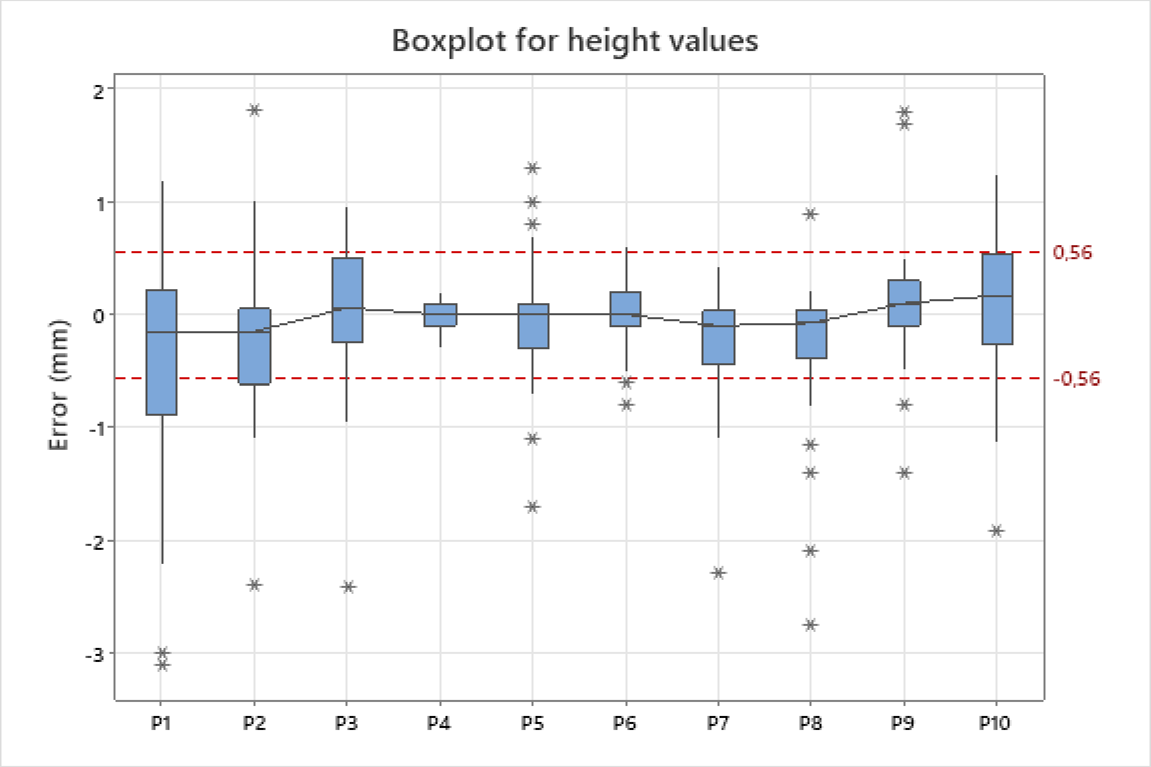


Figure 15 - Centre and spread analysis of height error

Figure 15 shows some interesting details such as every median being around zero (error true value). Besides that, it shows that for every participant, 75% of the measurements lie under 0.56 mm. In fact, only participant 1 showed an interquartile range slightly higher than 1 mm, as it could also be seen in Table 2.

The ligament correspondent analysis is found at Figure 16 where the limits are the mean of the standard deviation's values for each participant as precision indicator which is 0.88 mm. This value of precision indicator takes into account the P5 outlier of 15.5 mm. The value would drop to 0.71 mm without it and conclusions would not be different, so the original 0.88 mm will be used. But in order to improve visualization, this outlier will be omitted on Figure 16.

It can be concluded from Figure 16 that every measurement error obeys standard deviation limit (precision indicator) and have their mean around zero. This means that even though there are outliers, the measurements were at least 75% precise and accurate.

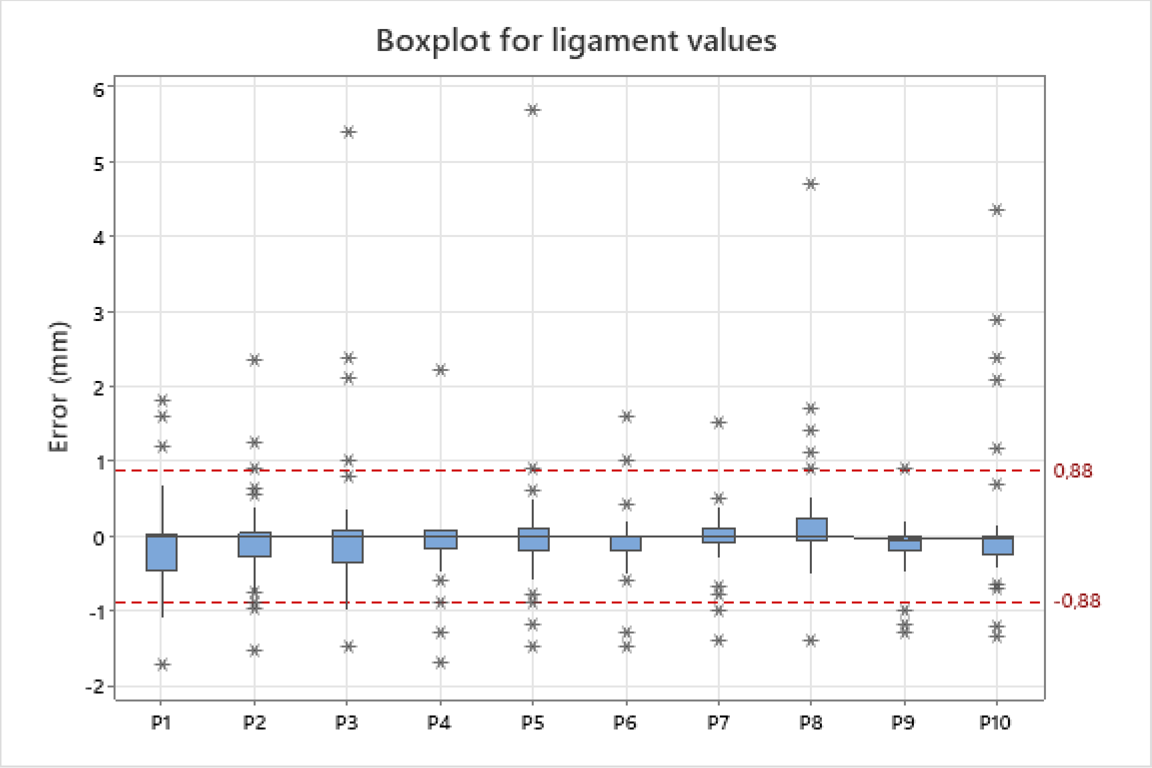


Figure 16 - Centre and spread analysis of ligament error

It is interesting to point that while height measurements also showed mean values around zero, the IQRs were higher than the ones on ligament chart. This means that values were more spread regarding height errors than ligament errors.

In fact, this behaviour can be seen in the histogram representations on Figures 17 and 18. All ranges on x axis are the same for all participants in order to shade light to the difference of their frequencies. Regarding height frequencies, there is a clear difference between P1 and P4 behaviours. The height measurement error of participant 4 are visually concentrated at zero while P1's values were much more spread around zero, as shown previously. Histograms are another way of representing some of the conclusions made so far.

Regarding ligament error values, in Figure 18 it is possible to see the difference of P9 and P5, for example. While the frequency under a normal distribution of P5 error values are wide, mostly because of the 15.5 mm outlier, P9's performance seems to be the most precise and accurate one.

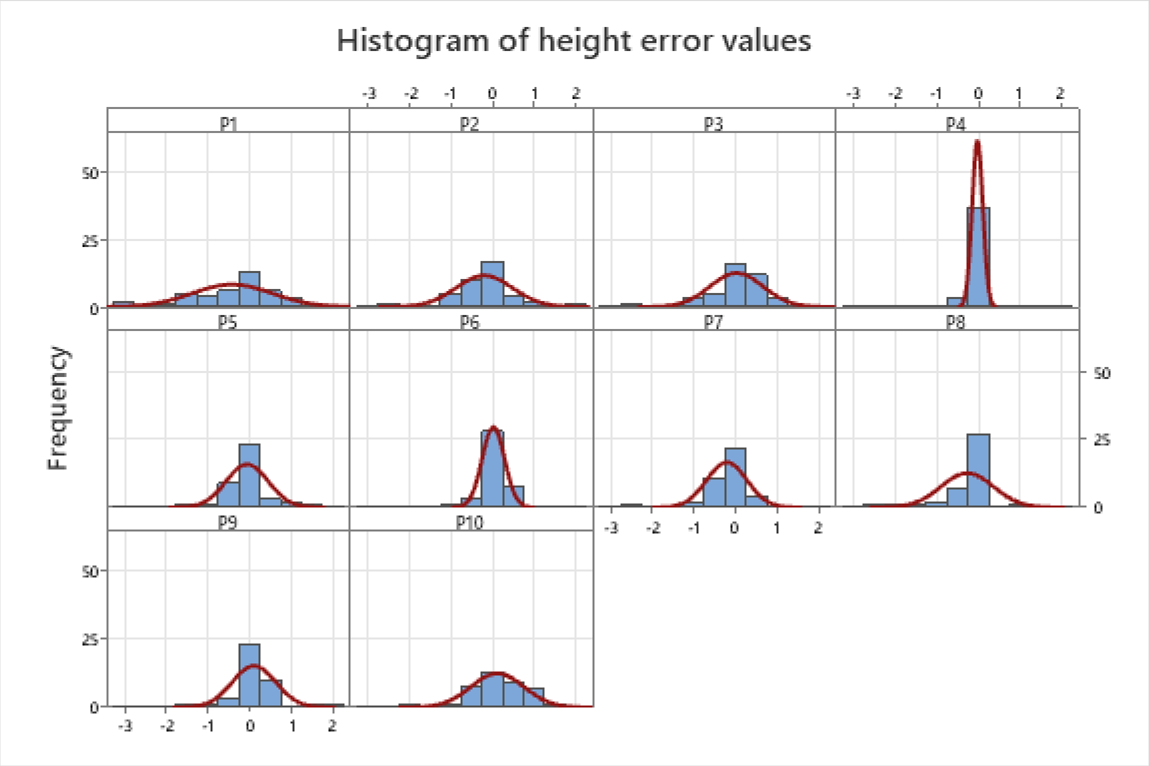


Figure 17 - Histogram of height error values under a normal distribution

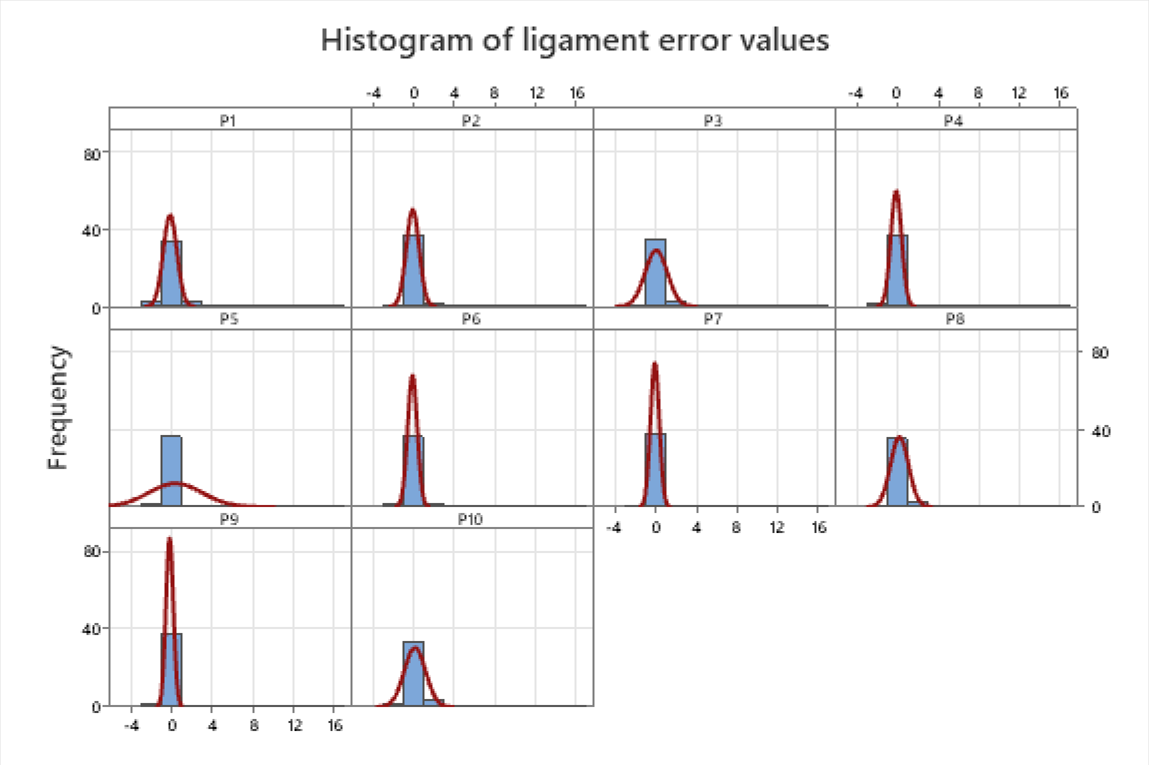
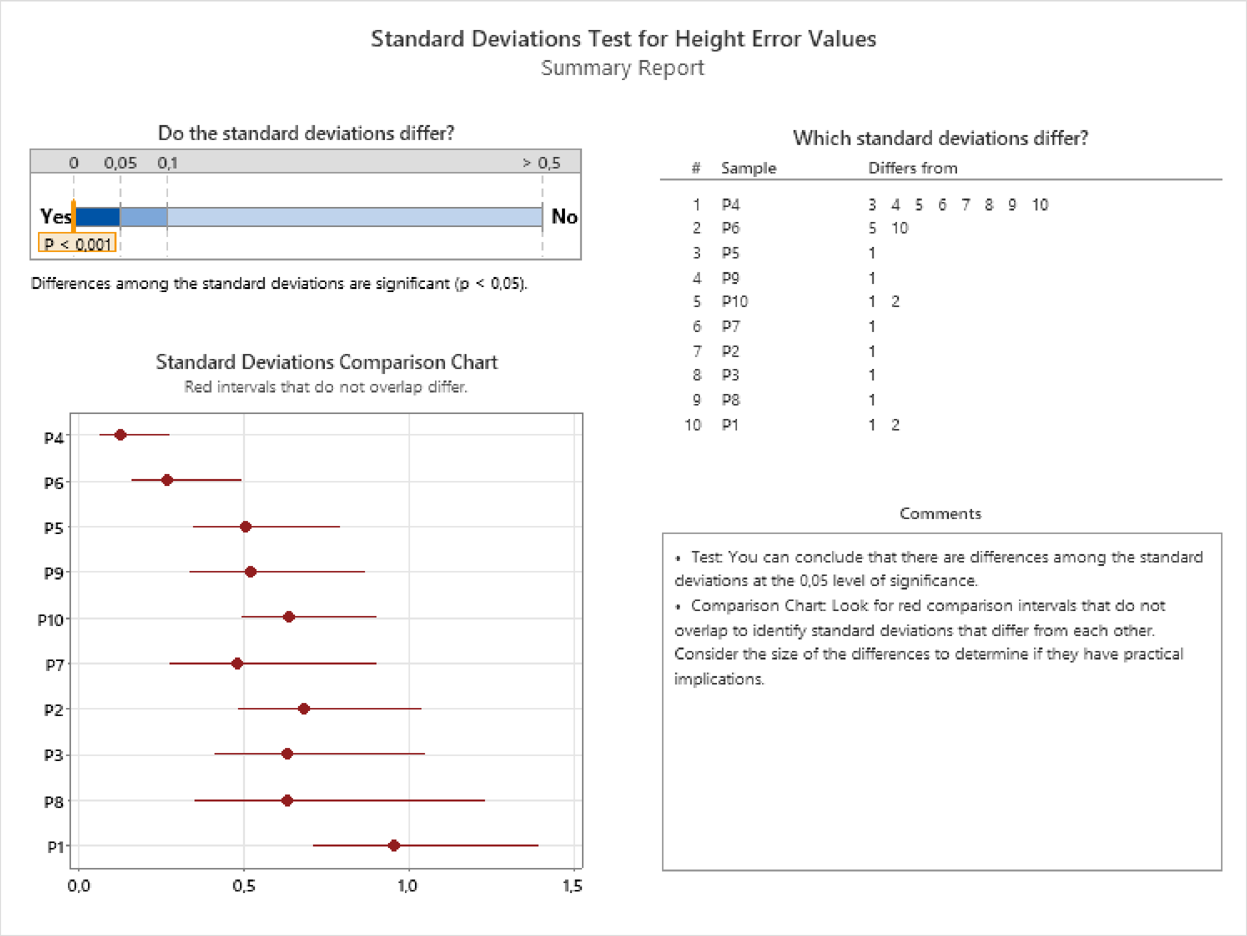


Figure 18 - Histogram of ligament error values under a normal distribution

Although visualization of data could indicate differences in participants' performances, the question would be if these differences are statistically significant.
It is particularly useful to check if there were significant differences between participants' performances because it is what happens in real life: each inspection professional has his/her capability, procedure and/or method. So, establishing if these random, worldwide participants had real differences in their sizing results and therefore on the error values, is valuable information.

In order to check for statistical differences, standard deviation tests were carried out in both variables' data sets, comparing each participant. The standard deviation test is a powerful tool to compare data sets when the data is continuous and the objective is to compare more than 2 samples. In the present case, 10 samples are compared (P1 – P10) and each sample has 40 data each (40 flaws).

The standard deviation test is, in few words, a hypothesis test that compares two or more samples. The question that it answers is if at least two standard deviations intervals do not overlap. For that, it is necessary to stablish the confidence level. The confidence level in this case was set as 95%, meaning that conclusions made using this test have a 95% chance of being correct and 5% otherwise. For a confidence level of 95%, a value called p-value is set as being 5%, or 0.05 and this is the number that is used for rejecting or not rejecting the hypothesis that the samples are equivalent.



Standard Deviations Test for Height Error Values
Summary Report

Do the standard deviations differ?

Differences among the standard deviations are significant (p < 0,05).

Standard Deviations Comparison Chart
Red intervals that do not overlap differ.

Which standard deviations differ?

#	Sample	Differs from
1	P4	3 4 5 6 7 8 9 10
2	P6	5 10
3	P5	1
4	P9	1
5	P10	1 2
6	P7	1
7	P2	1
8	P3	1
9	P8	1
10	P1	1 2

Comments

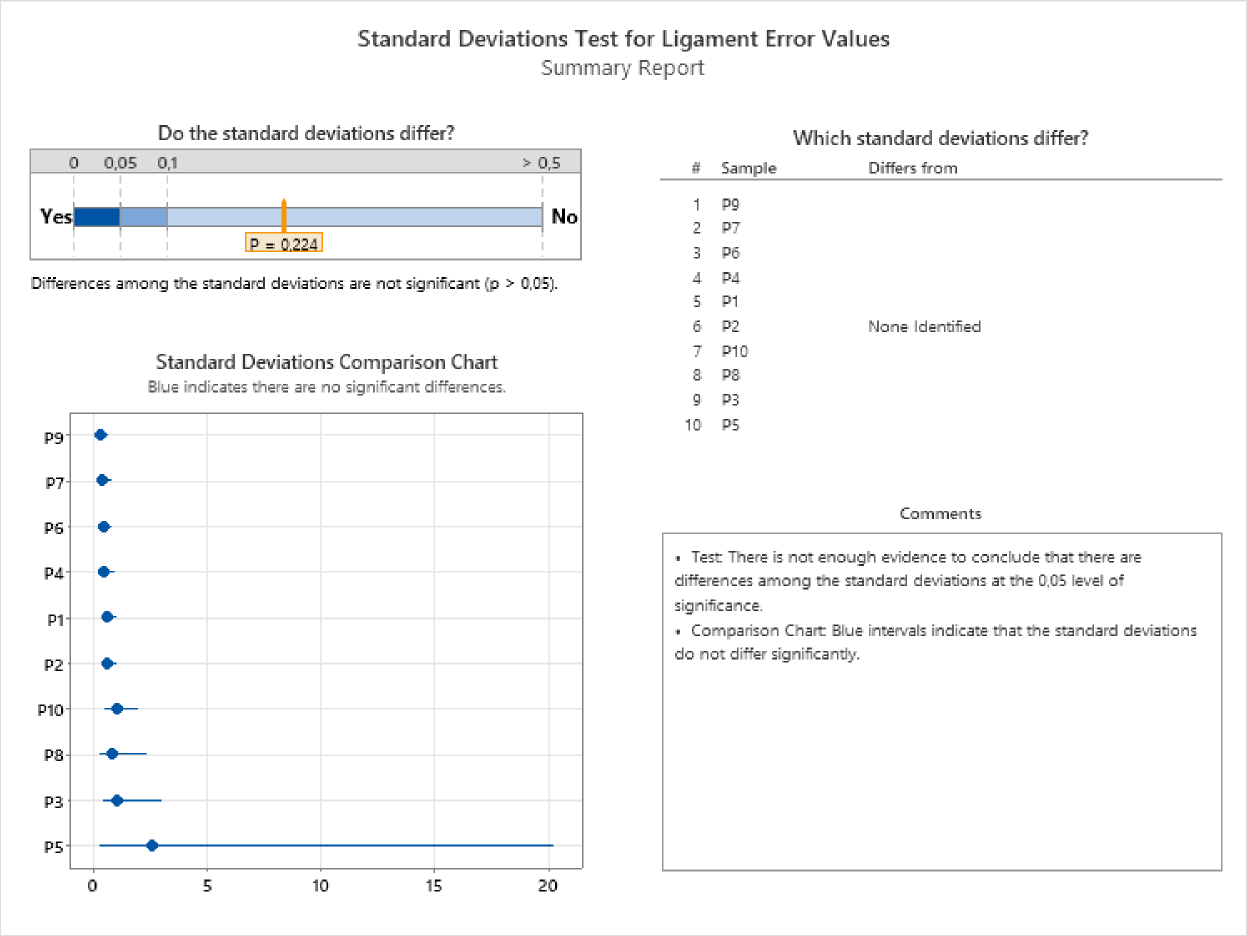
- Test: You can conclude that there are differences among the standard deviations at the 0,05 level of significance.
- Comparison Chart: Look for red comparison intervals that do not overlap to identify standard deviations that differ from each other. Consider the size of the differences to determine if they have practical implications.

Figure 19 - Standard deviation test for height error values between participants

Figure 19 shows the standard deviation test summary report between participants, regarding height measurements error. It is possible to conclude that there are significant differences among participants. For example, the list in Figure 19 on the top and right, shows that P4 differs from # 3, 4, 5, 6, 7, 8, 9 and 10. These # represents P5, P9, P10, P7, P2, P3, P8 and P1 respectively. At the same time, P8 (#9) is significant different only to P4 (#1).

The fact that there are significant differences between participants is very good news because it means that even with different height sizing methods and error values, they all showed results both accurate and precise.

As for ligament results, Figure 20 shows that no statistical difference could be observed even when the P5's 15.5 mm outlier was considered. The p-value for both cases, with and without the outlier, was the same: 0.224, which is bigger than 0.05 meaning that it is not possible to reject the hypothesis that there are differences between the samples, considering a confidence level of 95%.

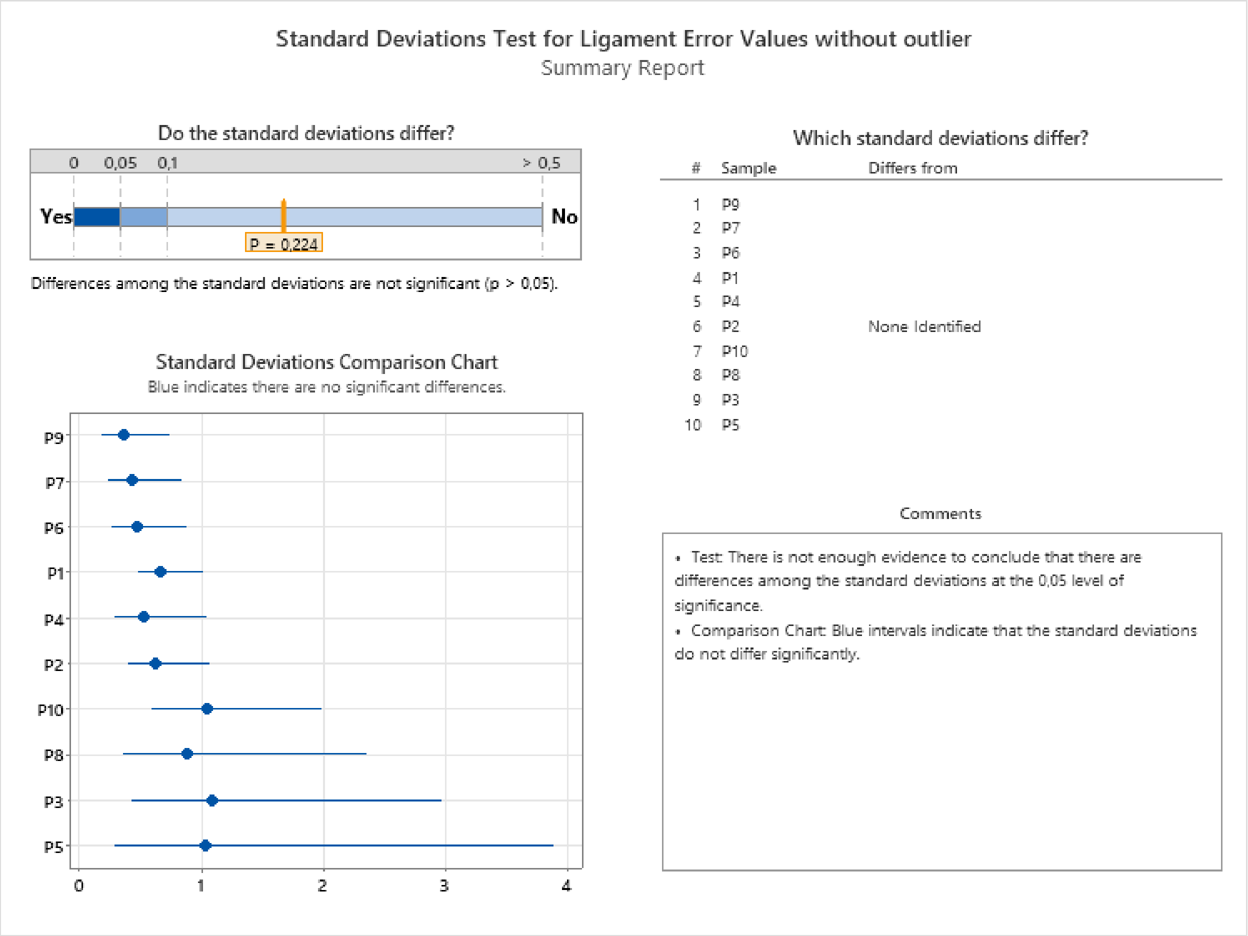


Differences among the standard deviations are not significant (p > 0,05).

#	Sample	Differs from
1	P9	
2	P7	
3	P6	
4	P4	
5	P1	
6	P2	None Identified
7	P10	
8	P8	
9	P3	
10	P5	

- Test: There is not enough evidence to conclude that there are differences among the standard deviations at the 0,05 level of significance.
- Comparison Chart: Blue intervals indicate that the standard deviations do not differ significantly.

(a)



Standard Deviations Test for Ligament Error Values without outlier

Summary Report

Do the standard deviations differ?

Differences among the standard deviations are not significant (p > 0,05).

Standard Deviations Comparison Chart

Blue indicates there are no significant differences.

Which standard deviations differ?

#	Sample	Differs from
1	P9	
2	P7	
3	P6	
4	P1	
5	P4	
6	P2	None Identified
7	P10	
8	P8	
9	P3	
10	P5	

Comments

- Test: There is not enough evidence to conclude that there are differences among the standard deviations at the 0,05 level of significance.
- Comparison Chart: Blue intervals indicate that the standard deviations do not differ significantly.

(b)

Figure 20 - (a) Standard deviation test for ligament error values between participants with the 15.5 mm outlier of P5 (b) and Standard deviation test for ligament error values between participants without the 15.5 mm outlier of P5

Until this point, all data is being analysed regardless of their order. A slightly different display can be made if instead of organising the data with respect to the sample number for each participant, the samples were arranged in order of error magnitude. If the sample is put in order with the first data the lowest number on each data sample (participant) and then growing until the highest number (higher error value).

Treating the data that way, it is possible to calculate the cumulative relative frequency which is very useful to determine what percentage of the sample each error value occupies. Figure 21 represents the cumulative relative frequency of all participants regarding height error sizing. It can be seen that most data are comprehended between -1.0 mm and 1.0 mm.

For better visualization, Figure 22 shows each one separately, still for height error. It is worthy to mention that the y axes are all the same but the x axes are relative to each sizing error interval. It can be seen that while participants 4 and 10 have their curves crossing the zero mark on the 50%, which means that half of their errors was equal or below zero, participant 2 had the curve crossing the point of 50% around -0.7 mm. That means that P2 had 50% of his/her data below -0.7 mm height error.

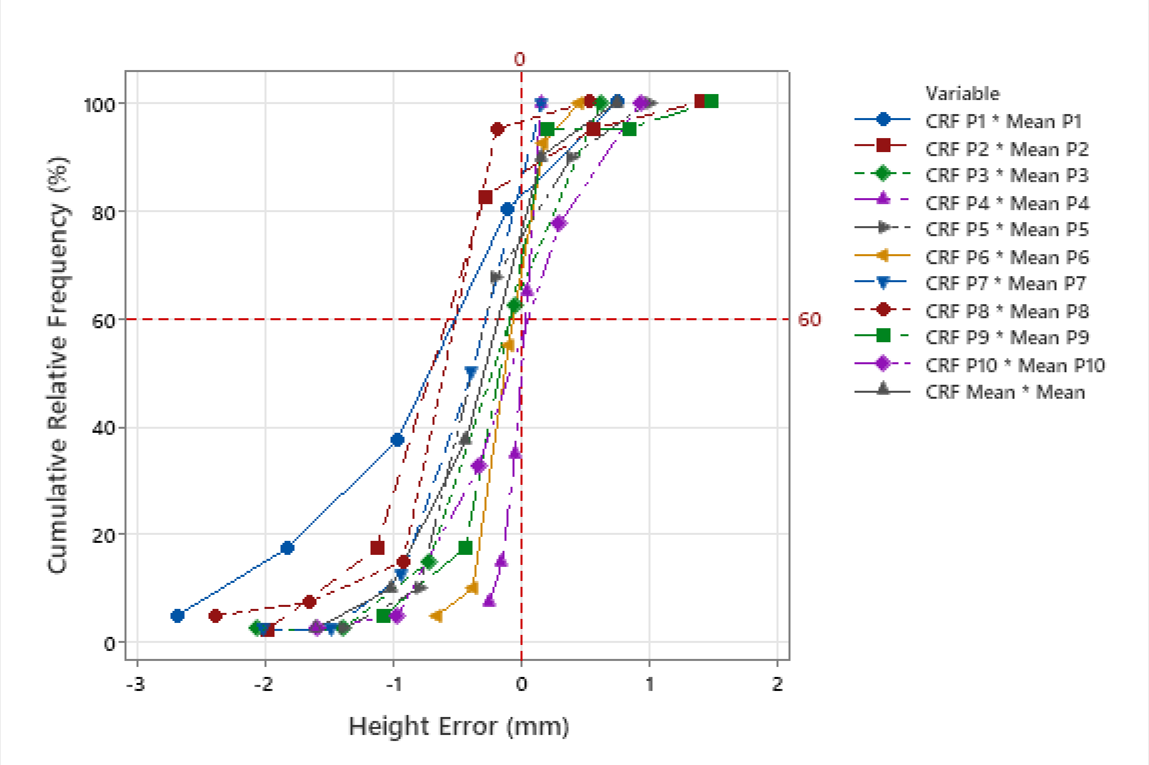


Figure 21 – Cumulative relative frequency for all participants regarding height error values

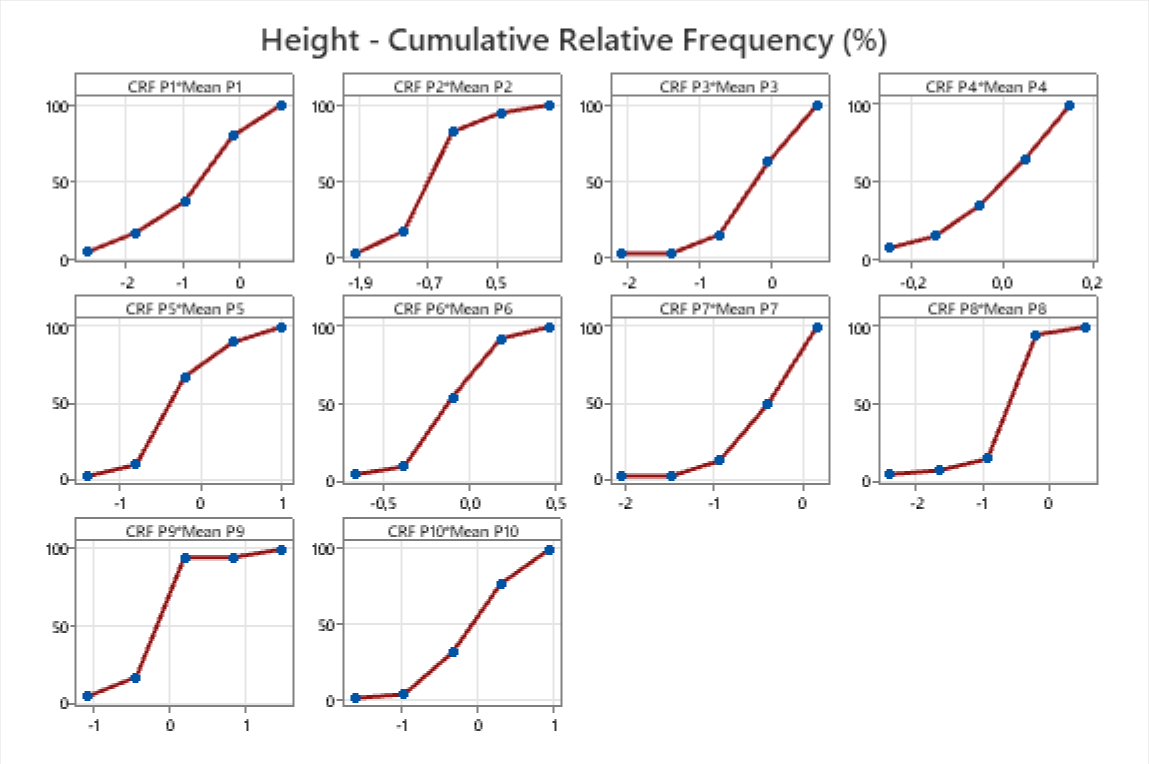


Figure 22 – Cumulative relative frequency for each participant regarding height error values

The participants used different CIVA versions and methods in order to size the flaws. While some used 2021 version using classical TFM sizing method, others used 2023 version and the coherence factor (CF) option.

Of the 10 participants, four used CIVA 2021 and classical TFM sizing method while 4 others used CIVA 2023 and coherent factor options. The other two participants did not use CIVA to size the flaws' height or used both methods: classical and CF. For that reason, these two participants were left out of this particular comparison.

Figure 23 analyses the data of those who used CIVA 2021 and classical TFM sizing method. The mark of 70% was traced for comparison purposes only. It is undeniable that this sizing method shows a good level of accuracy since 70% of sizing results, in mean, are around zero, according to the mean values line in grey.

The mean lines that are going to be used in order to qualify the cumulative relative frequencies regarding CIVA version and sizing methods were calculated based only at the specific participants of each group. That means that the mean line is not based on the general results but specific on those that share the same characteristics.

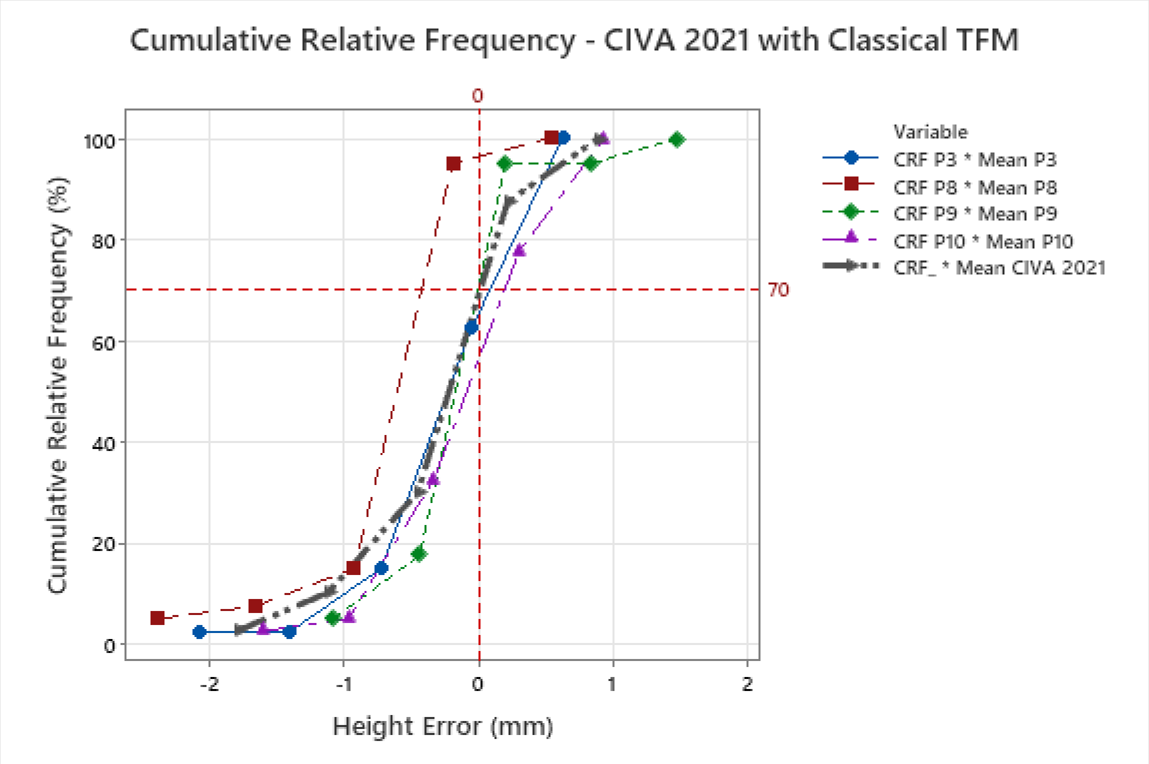


Figure 23 - Cumulative relative frequency regarding CIVA 2021 and classical TFM height sizing method

However, when the alternative samples are analysed, those that used CIVA 2023 and CF options, it is very clear that results are improved. Figure 24 shows that at the mark of ~83%, all participants converge closer to zero error. For P4, P6 and P7, the worst-case scenario is that ~83% of the data is concentrated between -2 mm and zero.

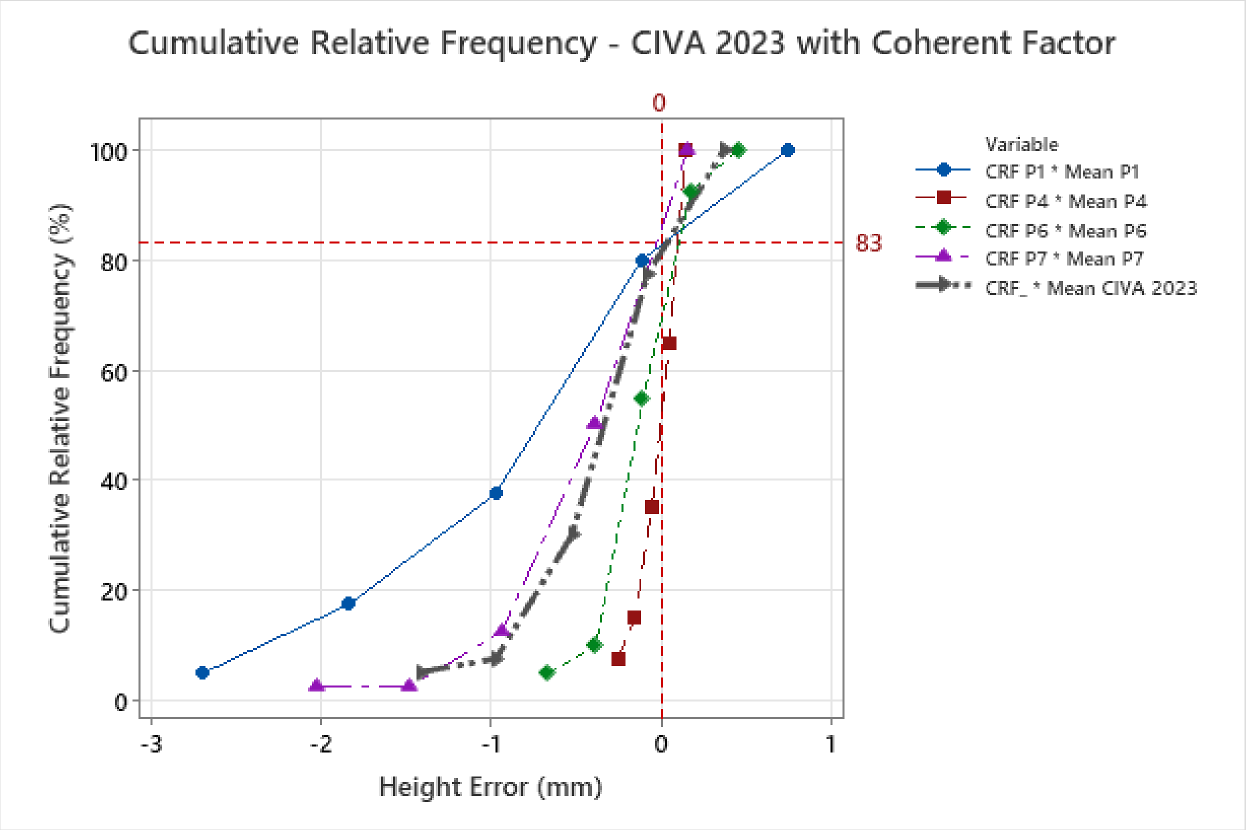


Figure 24 - Cumulative relative frequency regarding CIVA 2023 and coherent factor for height sizing

Performing the same analysis for ligament variable, the issue of the 15.5 mm outlier for ligament measurement by P5 is again addressed. Figure 25 shows the cumulative relative frequency for ligament errors and because of this specific outlier, the scale of the graphical representation of the other participants becomes out of proportion. Having said that, this particular outlier is going to be again left out for better data visualization (see Figure 26). The data distribution of P5 data in that case was recalculated considering 39 flaws instead of 40.

Unlike the cumulative relative frequency distribution regarding height, the difference between distribution regarding ligament error showed to be much narrower. This phenomenon was already observed on previous results. One of the interesting things to point out is that from Figure 27, which show only the mean value of ligament results, 80% of all data are between -0.93 mm and zero error values. Which means that almost all data are comprehended between zero error and a value near the ligament precision indicator that is 0.88 mm, showing good tendency of height accuracy.

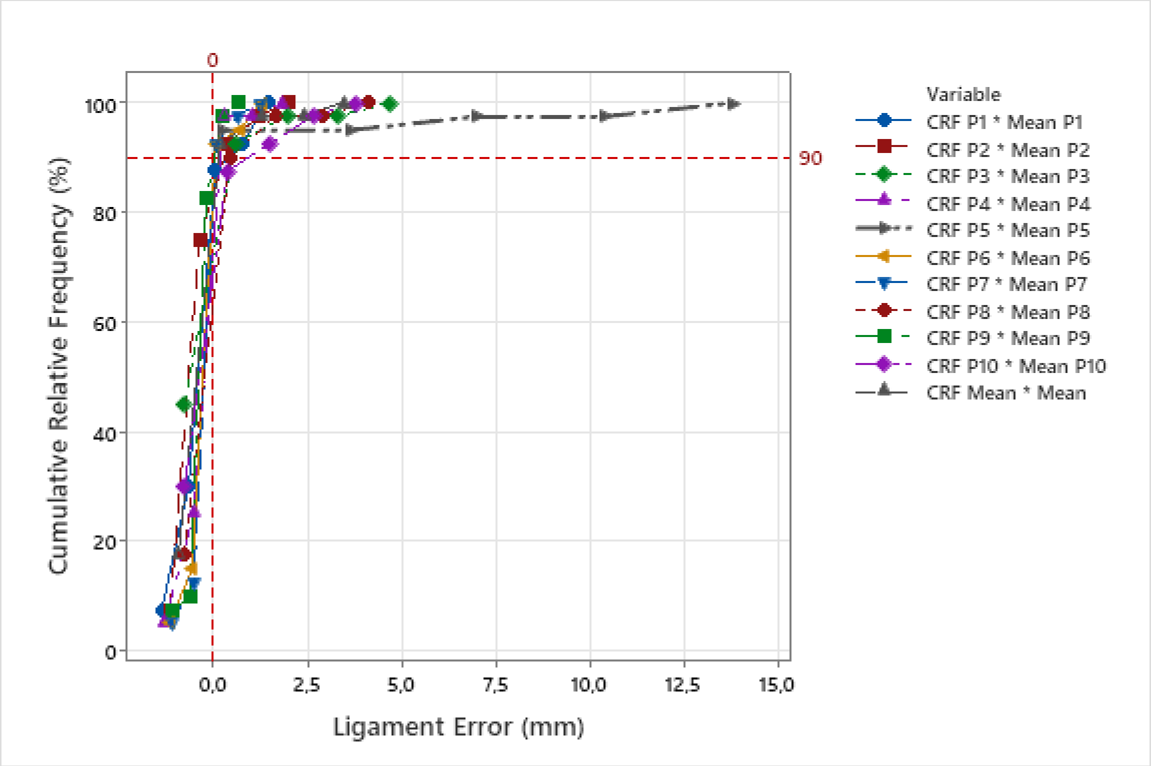


Figure 25 - Cumulative relative frequency for all participants regarding ligament error values

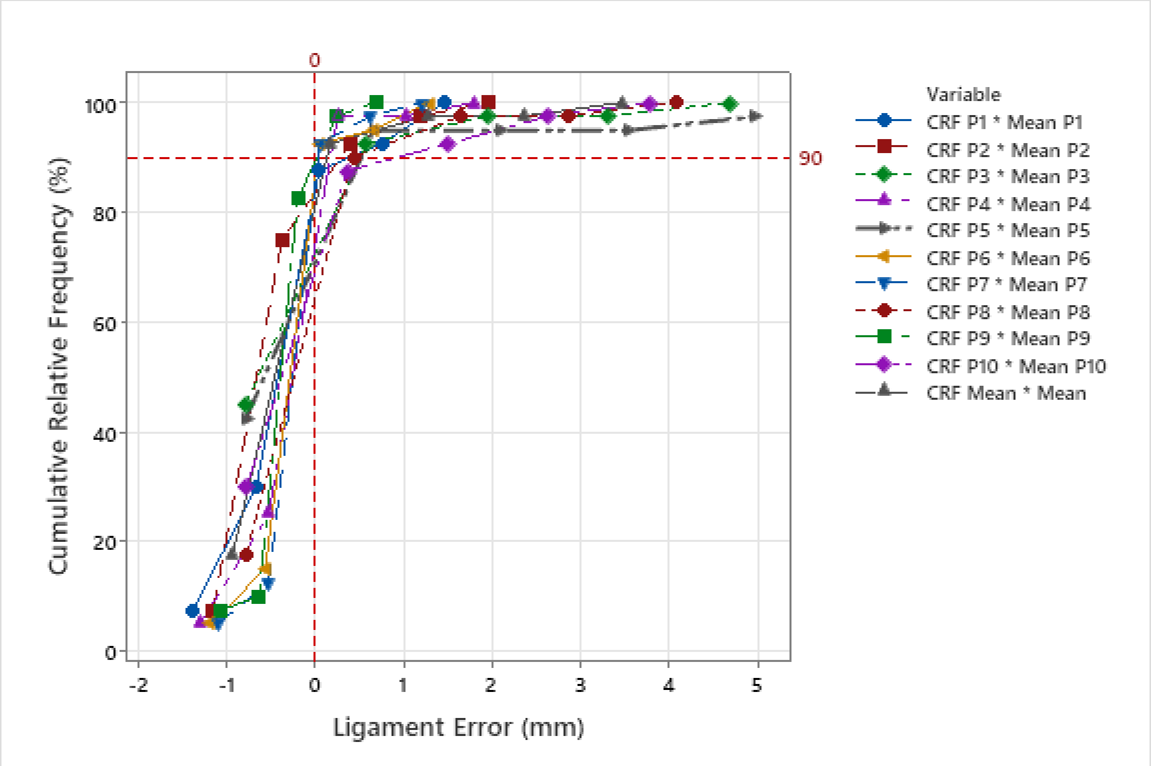


Figure 26 - Cumulative relative frequency for all participants regarding ligament error values **with** censored outlier

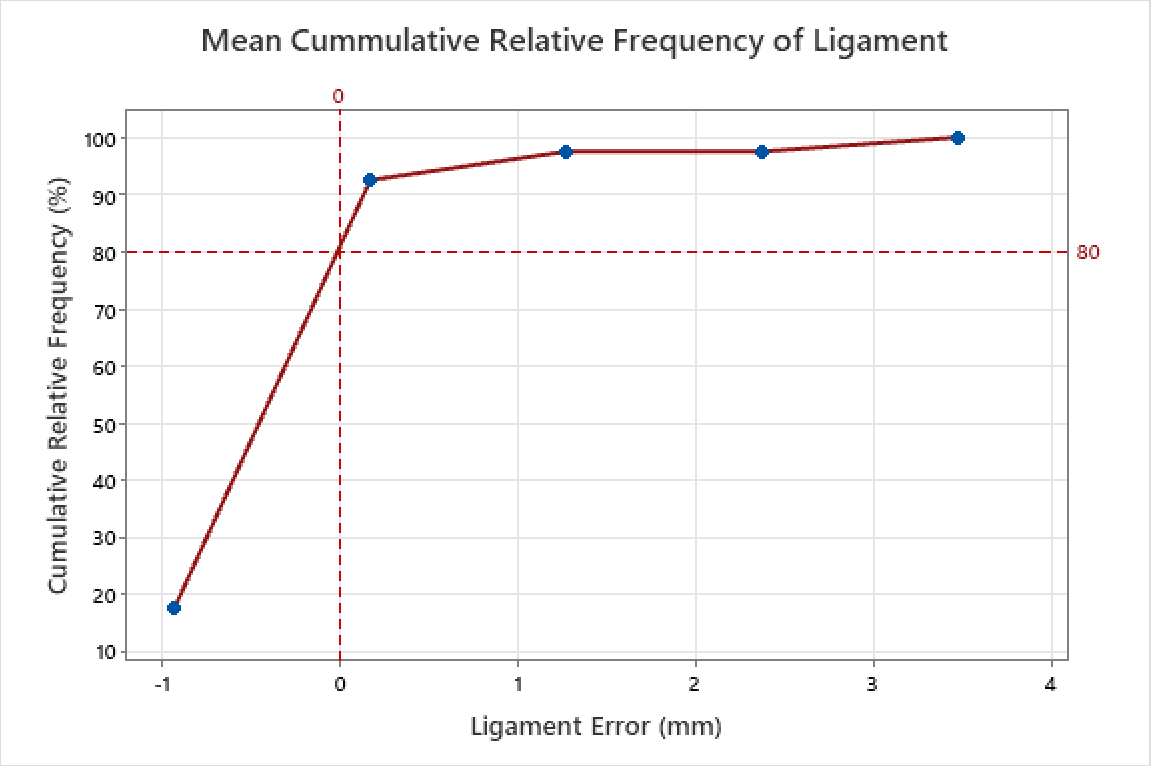


Figure 27 - Mean Cumulative relative frequency for all participants regarding ligament error values **with** censored outlier

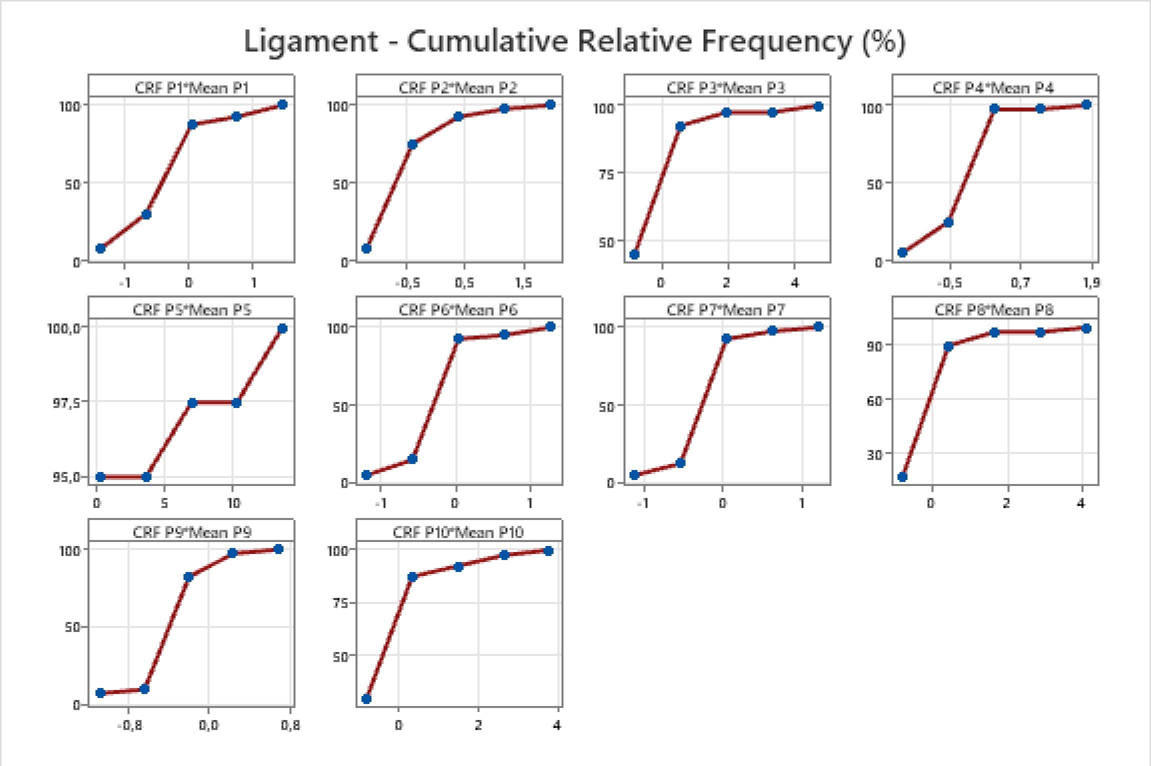


Figure 28 - Cumulative relative frequency for each participant regarding ligament error values with **no** censured outlier

When the participants are observed individually, as is shown by Figure 28, it is very clear that there were similarities that stood out such as P1, P6, P7 and P9. All four participants present a ligament error interval between around ± 1.0 mm with a little less than 100% of the data concentrated between -1.0 mm and zero.

The more symmetrical the graphs (Figure 28) are around zero, regardless the interval, in the way that the curve crosses the 50%, it means that there were half of the data that was oversized and half undersized. P6, for example, presented very few data between zero and 1.0 mm since almost 100% of the data was between -1.0 mm and zero. On the other hand, P8 presented 60% of negative ligament size error and 90% of positive near zero and less.

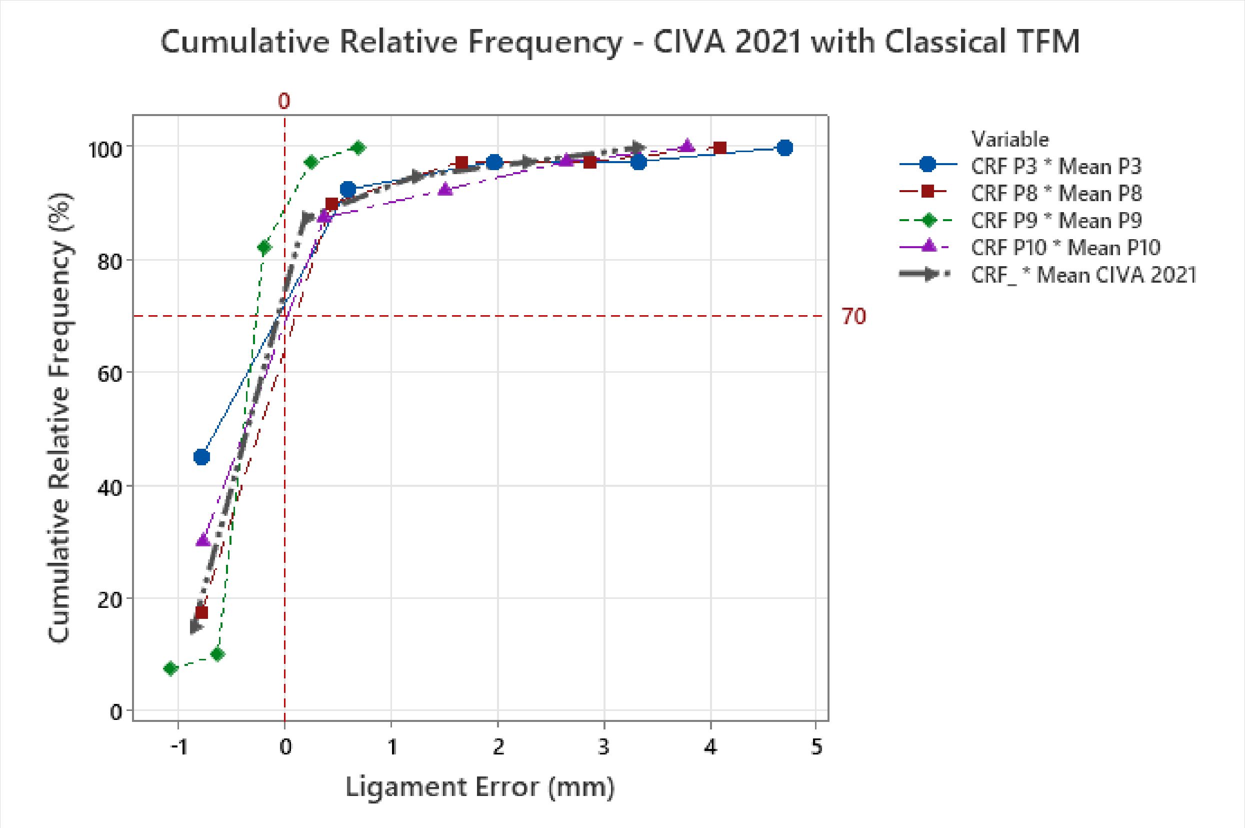


Figure 29 - Cumulative relative frequency regarding CIVA 2021 and classical TFM ligament sizing method

Figure 29 shows the cumulative percentage of data for the participants that used CIVA 2021 with classical TFM sizing method. It can be concluded that participants presented 70% of their data, in mean, between zero error and -1 mm ligament error. Although the result is numerically the same as shown in Figure 27, the mean values are not the same since the samples are not the same (different groups of participants). When these results are compared with CIVA 2023 and the use of coherence factor, as can be seen in Figure 30, almost the same behaviour is observed: 85% of ligament errors, in mean, also respected the same interval from -1.0 mm to zero error.

These results show that for ligament determination, both CIVA 2021 and 2023 and both classical TFM and coherence factor, showed practically the same efficiency. Both were accurate as well as precise in the same way. This conclusion only corroborates the standard deviation tests that stated that there were no significant statistical differences between participants regarding ligament.

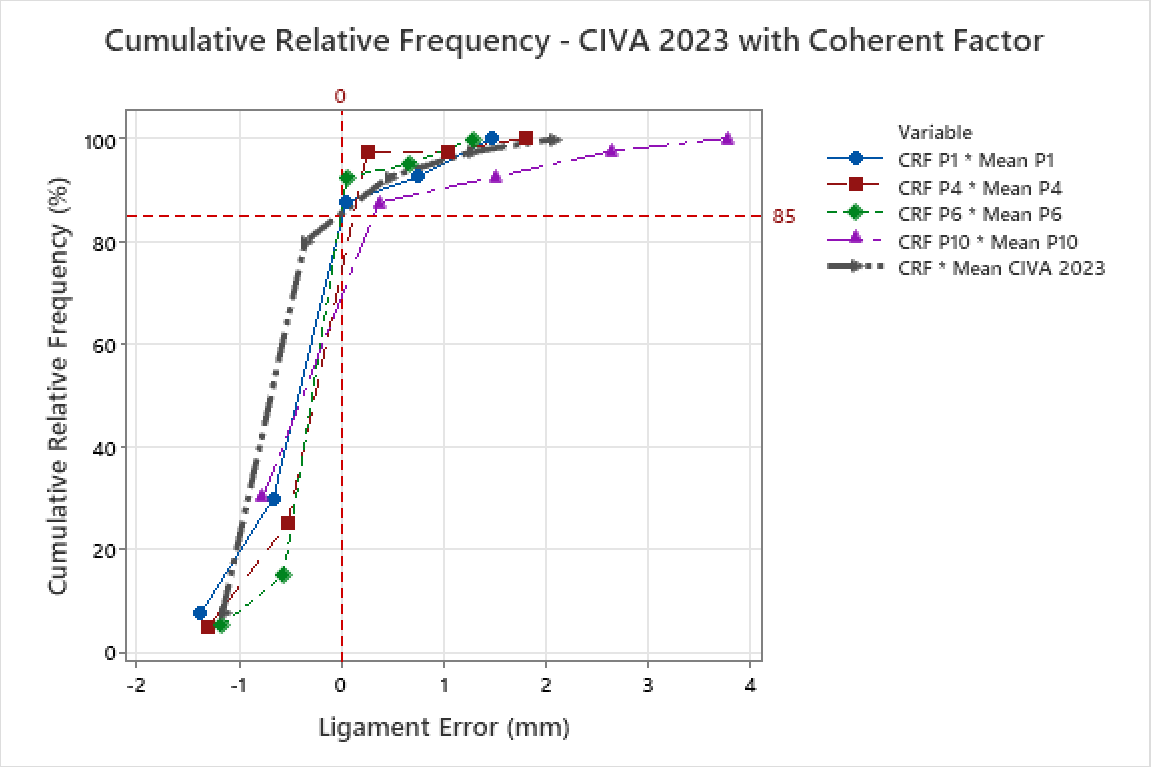


Figure 30 – Cumulative relative frequency regarding CIVA 2023 and coherence factor for ligament sizing

For an overall final visualization of the data gathered during this round-robin step, Figure 31 shows the parallel coordinates plot without the P5 outlier. When compared with Figure 32, it becomes clear, once again, that ligament values showed to be accurate and precise as well as height sizing measurements.

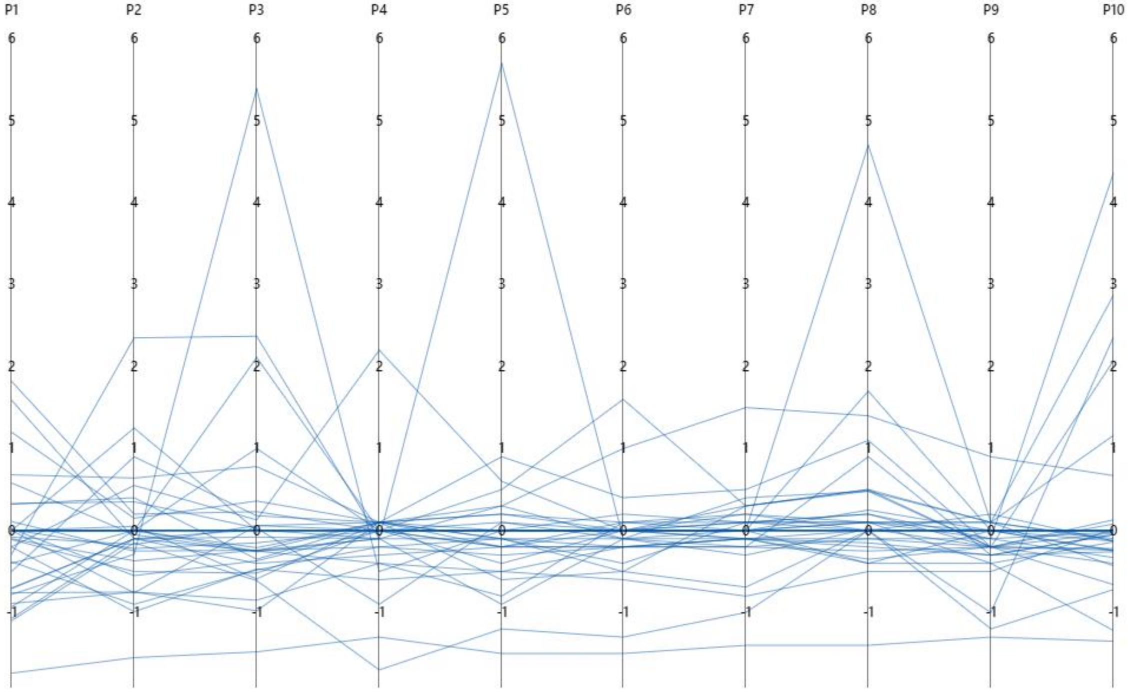


Figure 31 - Parallel plot showing the overall data scattering for ligament values

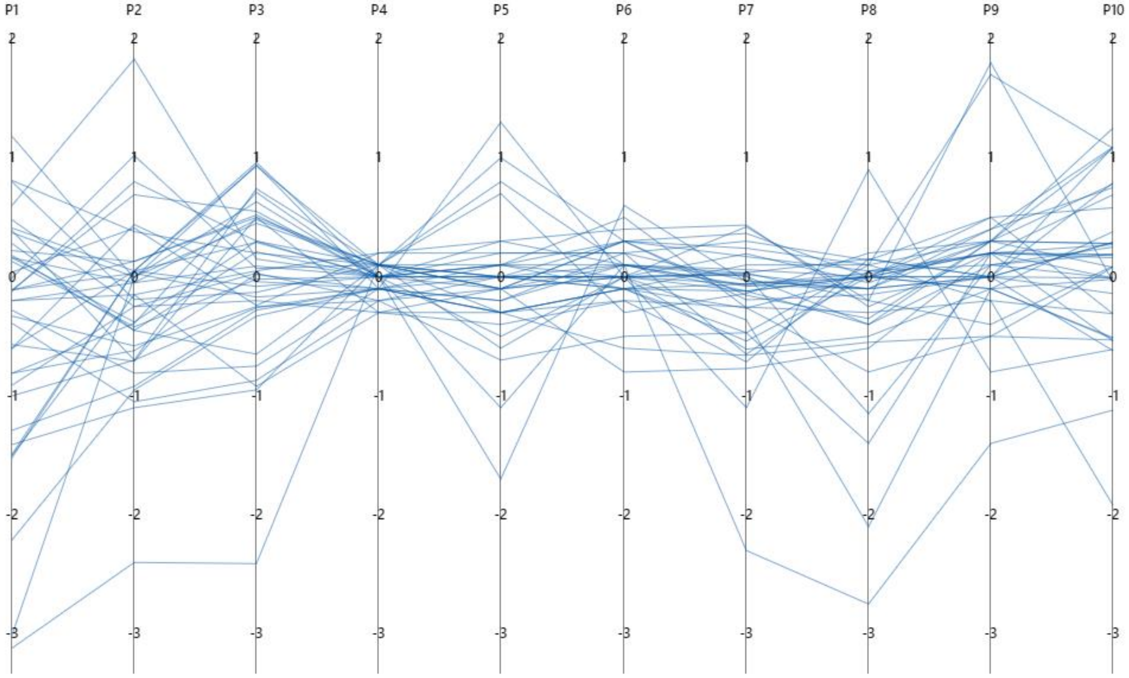


Figure 32 - Parallel plot showing the overall data scattering for height values

6. Conclusions and Future Work

The evidence from the round-robin using CIVA-simulated weld flaws and TFM sizing tools suggests that TFM can provide statistical sizing capabilities that are superior to any of the alternative traditional ultrasonic techniques (as described in BS-7910). Standard deviations of error in this round-robin approach the accuracies achieved using destructive test methods.

The average standard deviation of height sizing for all participants in this round-robin was 0.56mm. The average standard deviation of ligament sizing was 0.88mm.

Trends in individual performances suggest that experience and tools available for analysis can have an effect on the precision that can be achieved using TFM. Statistics suggest that the use of coherence factor TFM can produce better precision in sizing. Appendix 2 illustrates how the selection of modes and TFM algorithm can play a rôle in the image quality suitable for sizing analysis.

For NDT reliability qualifications, where PODs are to be used, the simulations suggest that the true values from the TFM sizing would provide more accurate values than many of the flaw manufacturers estimates.

Future work is planned to validate the results obtained by these simulations. Lab data are planned that would use FMC (or PWI) acquisition techniques that are post-processed with TFM algorithms. When results of the sizing from simulated data are combined with lab data, it is expected that a guideline document can be prepared to aid in preparations of POD programmes where flaw sizes can be gauged using TFM algorithms.

7. Acknowledgements

This project involved a lot of volunteer work. It would not have been possible without the participants and their companies' contribution of time and effort. We would like to thank the following individuals and their companies.

Claus Aroldo	Nucleoeléctrica Argentina S.A.
Mohammad Bajgholi	Université ÉTS Montréal, Canada
Ewen Carcreff	TPAC (The Phased Array Company), France
Michel Cence	Onet Technologies, France
Vincent Chardome	AIB Vinçotte, Belgium
Gustav Holmer	Kiwa AB, Sweden
Weina Ke	Onet Technologies, France
Prathik Prabhakara	BAM (Federal Institute for Materials Research and Testing) Germany
Frédéric Reverdy	EddyFi, Canada
Gilles Rousseau	Hydro Québec, Canada

We would also like to thank Fabrice Foucher and Extende (France) for always being there to help us with any challenges we encountered with CIVA.

Appendix 1 Round-Robin Flaw List

Item	Bevel	Wall Thickness	Land	Gap	Plate/Pipe	Height (vertical)	Length	Ligament (near/far)	Type
1	60° V	10	2	2	Plate		15		IP
2	60° V	10	2	2	Plate		15.3		LoF
3	60° V	10	2	2	Plate		10.7		Toe Crack
4	60° V	12.5	2	2	Pipe (6”)		13.3		LoF
5	60° V	12.5	2	2	Pipe (6”)		17		IP
6	60° V	12.5	2	2	Pipe (6”)		18.8		Toe crack
7	60° V	19	2	2	Pipe (8”)		12.7		LoF
8	60° V	19	2	2	Pipe (8”)		10		Under-bead crack
9	60° V	19	2	2	Pipe (8”)		7.9		LoF
10	60° V	19	2	2	Pipe (8”)		7.8		LoF
11	55° X	25	3	3	Plate		9		IP
12	55° X	25	3	3	Plate		13		IP
13	55° X	25	3	3	Plate		14		Toe Crack
14	55° X	25	3	3	Plate		11.1		LoF
15	55° X	25	3	3	Plate		12.9		LoF
16	60° X	32	4	3	Plate		13.9		LoF
17	60° X	32	4	3	Plate		14		LoF
18	60° X	32	4	3	Plate		9		Centreline crack
19	60° X	32	4	3	Plate		10.1		IP
20	60° X	32	4	3	Plate		20		Interpass at IP
21	60° X	32	4	3	Plate		10		LoF
22	3° U	28	2	1	Pipe 14”		13		IP Root hilo
23	3° U	28	2	1	Pipe 14”		6		BT Root hilo
24	3° U	28	2	1	Pipe 14”		9		HotPass LoF
25	3° U	28	2	1	Pipe 14”		8.1		HotPass LoF
26	3° U	28	2	1	Pipe 14”		13.8		LoF
27	3° U	28	2	1	Pipe 14”		5.9		LoF (triangular)
28	3° U	28	2	1	Pipe 14”		9		Stop-start
29	3° U	28	2	1	Pipe 14”		7		Stop-start
30	3° U	28	2	1	Pipe 14”		6.7		LoF
31	3° U	28	2	1	Pipe 14”		6.6		Centreline crack
32	CRC	19	1.27	0	Pipe 48”		11.7		Missed Edge
33	CRC	19	1.27	0	Pipe 48”		9.5		LoF Fill
34	CRC	19	1.27	0	Pipe 48”		7.5		F1/HP (with silica)
35	CRC	19	1.27	0	Pipe 48”		7.2		HP
36	CRC	19	1.27	0	Pipe 48”		21		HP/LCP
37	CRC	19	1.27	0	Pipe 48”		8.4		LCP
38	CRC	19	1.27	0	Pipe 48”		16		LCP/Root
39	CRC	19	1.27	0	Pipe 48”		16		Misfire root
40	CRC	19	1.27	0	Pipe 48”		4		Burn Through

Appendix 2 Comparing Image Quality for Sizing

With knowledge of the flaw type that is being sized, it is feasible to select appropriate modes to optimise the image for TFM sizing. As an example, Figure A1 illustrates a simple planar flaw simulating lack of fusion is placed in a 20mm thick U-bevel butt weld. The flaw is a nominal 5mm x20mm rectangle aligned with the 10° bevel angle. This gives the flaw a 4.9mm vertical extent.

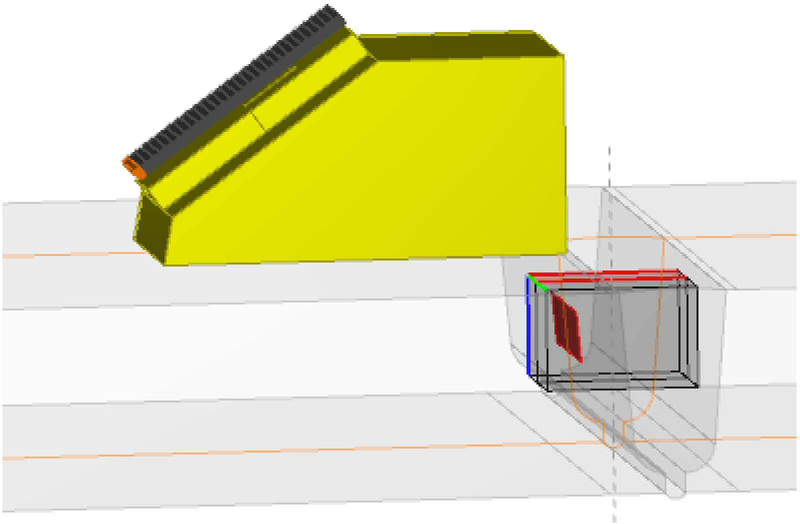


Figure A1 Planar flaw in U-Bevel weld

Potentially useful modes would be selected to include paths that have a pulse that produces a signal returning to the probe after direct interaction (TT and TTTT) that might provide tip echoes, and a mode that would confirm the planar nature of the flaw (TTT). These are described in Figure A2.

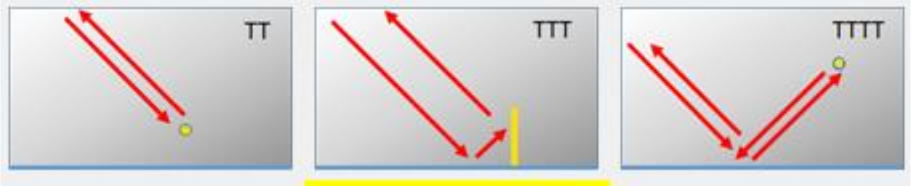


Figure A2 Mode options to optimise TFM image

CIVA allows for the individual modes to be imaged as well as an image that combines the amplitudes in a separate image using the maximum amplitude from each, the average of all of the modes or the sum of all of the modes. Holloway [11] demonstrated how details of facets in TFM reconstruction of flaw images is improved when multiple modes are used, so we select to display the sum of the modes in addition to the individual mode images.

Using the Classical TFM algorithm with a 0.1x0.1mm grid and the Envelope signal with Times of Flight by Beam Computation, the images with the weld overlay (including the CIVA flaw representation) show how the differences in presentation can enhance or detract from sizing analysis. Results using the Classic TFM algorithm are shown in Figure A3.

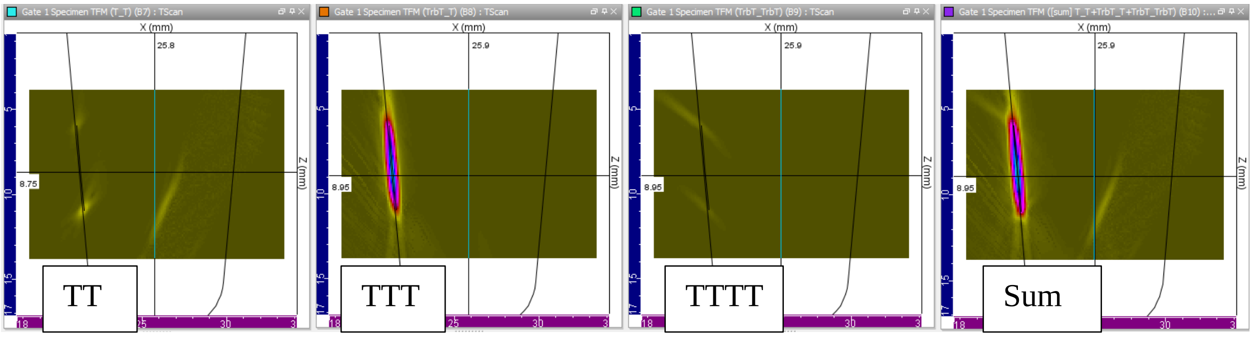

Gate 1 Specimen TFM (T_T) (B7) : TScan
Gate 1 Specimen TFM (TrbT_T) (B8) : TScan
Gate 1 Specimen TFM (TrbT_TrbT) (B9) : TScan
Gate 1 Specimen TFM ([sum] T_T+TrbT_T+TrbT_TrbT) (B10) :...
X (mm)
Z (mm)
25.8
25.9
8.75
8.95
TT
TTT
TTTT
Sum

Figure A3 Classic TFM using TT, TTT, TTTT and Sum of all 3 modes

In the TT image an artefact can be seen near the weld centreline and weak tip echo signals are associated with the ends of the flaw but they appear as extended lines. The TTT image indicates a strong planar indication due to the ideal specular path in the TTT mode. The TTTT mode produces similar weak lines associated with the tips and no artefact is present. With the ability to sum the amplitudes from the three images, there appears to be no particular advantage. The strong specular reflection in the sum image occludes the weak tip signals. Many technicians rely on a dB drop to size the flaw from the strong specular image. In this case, the -6dB sizing would result in a flaw 5.1mm high and at -3dB the height would be estimated at 3.8mm.

CIVA 2023 has added several new TFM algorithms that can be used to help identify tip echoes more precisely. Using the same 3 modes plus their summation we then examine the algorithm identified as Coherence Factor (CF). This was described by Suttcliffe [12] describing an equation where the coherent energy in a synthetic aperture imaging sum is divided by total energy for the delayed signals in the synthetic aperture imaging sum. The end result of using CF is to supress noise and artefacts. Using the CF algorithm in CIVA 2023 produces an artefact-free image with more pronounced tip indications as seen in Figure A4. The Tip indications are most clearly defined in the TT mode using CF and allows sizing of the planar flaw as 4.95mm.

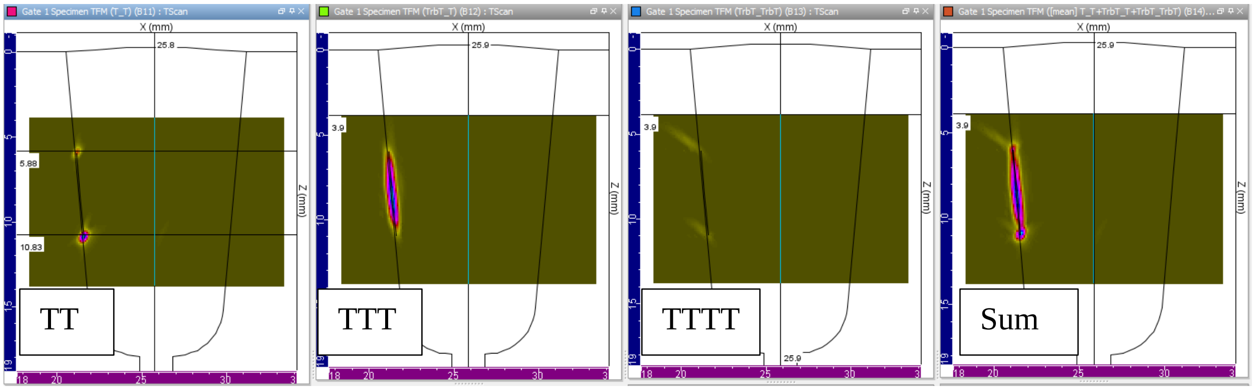

Gate 1 Specimen TFM (T_T) (B11) : TScan
Gate 1 Specimen TFM (TrbT_T) (B12) : TScan
Gate 1 Specimen TFM (TrbT_TrbT) (B13) : TScan
Gate 1 Specimen TFM ([mean] T_T+TrbT_T+TrbT_TrbT) (B14)...
X (mm)
Z (mm)
25.8
25.9
5.88
10.83
3.9
TT
TTT
TTTT
Sum

Figure A4 Coherence Factor TFM using TT, TTT, TTTT and Sum of all 3 modes

References

1. Ginzel, E., Bajgholi, M.E., Burrowes, M. Guimarães, M., Foucher, F., Rousseau, G., Viens, M., Total Focusing Method used for flaw sizing in probability of detection determination, NDT.net, July 2022, https://www.ndt.net/article/ndtnet/papers/TFM_used_for_flaw_sizing_for_probability_of_detection_determination.pdf
2. ASME Boiler and Pressure Vessel Code, Section V Art. 14, The American Society of Mechanical Engineers, New York, USA 2021.
3. ASTM E2862-18 Standard Practice for Probability of Detection Analysis for Hit/Miss Data, ASTM International, 2018
4. ASTM E3023-21 Standard Practice for Probability of Detection Analysis for â versus a Data, ASTM International, 2021
5. DNVGL ST-F101 Submarine Pipeline Systems, DNVGL-Norway 2017
6. ENIQ Recommended Practice #5, Guidelines for the design of test pieces and conduct of test piece trials, Publications office of the European Union, Netherlands, 2011
7. MIL-HDBK 1823A Nondestructive Evaluation System Reliability Assessment, Department of Defence, United States of America, 2009
8. Voillaume, H., Costain, J., Murray, D., Phased Array Ultrasound and Total Focusing Method for the detection and sizing of HTHA and HIC in critical pressurized components, NDT.net March 2021, https://www.ndt.net/article/nde-india2019/papers/Voillaume.pdf
9. Peng, C., Bai, L., Zhang, J., Drinkwater, B.W., The sizing of small surface-breaking fatigue cracks using ultrasonic arrays, NDT and E International 99, Elsevier (2018)
10. British Standard BS 7910:2019, Guide to methods for assessing the acceptability of flaws in metallic structures, BSI Group, 2019
11. Holloway, P., Ginzel, E., TFMi: Using Intermodal Analysis to Improve TFM Imaging, https://www.ndt.net/article/ndtnet/papers/TFMi_-_Using_Intermodal_Analysis_to_Improve_TFM_Imaging.pdf, NDT.net May, 2021
12. Sutcliffe, M., Charlton, P., Weston, M., Multiple virtual source aperture imaging for non-destructive testing, Insight Vol 56 No 2 February 2014