Statistical Assessment of the Laboratory Performance Achieved in Charpy Impact and Tensile Testing Scenarios: A Comprehensive Proficiency Testing Approach

Askarli Hasan
University of Pannonia
Csaba Hegedűs
University of Pannonia
Zsolt T. Kosztyán
kosztyan.zsolt@gtk.uni-pannon.hu

University of Pannonia

Research Article

Keywords: Charpy impact test, proficiency testing, interlaboratory comparison, tensile testing, laboratory performance evaluation, z score analysis, mechanical testing variability

Posted Date: May 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9462882/v1

Additional Declarations: No competing interests reported.

Statistical Assessment of the Laboratory Performance Achieved in Charpy Impact and Tensile Testing Scenarios

Askarli Hasan[1], Csaba Hegedűs[1], Zsolt T. Kosztyán[1,2,3*]

[1]Department of Quantitative Methods, University of Pannonia, Egyetem str. 10, Veszprém, 8200, Hungary.
[2]Instiute of Advances Studies, Kőszeg (iASK), Chernel str. 14, VKőszeg, 9730, Hungary.
[3]Wekerrle International University, Jázmin str. 10, Budapest, 1083, Hungary.

*Corresponding author(s). E-mail(s): kosztyan.zsolt@gtk.uni-pannon.hu;
Contributing authors: haskarli@phd.gtk.uni-pannon.hu;
hegedus.csaba@gtk.uni-pannon.hu;

Abstract

In this study, measurement performance in mechanical testing of materials was examined by analyzing proficiency testing data within a statistical evaluation framework that was aligned with the requirements of related international standards. Six laboratories participated in two autonomous proficiency testing schemes that were performed under monitored and standardized test conditions, involving both absorbed impact energy and room-temperature tensile properties such as yield strength, ultimate tensile strength, and percentage elongation. Homogenized and stability-verified samples were used for the measurements to reduce the errors induced by material variations. Statistical tools and visualizations were used to reveal parameter-specific behavior differences, highlighting that some mechanical properties may not follow a Gaussian distribution. When supporting the effectiveness of proficiency testing programs in terms of laboratory competence, quantifying between-laboratory variability and identifying methodological factors that affect the reliability of measurements are important. A unified analytical framework that can support the improved harmonization of mechanical testing practices and strengthen the metrological and methodological reliability attained across different testing networks is also presented in this study.

Keywords: Charpy impact test, proficiency testing, interlaboratory comparison, tensile testing, laboratory performance evaluation, z score analysis, mechanical testing variability

1 Introduction

A main aspect of materials engineering, integrity assessments, and industrial-level quality assurance scenarios is the evaluation of mechanical parameters through standardized testing procedures. Monitoring the performance and competence of testing and calibration laboratories, as well as harmonizing measurement practices across different national and international networks, is essential for satisfying interlaboratory comparability and traceability requirements in an increasingly interconnected global manufacturing environment, (Panteghini and Krintus [1]). For this reason, proficiency testing schemes play an essential role in supporting laboratory-level traceability and comparability.

Mechanical engineering tests are sensitive to variations in environmental conditions, equipment calibration levels, the techniques and expertise of operators, and sample preparation procedures (Bouhouche et al. [2]). Therefore, to ensure the reproducibility and metrological reliability of mechanical testing methods, proficiency testing programs are conducted in accordance with ISO/IEC 17043:2023 [3] and ISO 13528:2022 [4]. In Charpy impact and tensile testing cases, critical parameters such as the absorbed impact energy, yield strength, ultimate tensile strength, and percentage elongation are used to characterize metallic materials and support industrial decision-making processes in material selection scenarios, compliance with regulatory requirements, and product certification tasks.

During such measurements, the performance achieved in routine use cases may vary significantly between different laboratories, partly because of the sensitivity of material deformation mechanisms and partly because of the inconsistencies that can arise even under normal conditions.

According to recent proficiency testing activities conducted regarding the Charpy impact test based on ISO 148-1:2016 [5] and metallic tensile testing according to ISO 6892-1:2019 [6], evaluating frameworks that combine homogeneity and stability assessment, assigned value determination, and performance evaluation procedures is necessary. The homogeneity and stability criteria prescribed by ISO 13528:2022 [4] can be evaluated in accordance with the target distribution, ensuring that variations in the reported impact energies result from laboratory performance and not material inconsistencies.

In tensile proficiency testing programs, where laboratories perform replicated measurement processes such as determining yield strength, tensile strength, and elongation values, a standardized framework can be applied using predefined plate specimens under controlled conditions. The imposed homogeneity and stability requirements can be fulfilled, and assigned values can be derived by implementing bias-corrected formulations that are aligned with ISO 13528:2022 [4] procedures.

The existing studies, Coucke and Rida Soumali [7], Visser [8], and Ehrmeyer and Laessig [9], have highlighted significant limitations, as most proficiency reports include only single-number scoring schemes (z scores) and provide limited report-level benchmarking capabilities or poor statistical diagnostics for implementing performance assessments.

Additionally, some existing studies have compared the use of basic scoring and highly advanced statistical estimators. Tsamatsoulis and Cofino et al. have discussed the importance of selecting an appropriate proficiency testing evaluation method, noting that variations may not be properly benchmarked unless appropriate statistical tools are applied.

In long-term technical assessments accompanying proficiency testing datasets, the assumptions of normality are not always satisfied for some mechanical parameters, such as the yield strength, which deviates from Gaussian behavior.

Proficiency test reports are limited in their ability to comprehensively benchmark interlaboratory performance, as they poorly utilize statistical assessment tools to detect performance variations.

On the basis of this gap, the following research aim (RA) can be stated.

RA - The aim of this research is to establish a unified framework that enhances both the analytical and practical relevance of proficiency programs in mechanical tests while supporting measurement harmonization, accreditation, and metallic characterization consistency.

To achieve this goal, analytical and statistical assessments are used to identify the possibility of methodical and technical variations between the results of laboratories, a one-way ANOVA, the Kruskal–Wallis, normality testing, and R&R with graphical analytics for supporting the Z scores and performance results of laboratories.

RQ - Which statistical tools or comprehensive interlaboratory performance benchmarking methods can be used to detect laboratory performance variations in terms of proficiency testing results for mechanical testing purposes?

An underexplored methodological gap is that proficiency testing programs rely on parametric statistics for performance evaluation tasks, while assessments of the suitability of assumptions for mechanical test data remain limited, especially when specimen-related heterogeneity and deformation mechanism instability may lead to multimodal or skewed distributions. Enhanced strategies and systematic tools are needed to better detect outliers and systematic errors (Bisson et al. [12]). The broader scientific discourse lacks a comprehensive analysis of how statistical deviations should affect proficiency testing designs, assigned value calculations, decision rules, and reliability metrics. Therefore, a complementary nonparametric evaluation is needed to verify the robustness of interlaboratory comparisons.

These considerations require a measurement science-oriented examination of the proficiency testing processes that are applied to Charpy impact and tensile testing scenarios. The objective of this study is to synthesize experimentally derived proficiency testing data using statistical evaluation methods to identify between-laboratory variability, assess the suitability of the conventional and nonparametric statistical tools

for evaluating interlaboratory differences, and detect the possibility of procedural or environmental factors that may systematically affect deviations. Integrated evidence is derived from proficiency schemes, where the homogeneity, stability, and distribution logistics of the samples are fully checked and monitored.

In the proposed framework, comprehensive normality tests, an ANOVA, a Kruskal–Wallis analysis, and distribution visualization techniques are used to address unresolved questions about the reliability, comparability, and interpretability of mechanical test results concerning different laboratory functions.

As the main contribution of this research, the services of proficiency testing providers can be improved by providing comprehensive benchmarking capabilities in accordance with the application of analytical and statistical assessments.

2 Literature Review

Proficiency testing is an internationally recognized and accepted key element of conformity assessment and laboratory quality assurance processes. ISO/IEC 17043:2023 [3], as an international standard, sets requirements for how proficiency testing schemes are designed, operated, and reported. It formally defines proficiency testing as an "evaluation of participant performance against pre-established criteria by means of interlaboratory comparisons".

Alper noted the importance of conducting proficiency testing for verifying the accuracy of measurements, demonstrating the competence of laboratories, and highlighting areas that need improvement by evaluating the performance of laboratories using a comparison between different test procedures under the same methodology. The updated version of ISO/IEC 17043 is aligned with the ISO/IEC 17025:2017 and ISO 13528:2022 standards, providing more accurate and trustworthy proficiency testing approaches. Interlaboratory comparisons are essential for providing objective and strong evidence of the' performance achieved by laboratories under controlled environments, and this has also been supported by Ilinca et al..

In proficiency testing cases, the main focus is on benchmarking measurement performance and ensuring that laboratories obtain mutually comparable results. Vander Heyden and Smeyers-Verbeke [15] provided a general definition, stating that interlaboratory studies, which are sometimes also called collaborative studies or ring tests, are studies in which several laboratories analyze the same material(s). However, even when laboratories use unified standard methods, strict testing requirements are necessary to reduce the induced variability; Arrhenius et al. agreed with this idea. Sipkens et al. reported that one of the main challenges is variability, which is difficult to explain, and to increase the consistency of the output results, standard verification and calibration procedures are needed. Differences between the performances of various laboratories are related to their equipment calibration settings and operator performance; in this regard, Cherie et al. mentioned that regular proficiency testing evaluations can be used for the identification of analytical errors.

Proficiency testing outcomes depend on various factors, such as the competence of proficiency testing providers and controlled scheme monitoring. ISO/IEC 17043:2023

[3] "specified general requirements for the competence of providers of proficiency testing schemes and for the development and operation of proficiency testing schemes". This work explained the responsibilities, requirements, and demanding monitoring processes that are associated with proficiency testing when performed in accordance with internationally accepted rules. In accordance with this framework of proficiency testing, different studies have had distinct purposes: "Laboratory-performance or proficiency studies focus on the laboratory with the aim of assessing the proficiency of the individual laboratories."

Additionally, performance evaluation and reporting constitute key stages of proficiency testing, and the imposed standard demands that results be clearly and accurately visualized in a way that allows participant laboratories to understand their performance. ISO/IEC 17043:2023 [3] also stated that "Proficiency test reports shall be clear and comprehensive and include data covering the results of all participants, together with an indication of the performance of individual participants".

Methodological approaches for proficiency testing can be divided into two perspectives: general and application-specific techniques. In the analytical context, proficiency testing schemes are not only tools for accreditation but also mechanisms for identifying systematic errors, improving methods, and strengthening confidence in reported data. While discussing method selection, importantly, between-laboratory reproducibility strongly depends on the implemented procedures and methodology. It has also been mentioned by Vasselon et al. that different methodologies can result in the noncomparability of laboratory results.

Finally, individual Z scores can be calculated to indicate the performance of each tested laboratory. The Z score is related to the standard deviation of the examined proficiency testing assessment and the corresponding assigned value. Tsamatsoulis also supported the notion that it is important to use robust statistical methods or appropriate methods to detect statistical outliers. Z scores are widely used in proficiency testing scenarios, where results are accepted as satisfactory when they fall within predefined limits (Cavalli et al. [20]).

Furthermore, in terms of continuous improvements, corrective actions must be taken after conducting proficiency testing. Proficiency testing is intended to provide evidence that supports corrective actions but does not mandate them by itself.

In terms of achieving technical validity, comparability, and fairness across different proficiency testing participation conditions, proficiency testing schemes are not arbitrary but rather are derived from standardized principles. Instead of isolated accuracy, the main aim is to assess how laboratories perform under routine conditions, which leads to the fundamentally performance-oriented nature of proficiency testing. This involves the use of interlaboratory comparisons to determine the performance of participants (which may be laboratories, inspection bodies, or individuals) in specific tests or measurements and to monitor their continuing performance.

In mechanical tests of materials, the measured properties depend on both their underlying microstructures and the engineering characteristics of the test configuration, which is why the mechanical material tests are inherently variable. As a starting point, Beckert et al. emphasized the centrality of such tests for engineering decisions,

noting that "Mechanical tests are common in industrial activity" and that "The specification of mechanical properties is essential for the adequate choice of material, as well as for the design and manufacturing of components and products." Following this, ISO 148-1:2016 [5] provided a predefined sample size for the unit under test to obtain highly accurate measurement results. In laboratory performance evaluations, consensus values derived from participant measurements are used (Possolo [22]).

The results of mechanical tests, such as fracture and impact behavior tests, are strongly influenced by the underlying material conditions and loading–storage configurations. In terms of between-test and between-laboratory dispersion effects, small differences are present between the responses of samples to environmental conditions, especially temperature, even when standards are followed. These kinds of influences cause slight changes in the behaviors of material properties.

Additionally, the size and geometry of the unit under test influence the mechanical testing process, leading to variability. Tests such as yield strength, ultimate tensile strength, and elongation tests are affected by variations. Beckert et al. concisely summarized this multifactor nature, stating that "The results of the tensile test are influenced by factors related to the material, specimen, test equipment, test procedure, and calculation of mechanical properties."

In tensile testing and proficiency testing schemes, destructive testing, machining costs, and the limited availability of materials constrain sample sizes to a few replicates per test condition.

The present study follows a strategy involving reproducibility, homogeneity, and stability checks, as well as a comparative evaluation of Charpy impact and tensile tests.

3 Background

3.1 Background for the Charpy Impact Test

The test temperature is specified as an interval of 23 °C ± 5 °C, which was noted as the optimal performance temperature in ISO 148-1:2016 [5]. If the temperature is specified by the examined laboratories, it may be varied within ± 2 °C to appropriately narrow the interval.

The assigned measurement interval is between 21 °C and 25 °C (within ±2 °C); however, the best practice suggests using a value of approximately 20 °C as the test temperature for obtaining optimal results. Chao et al. reported that the Charpy impact test is sensitive to temperature, especially in the ductile-to-brittle region, where a reduction in the absorbed energy and an increase in brittleness are strongly related to a decrease in temperature. The samples are maintained at an average temperature of 20 °C, which represents the standard environmental conditions for calibration measurements. The standard provided by ISO 148-1:2016 [5] explicitly highlights the fact that "Because the impact values of many metallic materials vary with temperature, tests shall be carried out at a specified temperature." To evaluate the stability of the test, the samples are stored under these conditions for a duration of three hours. For the initial approach, the interval should be within the limits of 18 °C and 28 °C; however, the risks associated with a wide temperature range must be considered. The sample proportions are 55*10*10 mm, as shown in Figure 1, which are taken from the

correct sample size table presenting in ISO 148-1:2016 [5]. Charpy testing must be carried out using a standardized unit under test with dimensions of 55 × 10 × 10 mm; this requirement was posed by ISO 148-1:2016 [5], which states that "The standard test piece shall be 55 mm long and of square section, with 10 mm sides."

Sadowski et al. mentioned the importance of material selection; consistent and reproducible performance comparisons between laboratories are strongly linked to appropriate measurement material selection processes.

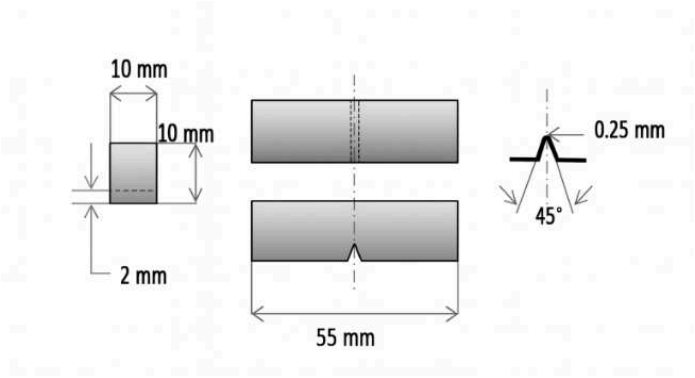

Fig. 1 Illustration of the samples

In practice, it is suitable to perform measurements with a set of three samples under the same environmental conditions (using the same equipment and operator). It is important to report the average absorbed and calculated energy levels; however, ISO 148-1:2016 [5] does not mandate testing three specimens in three continuous measurements.

For this proficiency test, particularly with respect to specimen handling, Model 2 (Simultaneous) is adopted in accordance with the ISO/IEC 17043:2023 [3] - Conformity Assessment: General Requirements for Proficiency Testing standard.

3.2 Background for the Tensile Test

Tensile tests are conducted to determine the elastic and plastic behaviors (mechanical properties) of materials under static loading to classify them according to their mechanical behaviors and to assist in material selection tasks. In this test, the strength values of standard tensile samples are measured. The preparation, qualification, dimensions, and tolerances of the test samples are determined in accordance with ISO 6892-1:2019 [6] Metallic materials - Tensile testing - Part 1: Method of test at room temperature standard.

A plate sample with a nominal thickness of 5 mm is selected as the tensile test sample. The samples are kept at room temperature and then subjected to testing. Each participant is asked to perform three measurements, and the average of the three measurements is taken as the final result.

With respect to the distribution of the specimens, Model 2 (Simultaneous) is chosen in accordance with the ISO/IEC 17043:2023 [3] Conformity assessment - General requirements for proficiency testing standard. In Model 2, all laboratories simultaneously receive samples within a limited time frame for performance testing. An equal test condition rule is also applied in this test method.

The processes of selecting, preparing, packaging, storing, and distributing the specimens, as well as determining the standard deviation of the program, performing homogeneity and stability checks, defining the assigned value and measurement uncertainty level, preparing result recording forms, establishing communication with the participants and ensuring data confidentiality, reporting and delivering the results to the participants, and handling objections, are carried out in accordance with the ISO/IEC 17043:2023 [3] Conformity assessment - General requirements for proficiency testing standard.

This test method is designed for conducting tensile testing on metallic materials and is based on tensile tests of metallic samples. The test is used to determine the elastic and plastic behaviors of the samples, which are classified according to their mechanical reactions, and the correct material is selected. The Key material properties are recorded according to the test requirements:

- Yield strength
- Ultimate tensile strength
- Percentage elongation after a fracture

All measurements are obtaining in accordance with the requirements of ISO 6892-1:2019 [6], which govern the creation of test samples, the test point selection process, the setting of tolerance limits, and the determination of the test temperature. The test temperature must be within the room temperature range, i.e., between 10 °C and 35 °C, unless otherwise specified. With respect to the test conditions, ISO 6892-1:2019 [6] defined the temperature interval within which tensile tests should be performed, stating that "The test shall be carried out at room temperature between 10 °C and 35 °C, unless otherwise specified". The tests are carried out under controlled conditions, which are given as an interval of 23 °C ± 5 °C.

The standard deviation of the program is selected from the 2022 proficiency testing conditions:

- Rp 0.2 σ_{PT}: 7.28
- Rm σ_{PT}: 8.2
- An 80 σ_{PT}: 2.63

Standard deviations from previous programs can be used, while ISO 13528:2022 [4] highlights that there is no fixed method but rather several methods for the calculation of the standard deviation of a program, including using experience with previous rounds of a proficiency testing scheme for the same measurement.

4 Methodology

4.1 Homogeneity and Stability in Charpy Impact Testing

A proficiency test is performed among six calibration and testing laboratories using test methods in accordance with ISO 148-1:2016 [5].

During proficiency testing, homogeneity signifies that all test samples distributed to the participating laboratories possess uniform properties, ensuring that any observed differences among their results arise solely from analytical performance rather than sample variability. Stability denotes the capacity of the test material to maintain its characteristics over a defined period and under specified storage and transportation conditions up to the point at which they are analyzed by the participants. The homogeneity and stability of the unit under test must involve a participant evaluation, which depends on the preparation and statistical treatment of test items that rely directly on the requirements of ISO 13528:2022 [4]: "The proficiency testing provider shall ensure that batches of proficiency test items are sufficiently homogeneous and stable for the purposes of the proficiency testing scheme."

The application of homogeneity and stability checks in proficiency testing reports occurs prior to conducting performance scoring. For Charpy impact testing, considering both homogeneity and stability is essential because the statistical reliability and material variability of a set of three measurements are high according to the 3 test samples; therefore, it is crucial to apply homogeneity and stability tests. The stability test can be expressed as follows:

$$S_{SD} = 0.3\,\sigma_{PT} \tag{1}$$

where S_{SD} denotes the standard deviation between groups of samples and σ_{PT} denotes the standard deviation of the proficiency testing (PT) program.

Twenty samples are chosen for homogeneity testing and are divided into two groups. The first group, consisting of ten samples, is used for the homogeneity test, while the remaining samples are used for the stability and repeatability tests. The homogeneity test can be expressed as follows:

$$|y_1 - y_2| \leq 0.3\sigma_{PT} \tag{2}$$

where y_1 is the average homogeneity test result and y_2 denotes the average stability test value.

The repeatability test has several requirements, such as performing measurements under the same environmental conditions, using the same equipment, and having the same operator conduct the tests multiple times (ideally, 10 tests are recommended). To calculate the assigned value for proficiency testing purposes, it is important to use the following equation:

$$x_{\mathrm{PT}} = x_{\mathrm{Homogeneity}} + (x_{\mathrm{Homogeneity}} \times x_{\mathrm{Bias\%}}) \tag{3}$$

where x_{PT} is the known value of the chosen sample, $x_{\mathrm{Homogeneity}}$ is the result of the homogeneity test and $x_{\mathrm{Bias\%}}$ is the deviation from the reference sample.

The assigned value is $x_{\mathrm{PT}} = 69.1$ J. This value represents the reference for comparing the results of different participants, as grounded in the definitions and procedures

of ISO 13528:2022 [4]: "assigned value: value attributed to a particular property of a proficiency test item." Its determination must be appropriate for the scheme design and objectives. Importantly, the assigned value is one of the critical parameters in proficiency testing and needs to be determined with higher accuracy and metrological reliability (Pommé and Spasova [25]). This value is used to enhance the accuracy of the conducted performance assessment.

The sample is used in proficiency testing, with its certified value represented as an interval of 70.0 J ± 7.6 J. For uncertainty calculation purposes, Equation (4) can be expressed as follows:

$$u_{\mathrm{PT}} = \sqrt{\mathrm{Bias}^2 + u^2_{\mathrm{Homogeneity}} + u^2_{\mathrm{CRM}}} \tag{4}$$

where u_{PT} is the uncertainty of the proficiency testing sample, Bias is the standard deviation of the reference sample, u_{CRM} is the standard uncertainty of the reference sample and $u_{\mathrm{Homogeneity}}$ is the uncertainty derived from the homogeneity test. The uncertainty determined from the homogeneity test $u_{\mathrm{Homogeneity}}$ can alternatively be represented as follows:

$$u_{\mathrm{Homogeneity}} = \frac{SD}{\sqrt{n}} \tag{5}$$

where SD is the standard deviation of the homogeneity test results and n is the number of measurements.

4.2 Homogeneity and Stability in Tensile Testing

Prior to the distributing the samples to the participants and using them throughout the entire evaluation process, the "ISO 13528:2022 [4]: Statistical Methods Used in Proficiency Testing by Interlaboratory Comparisons standard", referenced in "ISO/IEC 17043:2023 [3]: Conformity Assessment - General Requirements for Proficiency Testing", is applied, beginning with homogeneity and stability assessments. The homogeneity evaluation is conducted in accordance with the equation given below:

$$S \leq 0.3\,\sigma \tag{6}$$

where S represents the standard deviation between groups and σ denotes the standard deviation of the program.

A total of 20 test samples are randomly selected from the prepared batch and divided into two groups. These samples are tested under repeatable conditions. Since the results satisfy the criterion expressed in Equation 1, the samples are confirmed to be homogeneous.

Following the receipt of the test results from the participants, stability tests are performed. For this purpose, 10 remaining samples are selected and tested under the same repeatability conditions. The stability assessment is carried out using the equation below:

$$|y_1 - y_2| \leq 0.3\,\sigma \tag{7}$$

where y_1 is the mean value obtained from the homogeneity tests and y_2 is the mean value obtained from the stability tests.

When the results are satisfactory, the samples are deemed to exhibit adequate stability.

5 Results and Discussion

5.1 Charpy Impact Test

Table 1 describes proficiency testing results obtained for six laboratories conducting Charpy impact testing, where the laboratories handle their reports related to the energy absorption values measured from replicated tests. Owing to confidentiality requirements, all laboratories are named by unique codes. Charpy impact tests are performed for each sample set, the recorded values are expressed in joules (J), and the average result obtained for each laboratory is subsequently calculated. According to the previous proficiency testing program designed in 2023, the standard deviation of the program is as follows:

$$\sigma_{\mathrm{PT}} = 5 \text{ J} \tag{8}$$

Performance evaluations of proficiency tests rely on standardized statistics such as z scores: "z score: a standardized measure of performance, calculated using the participant result, assigned value, and the standard deviation for proficiency assessment." A statistical analysis of the Z scores produced during proficiency testing can be represented as follows:

$$Z = \frac{X - X_{\mathrm{PT}}}{\sqrt{\sigma_{\mathrm{PT}}^2}} \tag{9}$$

where X is the result of an individual laboratory, and x_{PT} is the known value of the chosen sample.

In accordance with the performance assessments, the reported average values are compared with the estimated high and low limits, which are 62.4 J and 77.6 J, respectively. When these high and lower limits are considered, the acceptance performance range is established, within which the results of all laboratories are expected to fall. The computed Z scores of each laboratory are used to identify how many standard deviations the reported average lies from the assigned consensus mean.

ISO 13528:2022 [4] notes that most proficiency testing schemes evaluate whether 'the deviation of a laboratory from the reference value represents a cause for concern: "most schemes compare the participant"s deviation from an assigned value with a numerical criterion which is used to decide whether or not the deviation represents cause for concern." This directly supports the use of interpretation limits in proficiency testing reports, where $|z| \leq 2$ is generally considered satisfactory and reflects strong agreement with the consensus value (being closer to zero), while values exceeding ± 2 are considered unsatisfactory performance.

In this case, according to the degree of agreement with the consensus value, all laboratories yield average results that are within the acceptance interval. For example, laboratory L-350441 obtains an average of 70.3 J with a z score of 0.24, indicating

close alignment with the consensus mean. In contrast, laboratory L-514201 records a lower average of 64.7 J, with a z score of -0.88, indicating a slight negative deviation, although it is still within the acceptable range. The highest degree of agreement with the upper limit is that of laboratory L-452961, which results in an average of 76.8 J and a z score of 1.54, which is high but still within the acceptable range. Overall, the table demonstrates how individual laboratory results are quantitatively assessed for consistency and accuracy within an interlaboratory proficiency testing framework. The Charpy impact test results produced by the participating laboratories are shown in Table 1.

Table 1 Charpy impact test results of the participating laboratories

S/N	Lab Code	1st Test (J)	2nd Test (J)	3rd Test (J)	Low Limit	Upper Limit	Avg. Results (J)	Z–score (J)
1	L–350441	70.6	73.2	67.1	62.4 J	77.6 J	70.3	0.24
2	L–110324	68.9	67.5	62.1	62.4 J	77.6 J	66.2	- 0.58
3	L–611434	71.4	72.0	72.6	62.4 J	77.6 J	72.0	0.58
4	L–452961	76.4	76.9	77.0	62.4 J	77.6 J	76.8	1.54
5	L–514201	60.0	66.0	68.0	62.4 J	77.6 J	64.7	- 0.88
6	L–978104	74.1	75.0	75.9	62.4 J	77.6 J	75.0	1.18

A graphical representation of the results of the proficiency test clearly illustrates the variability exhibited in both cases; the results obtained within and among the six laboratories are shown in Figure 2.

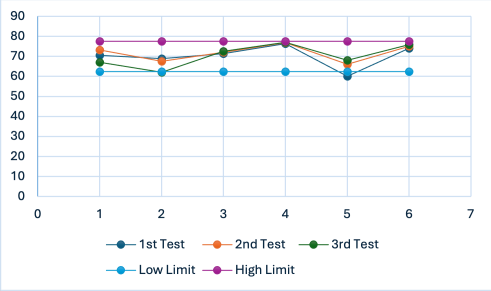

Fig. 2 Control chart showing the laboratory-averaged absorbed energy values relative to predefined acceptance limits

It is obvious from the chart that the results of all laboratories align closely with each other, demonstrating acceptable repeatability for individual laboratories. The positions of the data points relative to the acceptance boundaries also show that all laboratories have reported values within the designated limits, with no instance of performance falling outside the required range.

The chart describes the use of control limits as a benchmark for monitoring the performance of laboratories. The results that are closer to the center of the acceptance band more strongly agree with the consensus value, whereas those that are near the lower and higher limits indicate a tendency toward systematic deviation, which is

still acceptable. The chart supports an objective assessment of interlaboratory comparability, highlighting both the robustness of the process and the reliability of the measurements.

Starting from scratch, the Shapiro–Wilk normality test is applied to determine whether the residuals of the given dataset follow a normal distribution. Its W value is 0.93908, and its p value is 0.2794.

There is no evidence to reject the null hypothesis of normality in this case, as the p value is greater than the established threshold of 0.05. That is, the residuals are not significantly different from the normal distribution, and the assumption of normality to which the ANOVA refers is supported. Next, a one-way ANOVA is conducted, and the factor "ID" shows a statistically significant interaction in the table, with a F value of 10.04 and a p value of 0.000576. The spread of the residuals is very small compared with the between-group variance, supporting the conclusion of variance between the groups. Overall, the assumptions of the ANOVA seem to be satisfied, as the normality test does not suggest that any violations occur. This adds weight to the statement that the observed differences among the laboratories are statistically trustworthy. The ANOVA results of the Charpy impact test are listed in Table 2.

Table 2 One-way ANOVA results obtained for the Charpy impact test data

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	342,045	5	68,409	10,04291656	0,000575926	3,105875239
Within Groups	81,74	12	6,811666667			
Total	423,785	17				

A graphical approach for determining whether the residuals derived from testing data follow a normal distribution, which is one of the assumptions required for the application of an ANOVA, is shown in Figure 3. In the Q–Q plot, the sample quantiles are plotted against the corresponding theoretical quantiles determined from a normal distribution. When the residuals are normally distributed, the points should fall approximately along the reference line, with only minor outliers expected at the extremes. The distribution of the plotted points illustrates a linear pattern through the center, supporting the notion that the data align well with the assumption of normality.

Overall, the Q–Q plot complements the statistical test by visually confirming that the residuals of the proficiency testing dataset follow a normal distribution and support the reliability of the ANOVA.

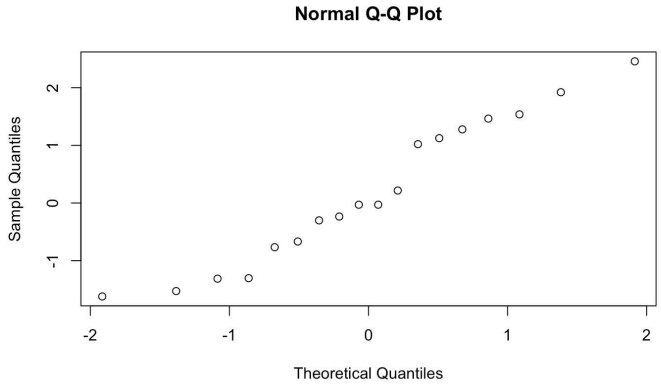

Fig. 3 Normal Q–Q plot for assessing the normality of the Charpy impact test results

The distribution of the proficiency test results among the six laboratories is shown in Figure 4. Each box represents the spread of the three replicated measurements produced for each laboratory. The central line represents the median value and the interquartile ranges, and the whiskers reflect the minimum and maximum results.

Overall, the plot provides a clear graphical summary of both the within-laboratory repeatability and between-laboratory comparability effects, reinforcing the conclusion that while all the results remain within the defined acceptance limits, systematic performance differences are evident among laboratories.

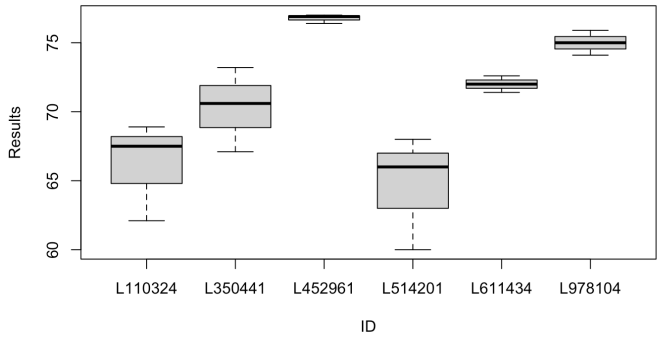

Fig. 4 Box plot illustrating the distribution of the Charpy impact test results across laboratories

As a result of the R&R testing process, the differences among the tests conducted within the same laboratory are small relative to the total variation, which demonstrates good operator reproducibility within the same laboratory in the Charpy impact test. However, laboratory-to-laboratory differences represent the largest portion of the

total variation. The mean values of each laboratory differ noticeably across various laboratories, resulting in poor interlaboratory reproducibility. Laboratories L-452961 and L-978104 show very low internal variations, whereas L-514201 and L-110324 exhibit the greatest internal variations, indicating weaker repeatability.

5.2 Tensile Test

5.2.1 Yield Strength

The data obtained from the six laboratory samples represent the proof stress of the tested material, which is the stress threshold at which permanent plastic deformation begins. For the acquisition of precise measurements, three replicates are performed in each laboratory, and these replicates are averaged before statistical evaluation is performed. It is determined that the average yield strength of 319.6 MPa indicates a central tendency of the results and that the standard deviation quantifies the differences between them. The low standard deviation observed in the σ_{PT} value of 7.28 MPa suggests that the results are uniform among all the participating laboratories and that the material exhibits stable mechanical behavior.

The Z scores calculated for each laboratory represent the standard deviation units between single average results and the overall mean. Positive Z scores indicate higher yield strength values than the mean, whereas negative Z scores indicate lower yields. A Z score near zero signifies close conformity with the overall average value. The majority of the Z scores in the samples are within acceptable ranges and indicate that the testing process is tightly controlled.

In contrast, one sample, Lab L-978104, whose average yield strength is 349.3 MPa, is anomalously high, which reveals an outlier, indicating a sharp deviation from the average over all the samples. Process tolerance limits are provided in the form of lower process tolerance (LPT) and upper process tolerance (UPT) values as guidelines for the acceptable limits of the desired value (Xpt). Since the target of 312.4 MPa is given, the calculated limits are 296.8 MPa and 343.5 MPa, respectively. Measurements within this range are considered compatible with the defined performance limits. All but one laboratory has yield strength results that are within the defined limits, indicating adequate control over the quality of the material and the testing procedure, according to the findings. Only one outlier significantly exceeding the UPT may be a result sample preparation, laboratory procedure, or material variability differences.

The yield strength measurements obtained in all the laboratories are generally homogeneous and show little or no variability with respect to the mean. The low σ_{PT} value indicates the repeatability and reliability of the proof stress determination procedure; the tolerance analysis confirms that the testing process remains within the limits of acceptable quality control.

A statistical analysis of this nature is key to verifying mechanical properties, providing product quality assurances and reinforcing the validity determined when evaluating interlaboratory results that are consistent with product performance. The yield strength results derived from the tensile proficiency testing procedure are listed in Table 3.

Table 3 Yield strength results derived from tensile proficiency testing

S/N	Lab Code	1st Test (MPa)	2nd Test (MPa)	3rd Test (MPa)	Avg. Results (MPa)	Z – score (MPa)
1	L – 350441	316.2	324.0	319.3	319.9	1.03
2	L – 110324	318.7	328.6	339.8	329.0	2.28
3	L – 611434	310.9	312.0	318.5	313.8	0.19
4	L – 452961	300.8	302.6	298.0	300.5	- 1.63
5	L – 514201	307.9	305.8	301.6	305.1	- 1.00
6	L – 978104	290.0	363.0	395.0	349.3	5.06

As a result of the Shapiro–Wilk test, the p value is less than 0.05, which indicates that the measurement results are not normally distributed. This suggests that there may be small systematic influences, such as calibration, measurement, test speed, environmental condition, or operator technique differences. Because the data do not follow a normal distribution, parametric results such as the one-way ANOVA results should be interpreted with caution. The one-way ANOVA results obtained for the yield strength data are shown in Table 4.

Table 4 One-way ANOVA results for yield strength data

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	4749,262778	5	949,8525556	1,865056506	0,174539713	3,105875239
Within Groups	6111,466667	12	509,2888889			
Total	10860,72944	17				

As a result of the ANOVA, the output has a p value of 0.175. Assuming normality, this indicates that there are no significant differences among the mean proof-stress results of the laboratories. However, given that the normality condition is violated, a nonparametric alternative is more appropriate.

Additionally, the results of the Kruskal–Wallis test show a p value of 0.090, which is still above the common significance threshold of 0.05. This finding also indicates that there are no statistically significant differences among the laboratories. The p value being close to 0.05 suggests that minor variations exist, and with more data, these variations might become clearer; however, at the current level, no strong evidence of bias is observed.

All laboratories are performing consistently. The observed differences are due to the natural uncertainty of mechanical testing and do not indicate that any laboratory is producing significantly higher or lower results. Therefore, no laboratory requires corrective action on the basis of these results, although routine good practices such as calibration checks and continued operator training procedures should be maintained. The data in Figure 5 visually confirm the findings of the Shapiro–Wilk test, i.e., that these data are not normally distributed.

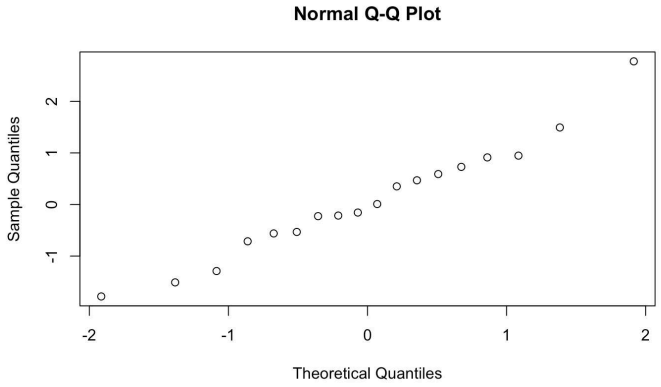

Fig. 5 Normal Q–Q plot for evaluating the distribution of the yield strength results

Assuming that the data of one sample are normally distributed, one can see that the points are in close proximity of a straight diagonal line. The points in this plot move visibly away from linearity, especially at the lower and upper ends of the distribution. Such curvature indicates that the distribution is skewed; some data are more extreme than expected and there is no normal distribution.

The fact that the points are not aligned with the straight-line trend only suggests that the normality hypothesis breaks down. The variance of the tensile proof-stress results comes from subtle measurement biases or specimen (process), machine stiffness, environment or operator action differences.

Hence, normality-based statistical methods such as the ANOVA must be approached with care, and a nonparametric alternative such as the Kruskal–Wallis test is more suitable. The distribution of the proof-stress results produced by each laboratory, including their values, is shown in Figure 6.

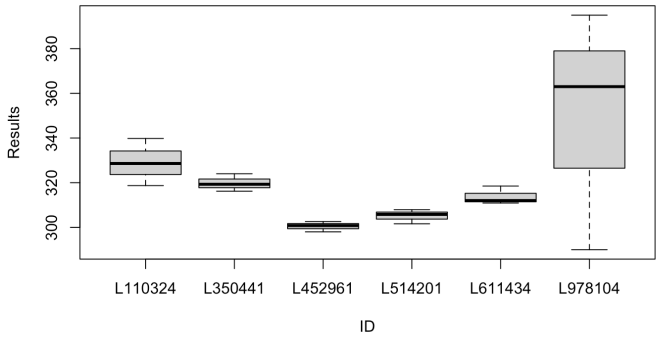

Fig. 6 Box plot showing the interlaboratory variability exhibited by the yield strength measurements

Most laboratories are clustered closely together with small interquartile ranges, indicating good test repeatability within each laboratory. The results of laboratories L-452961, L-514201, and L-611434 are grouped together, which implies that they have consistent conditions or controlled variables. More spreads are observed for laboratories L-110324 and L-350441 than in the other laboratories but are also within an acceptable mechanical testing variability range.

In contrast, laboratory L-978104 has a significantly broader distribution and has higher values than the other laboratories do. This could arise from varying the calibration of the tools, the sample preparation process, the strain rate, or the proof-stress offset. These differences, which are visible, are not conclusive indications that any laboratory is performing incorrectly; they are merely as expected with respect to tensile testing, and they are normal variations of that standard.

5.2.2 Ultimate Tensile Strength

The ultimate tensile strength describes the maximum stress that a material can withstand under tension before reaching its breaking point. All laboratories perform measurements three times on similar samples, and their averages reflect the consistency of the maximum strength determination.

The averages of each laboratory provide insight into the overall level of the tested series, with an average tensile strength of 417.9 MPa. The standard deviation is 6.5 MPa, which indicates that the results are spread around the mean value, indicating broadly comparable testing conditions across the different laboratories. The range of all Z scores is between -1.23 and +0.67, which is within the acceptance interval range of ±2. This variation may be attributed to the nature of the measurements, systematic effects, and differences in the measurement conditions. Overall, all the laboratories perform the tensile strength test appropriately, and the samples possess stable performance features that are suitable for the desired purposes. The ultimate tensile strength results derived from the tensile proficiency testing process are listed in Table 5.

Table 5 Ultimate tensile strength results derived from tensile proficiency testing

S/N	Lab Code	1st Test (MPa)	2nd Test (MPa)	3rd Test (MPa)	Avg. Results (MPa)	Z – score (MPa)
1	L – 350441	420.7	431.4	425.5	425.9	0.67
2	L – 110324	413.8	422.4	425.8	420.7	0.04
3	L – 611434	421.7	423.8	420.6	422.0	0.19
4	L – 452961	410.6	414.1	409.4	411.4	- 1.10
5	L – 514201	409.2	410.2	411.5	410.3	- 1.23
6	L – 978104	417.2	415.3	418.2	416.9	- 0.43

According to the result of a Shapiro–Wilk normality test, the p value is greater than 0.05, which means that the data follow a normal distribution. Therefore, parametric methods can be applied to compare the different laboratories.

Furthermore, for the one-way ANOVA test, the p value is less than 0.05, indicating statistically significant differences among the laboratory results.

Although all the laboratories follow the same standard, variations in their calibration processes, alignment schemes, or specimen preparation procedures can lead to differences in their results.

Additionally, the results of the Kruskal–Wallis test confirm that significant differences exist between the laboratories. The results of the laboratories differ beyond pure random variations even when normality assumptions are ignored and ranks are compared instead of means. The results of the one-way ANOVA conducted for the ultimate tensile strength data are shown in Table 6.

Table 6 One-way ANOVA results produced for the ultimate tensile strength data

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	568,9177778	5	113,7835556	8,632682824	0,001140649	3,105875239
Within Groups	158,1666667	12	13,18055556			
Total	727,0844444	17				

The normal Q–Q plot displays the sample quantiles of the tensile strength results against the theoretical quantiles of a normal distribution.

The data generally conform to a normal distribution. In this case, Figure 7 visually supports the results obtained from the Shapiro–Wilk test, in which the data are approximately normally distributed and suitable for the application of more parametric statistical methods.

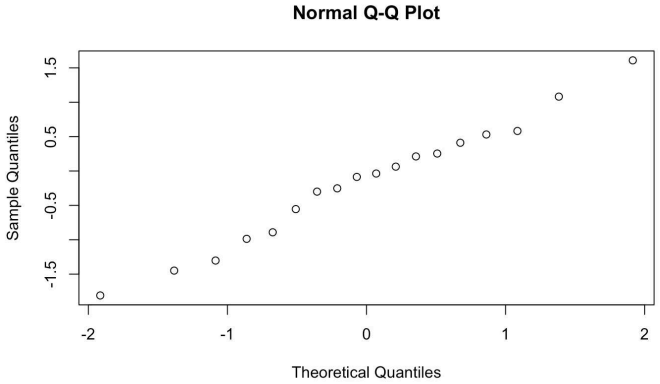

Fig. 7 Normal Q–Q plot of the ultimate tensile strength results

While the visualization in Figure 8 provides clarity regarding the statistical measurements, interlaboratory variations are observed; however, all of the laboratories are performing within the technically acceptable limits in terms of standardized tensile testing conditions.

The small interquartile ranges with small whisker lengths produced for laboratories L-452961, L-514201, and L-611434 suggest good repeatability and controlled testing conditions. Laboratories L-110324 and L-350441 show greater quartile differences, indicating large variability between their measurements; however, they are within an acceptable tolerance range.

A noticeable upward shift is observed in the median value of laboratory L-978104 relative to those of the other groups, with a slightly broader distribution.

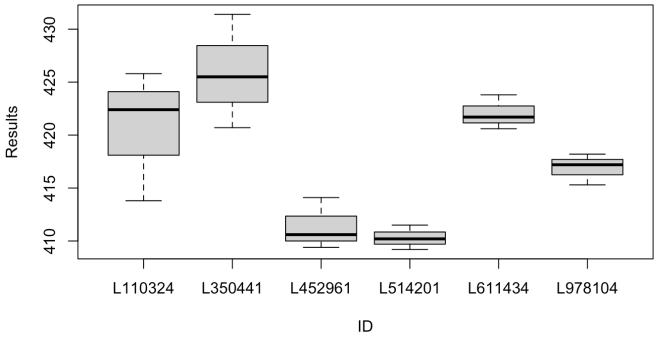

Fig. 8 Box plot illustrating the laboratory-wise distribution of the ultimate tensile strength results

5.2.3 Percentage Elongation After a Fracture

The percentage elongation exhibited after a fracture reflects the ductility of the tested metal, indicating how much deformation the metal can undergo before a failure occurs. All the laboratories conduct 3 tests under similar measurement conditions, and the average values of the measurements define how consistently each laboratory reports its material elongation performance results. The overall mean elongation of 32.6% represents the general ductility of the tested batches.

The standard deviation of 2.9% is greater than that of previous tests because of the sensitivity of the elongation measurement process to the specimen marking accuracy, gauge length formation, and fracture location; however, it remains within an acceptable range. The differences between the lower value of 26.2% and the higher value of 38.0% reflect ductility variability among the laboratories.

All the results fall within the acceptance interval of ± 2 relative to the Z score, which indicates satisfactory results without any major outliers. Laboratory L-350441 has the closest result to the assigned value ($Z = +0.99$), whereas laboratories L-514201 and L-110324 have the lowest elongation values and the greatest negative Z scores (-1.63 and -1.48, respectively) but remain within the acceptable range.

According to the measurements, the tested material exhibits a moderate level of ductility, which is important for forming processes and absorbing energy before a failure occurs. All measurements are within the acceptance interval for interlaboratory

mechanical testing and do not indicate any issues. The percentage elongation after fracture results determined from the tensile proficiency test are given in Table 7.

Table 7 Percentage elongation after fracture results determined from tensile proficiency testing

S/N	Lab Code	1st Test (%)	2nd Test (%)	3rd Test (%)	Avg. Results (%)	Z – score (%)
1	L–350441	38.0	35.5	37.1	36.9	0.99
2	L–110324	33.0	31.9	26.2	30.4	- 1.48
3	L–611434	32.1	31.3	35.4	32.9	- 0.53
4	L–452961	30.9	30.5	32.5	31.3	- 1.14
5	L–514201	30.5	29.3	30.1	30.0	- 1.63
6	L–978104	35.0	33.2	34.0	34.1	- 0.08

According to the results of the Shapiro–Wilk normality test, the p value is 0.9686, which is greater than 0.05, indicating that the data follow a normal distribution and that it is valid to use parametric comparison methods to compare the results.

The results of the ANOVA show that the p value is 0.00691, which is less than 0.05, indicating that there are significant differences among the measurement results. Although all the laboratories use the same standard and similar conditions when performing their measurements, the results vary. These differences may be attributed to specimen marking precision differences, surface finish inconsistencies, the localization of deformation near the fracture site, or extensometer positioning variations during the test.

The results of the Kruskal–Wallis test are similar to those of the ANOVA, with a p value of 0.02234, which is again less than 0.05. This means that even when the data are treated in a nonparametric manner, the differences between laboratories remain statistically meaningful. The one-way ANOVA-based percentage elongation after fracture results are listed in Table 8.

Table 8 One-way ANOVA-based percentage elongation after fracture results

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	102,2316667	5	20,44633333	5,587277972	0,006910393	3,105875239
Within Groups	43,91333333	12	3,659444444			
Total	146,145	17				

The data in Figure 9 indicate that the Shapiro normality test results are correct and that the percentage elongation data are normally distributed; thus, parametric analysis methods can be used for statistical comparison purposes.

It is obvious that toward the lower tail, there are slight deviations, where the smallest value falls below the line. However, these deviations are minor and do not suggest a significant departure from normality.

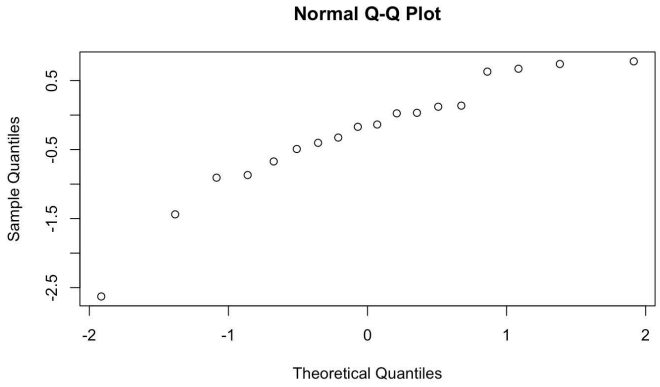

Fig. 9 Normal Q–Q plot for evaluating the normality of the elongation after fracture data

The variability of the measurement results obtained for each laboratory is shown in Figure 10. Laboratories L-452961, L-514201, and L-978104 produce narrow interquartile ranges and short whiskers, indicating high repeatability and consistency under the test conditions. Laboratory L-350441 provides good repeatable measurements while reporting higher values than those of the other laboratories.

Laboratory L-110324 exhibits the widest spread, with a noticeably lower minimum value, which indicates internal variability and suggests sensitivity to the main factors. Compared with the other laboratories, L-611434 has a higher range, with higher values.

Despite the interlaboratory variability observed in the box plots, all elongation values remain within the expected performance range for ductility measurements, which is consistent with the standardized tensile test procedures.

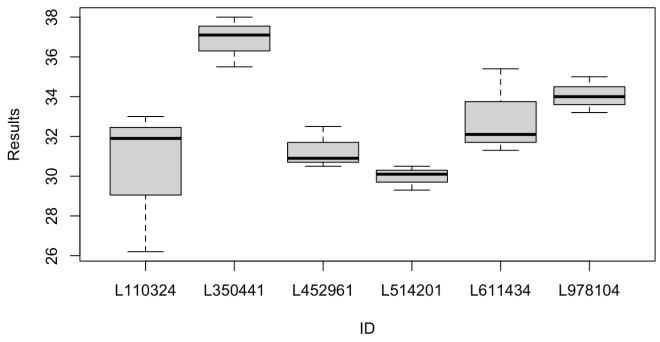

Fig. 10 Box plot showing the variability of the elongation after fracture measurements among the different laboratories

With respect to the R&R results of the tensile test, the measurements are acceptable regarding repeatability and operator-related effects; however, variations exist among the laboratories. Since the source of variation is neither operator- nor equipment-related, it is attributed to laboratory-related factors. Therefore, differences among the laboratory mean values indicate limited interlaboratory reproducibility, with laboratory effects representing the dominant contribution to the overall variation.

6 Conclusion

A proficiency testing program was successfully conducted, where six laboratories participated in an analytical and statistical performance comparison by conducting Charpy impact and tensile testing in accordance with international standards. Assessments of homogeneity and stability revealed that distributed samples were the first requirements of the proficiency test, along with ensuring that the measurements and observed performance differences were related to 'the practice of each laboratory rather than the variability of the examined material. A summary of the results, contributions, and implications is provided in Table 9.

Table 9 Summary of the results, contributions and implications of this study

Research Question	Research Result	Contribution	Implication
Which statistical tools or comprehensive interlaboratory performance benchmarking can be used to detect laboratory performance variations in proficiency testing results for mechanical testing?	Statistical assessments, graphical analysis and parametric methods were used in advanced evaluation of interlaboratory performance and resulted in variations between laboratories within acceptable technical limits.	Usage of both statistical and analytical assessment including Z-scores, Anova, Kruskal-Wallis, normality testing, R&R and graphical analytics highlights possibilities of methodical and technical variations between laboratories.	Proficiency testing service providers can support their evaluation results in reports by using comprehensive interlaboratory performance benchmarking.

In accordance with the Charpy impact test, all participating laboratories created absorbed energy values within the implemented acceptance interval. Statistical assessments, z scores and graphical analyses reflected strong interlaboratory agreement and a high level of repeatability. Although parametric methods such as ANOVAs resulted in significant differences between laboratories, these detected differences remained within acceptable technical limits and did not indicate poor performance by any laboratory. To support the robustness and reliability of the results across the network, control charts and distribution analyses were used.

All laboratories performed within the acceptance criteria, indicating satisfactory overall performance, while their results demonstrated distinct variability patterns.

The yield strength measurements exhibited the greatest dispersion, including outliers exceeding the upper tolerance limit, whereas the ultimate tensile strength and elongation results displayed tighter clustering around the predefined assigned values.

In both proficiency testing schemes, methodological challenges related to the reproducibility of the mechanical testing process were observed. Although all laboratories performed measurements under standardized protocols, variations were observed among the laboratories because of the cumulative influences of calibration discrepancies, environmental fluctuations, the strain-rate selection process, the extensometer alignment scheme, and postfracture gauge-length measurement practices. These factors are sources of uncertainty in mechanical testing research.

For tensile test parameters such as the yield strength, ultimate tensile strength, and elongation after a fracture, consistency was observed. Significant z scores were achieved by most laboratories, indicating conformance with the accredited values and acceptance criteria. Some deviations were observed in each parameter because of procedural or equipment-related biases. Nevertheless, the performance of all 'the laboratories was within the tolerance limits of proficiency testing, demonstrating competence in mechanical tests of metallic materials.

The combination of different statistical evaluation methods, including an ANOVA, the Kruskal–Wallis test, normality testing, and descriptive visualization, provided confirmation that the testing process was successfully controlled. The observed variations reflect the intrinsic sensitivity of mechanical testing, which stems from factors such as calibration, environmental stability, the strain rate, and operator techniques rather than systematic measurement errors. Proficiency tests performed by laboratories using ambient temperature test conditions resulted in measurements within the set limits, where all test results remained within the acceptance criteria defined by the assigned value and standard deviation of the program.

Overall, the program validated both the technical capabilities and the reliability of the participating laboratories. There is room for continuous improvement to support targeted reviews of variability sources, particularly in yield strength and ductility measurements, and procedures, calibration routines, and operator competency assessments can be maintained. Future studies involving larger numbers of participants, extended environmental control studies, and additional performance indicators will further strengthen the achievable benchmarking power and promote harmonized mechanical testing practices worldwide.

Declarations

- **Ethics and Consent to Participate declarations:** Not Applicable
- **Consent to Publish declaration:** Not Applicable
- **Funding:** This research received no external funding.
- **Clinical Trial Registration:** Not Applicable.
- **Competing interests:** The authors declare no competing interests.
- **Data Availability:** The datasets analyzed during the current study are not publicly available due to confidentiality agreements with Caspian Engineering Solutions

but may be available from the corresponding author upon reasonable request and with permission from Caspian Engineering Solutions.

References

[1] Panteghini, M., Krintus, M.: Establishing, evaluating and monitoring analytical quality in the traceability era. Critical Reviews in Clinical Laboratory Sciences **62**(3), 148–181 (2025)

[2] Bouhouche, S., Ziani, S., Mentouri, Z., Bast, J.: Uncertainty estimation of mechanical testing properties using sensitivity analysis and stochastic modelling. Measurement **62**, 149–154 (2015)

[3] ISO/IEC 17043:2023: Conformity assessment – general requirements for the competence of proficiency testing providers. Standard ISO/IEC 17043:2023, International Organization for Standardization, Vernier, CH (2023). https://www.iso.org/standard/80864.html

[4] 13528:2022, I.: Statistical methods for use in proficiency testing by interlaboratory comparison. Standard ISO 13528:2022, International Organization for Standardization, Vernier, CH (2022). https://www.iso.org/standard/78879.html

[5] Metallic materials – charpy pendulum impact test – part 1: Test method. Standard EN ISO 148-1:2016, International Organization for Standardization, Geneva, CH (2016). https://www.iso.org/standard/64598.html

[6] Metallic materials – tensile testing – part 1: Method of test at room temperature. Standard EN ISO 6892-1:2019, International Organization for Standardization, Vernier, CH (2019). https://www.iso.org/standard/78322.html

[7] Coucke, W., Rida Soumali, M.: Demystifying eqa statistics and reports. Biochemia medica **27**(1), 37–48 (2017)

[8] Visser, R.G.: Interpretation of interlaboratory comparison results to evaluate laboratory proficiency. Accreditation and quality assurance **10**(10), 521–526 (2006)

[9] Ehrmeyer, S., Laessig, R.: An evaluation of the ability of proficiency testing programs to determine intralaboratory performance. peer group statistics vs clinical usefulness limits. Archives of pathology & laboratory medicine **112**(4), 444–448 (1988)

[10] Tsamatsoulis, D.: Comparing the robustness of statistical estimators of proficiency testing schemes for a limited number of participants. Computation **10**(3), 44 (2022)

[11] Cofino, W.P., Crum, S., Vark, W., Molenaar, J.: Robustness comparison of three

statistical methods commonly used in proficiency tests. Accreditation and Quality Assurance **30**(5), 507–519 (2025)

[12] Bisson, K.R., Beharry, A., Blais, N., Carter, M.D., Cheema, P.K., Desmeules, P., Garratt, J.G., Melosky, B., Lo, B., Snow, S., *et al.*: Novel approach to proficiency testing reveals significant variations in biomarker practice leading to critical differences in lung cancer management. JTO Clinical and Research Reports **6**(7), 100837 (2025)

[13] Alper, M.P.: Enhancing proficiency testing: Exploring the innovations in iso/iec 17043: 2023. Mapan **39**(2), 221–227 (2024)

[14] Ilinca, R., Luțescu, D.A., Sfeatcu, R.I., Gherlan, I., Dănciulescu-Miulescu, R.-E., Țâncu, A.M.C.: A comprehensive review of proficiency testing/interlaboratory comparisons policies of the ea-mla signatories applicable to medical laboratories. Revista Romana de Medicina de Laborator **32**(2), 123–134 (2024)

[15] Vander Heyden, Y., Smeyers-Verbeke, J.: Set-up and evaluation of interlaboratory studies. Journal of Chromatography A **1158**(1-2), 158–167 (2007)

[16] Arrhenius, K., Morris, A., Hookham, M., Moore, N., Modugno, P., Bacquart, T.: An inter-laboratory comparison between 13 international laboratories for eight components relevant for hydrogen fuel quality assessment. Measurement **230**, 114553 (2024)

[17] Sipkens, T.A., Mehri, R., Perez Calderon, R., Green, R.G., Oldershaw, A., Smallwood, G.J.: Interlaboratory comparison of particle filtration efficiency testing equipment. Journal of Occupational and Environmental Hygiene **22**(4), 259–273 (2025)

[18] Cherie, N., Deress, T., Wolde, M., Teketelew, B.B., Tamir, M., Angelo, A.A., Terekegne, A.M., Chane, E., Nigus, M., Berta, D.M., *et al.*: Performances and determinants of proficiency testing in clinical laboratory services at comprehensive specialized hospitals, northwest ethiopia. Scientific Reports **14**(1), 7745 (2024)

[19] Vasselon, V., Rivera, S.F., Ács, É., Almeida, S.B., Andree, K.B., Apothéloz-Perret-Gentil, L., Bailet, B., Baričević, A., Beentjes, K.K., Bettig, J., *et al.*: Proficiency testing and cross-laboratory method comparison to support standardisation of diatom dna metabarcoding for freshwater biomonitoring. Metabarcoding and metagenomics **3**, 1 (2025)

[20] Cavalli, F., BOROWIAK, A., DOUGLAS, K., *et al.*: Results of the Second Comparison Exercise for EU National Air Quality Reference Laboratories (AQUILA) for TC, OC and EC Measurement (2011), (2013)

[21] Beckert, S.F., Domeneghetti, G., Bond, D.: Using historical results obtained in

the tensile tests for type a evaluation of uncertainty. Measurement **51**, 420–428 (2014)

[22] Possolo, A.: Interlaboratory consensus building challenge. Springer (2020)

[23] Chao, Y.J., Ward Jr, J., Sands, R.G.: Charpy impact energy, fracture toughness and ductile–brittle transition temperature of dual-phase 590 steel. Materials & design **28**(2), 551–557 (2007)

[24] Sadowski, A.J., Rotter, J.M., Reinke, T., Ummenhofer, T.: Statistical analysis of the material properties of selected structural carbon steels. Structural Safety **53**, 26–35 (2015)

[25] Pommé, S., Spasova, Y.: A practical procedure for assigning a reference value and uncertainty in the frame of an interlaboratory comparison. Accreditation and quality assurance **13**(2), 83–89 (2008)