Celal Bayar University Journal of Science

A Comparative Analysis of Robust Statistical Methods by ISO 13528 for the Evaluation of Tensile Proficiency Testing Results

Kemal Kuş[1*], Bülent Aydemir[2]

[1] VESTEL Beyaz Eşya Tic.San.A.Ş., Keçili/Şehzadeler / Manisa,Türkiye
[2] Duzce Universitesi, Mühendislik Fak. Mekatronik Müh.Böl. Duzce, Turkiye
* kemal.kus@vestel.com.tr
*Orcid: 0000-0001-8189-0196

Received: July 21, 2025
Accepted: October 31, 2025
DOI: 10.18466/cbayarfbe.1747154

Abstract

The objective of this study is to evaluate and compare the z-scores calculated for mechanical parameters obtained from interlaboratory basis tests using five robust statistical methods defined in ISO 13528:2022[1]: normalized interquartile range (nIQR), median absolute deviation (MADe), Algorithm A, Q/Hampel, and Qn. These methods are assessed in terms of their ability to suppress outliers, their effect on score variance, and their discriminatory power between laboratories. By analyzing the strengths and limitations of each method across different parameter types (such as Rp0.2, Rm, A%, and modulus of elasticity), this study aims to demonstrate the critical role that robust statistical method selection plays in ensuring fair, consistent, and technically reliable evaluation of laboratory performance in proficiency testing schemes.

Keywords: Proficiency Test (PT), Z score, ISO 13528, ISO 17043, ISO 17025, Tensile test

1. Introduction

According to the ISO/IEC 17025 standard [3], accredited testing laboratories are required to participate in proficiency testing programs. Clause 7.7 of ISO/IEC 17025 refers to ISO/IEC 17043[2] for the requirements and specifications regarding proficiency testing. ISO/IEC 17043 outlines the conditions for participants, the responsibilities and requirements of proficiency testing providers, the structure of the proficiency testing scheme, reporting principles, and the recommended statistical approaches to be used [2]. In addition, ILAC P9:01/2024 [6] requires accredited laboratories to regularly participate in proficiency testing and/or interlaboratory comparisons in order to demonstrate the validity of their results. It is also possible for proficiency testing providers to be accredited by this standard. ISO/IEC 17043 allows several methods to evaluate proficiency testing results, such as the z-score, En-score, z′-score, and zeta-score [4],[9]; among these, the z-score is the most widely used as below formula.

$$z_i = \frac{(x_i - x_{pt})}{\sigma_{pt}} \qquad (1)$$

x_i participant result
x_{pt} assigned value
σ_{pt} standard deviation of the results

Although ISO/IEC 17043 defines some general calculations, it refers to ISO 13528 as the main standard for most statistical evaluations [2],[1].

In this study, the performance of 15 laboratories that participated in a proficiency test in the field of tensile testing was evaluated using the z-score approach. To determine the z-scores, various robust statistical methods were applied, and the differences among these methods were investigated.

The z-score evaluation was performed for the tensile test parameters: 0.2% proof strength (Rp0.2), tensile strength (Rm), elongation (A%) and modulus of elasticity (mE). In addition, z-score distributions were obtained for each test parameter using five different robust statistical methods:

1. Median + normalized interquartile range (nIQR)
2. Median + median absolute deviation (MADe)
3. Algorithm A- An iterative robust method specific to ISO 13528
4. Q/Hampel method- combining the Q method for robust standard deviation and the Hampel estimator for robust mean
5. Qn method – a robust scale estimator proposed by Rousseeuw and Croux

The z-score distributions calculated using these methods were compared in terms of distribution width and variance, sensitivity to outlier suppression, inter-laboratory discrimination capability, and statistical reliability.

2. Definition of Robust Statistical Methods

This section describes the robust statistical methods provided in Annex C of ISO 13528 [1] and details the calculation procedures. Robust statistical methods such as MADe and Qn have also been extensively studied in the statistical literature [11], providing theoretical justification for their efficiency and breakdown properties.

2.1. Normalized Interquartile Range (nIQR) Method

The normalized interquartile range (nIQR) is calculated by multiplying the difference between the 75th percentile (third quartile, Q3) and the 25th percentile (first quartile, Q1) by the constant 0.7413. In other words:. $nIQR=0.7413\times(Q3-Q1)$. Here, Q1 represents the 25th percentile (first quartile), and Q3 represents the 75th percentile (third quartile) of the dataset. If the 25th and 75th percentile values are identical (e.g., when the dataset contains many repeated values in the middle range), $Q3-Q1=0$, and therefore, the nIQR value may also be zero. In such cases, alternative approaches should be applied to calculate a robust standard deviation, such as the classical standard deviation after excluding outliers or other robust methods suggested in the literature.

The calculation steps of this method can be expressed as follows:
a) Sorting the Data: Arrange all laboratory results in ascending order.
b) Determining Quartile Values: Identify the first quartile (Q1) and third quartile (Q3) of the ordered dataset (i.e., the values below which 25% and 75% of the data fall, respectively).
c) Calculating the Interquartile Range: Compute the interquartile range as Q3-Q1.
d) Normalization Factor: Multiply the obtained difference by 0.7413 to obtain the nIQR value. This constant is used to convert the interquartile range to an equivalent standard deviation under the assumption of a normal distribution.

Conceptually, the nIQR method provides a robust estimate of the spread (dispersion) of a dataset. The interquartile range (IQR) represents the width of the middle 50% of the data and is not influenced by extreme outliers. By scaling this range with the constant 0.7413 (1/1.349 under normality), the IQR is transformed into a measure comparable to the standard deviation.
As a robust method, nIQR provides reliable results as long as no more than 25% of the dataset consists of outliers. In other words, if more than one-quarter of the data are extreme outliers, the nIQR estimate may become unreliable. Compared to methods with a higher breakdown point, such as MADe (which has a 50% breakdown point), nIQR tolerates fewer outliers; however, it is computationally simpler, requiring only a single ordering step and the calculation of two percentile values.
Special care should be taken in cases where there are many repeated values (e.g., when half of the laboratory results are identical), as both nIQR and MADe may yield zero values under such circumstances. In these cases, alternative estimation methods should be considered. Additionally, nIQR may exhibit a slight negative bias when the sample size is small (particularly with fewer than 30 laboratory results), meaning that the estimated spread may systematically underestimate the true variability.
In conclusion, the nIQR method provides a practical and fast robust dispersion estimate in proficiency testing datasets when the number of laboratories is reasonable and the proportion of outliers is limited.

2.2. Median Absolute Deviation (MADe) Method

MADe is the scaled version of the median absolute deviation (MAD) adjusted for a normal distribution. First, the MAD is calculated by taking the absolute deviation of each data point from the median of the dataset, followed by determining the median of these absolute deviations. This value is then multiplied by the constant 1.4826 (rounded to 1.483 in the standard) to obtain the MADe. In other words:
$MADe=1.483\times median(|xi-median(x)|)$

The constant 1.483 is used to make the expected value of the median absolute deviation equivalent to the classical standard deviation under the assumption of normality [1] (for a normal distribution, this constant is approximately 1.4826).

If half or more of the dataset consists of identical values, all values will be equal to the median, resulting in a median of absolute deviations equal to zero, and thus MADe = 0. In such cases, robust dispersion estimation cannot be performed; alternative methods, such as the nIQR method or the classical standard deviation after removing outliers, should be considered.

The calculation steps of this method can be expressed as follows:
a) Determining the Median: Calculate the median (central value) of the laboratory results.
b) Absolute Deviations: Determine the difference between each laboratory result and the median and calculate the absolute values of these deviations ($|di|$).
c) Median of Absolute Deviations: Arrange all $|di|$ values in ascending order and determine their median. This value represents the typical magnitude of deviation relative to the median of the dataset.

d) Scaling Factor: Multiply the median of the absolute deviations by 1.483 to obtain the MADe value. In this way, the median absolute deviation is scaled to be equivalent to the classical standard deviation under the assumption of a normal distribution.

Conceptually, the MADe method is one of the most widely used measures for robust standard deviation estimation. It takes the median as the central location and evaluates how far each value deviates from this median. Taking the median of these absolute deviations, it provides a measure of dispersion that is unaffected by extreme outliers.

The MAD statistic has a breakdown point of approximately 50%, meaning it remains stable even when up to half of the data are contaminated by extreme outliers. Compared to nIQR, MADe offers greater resistance to distortion; even when nearly half of the data are outliers, both the median and the median absolute deviation remain largely unaffected.

However, MADe also has certain limitations. Particularly when the number of laboratories is small (e.g., fewer than 20–30 participants), MADe estimates tend to be systematically lower (negatively biased), which may lead to inaccuracies in scoring. Additionally, when more than 50% of the data have identical values, the information content decreases, resulting in the issue of MADe = 0.

Under a normal distribution, MADe has an efficiency of approximately 37%, meaning its variance as an estimator of variability is higher (less precise) compared to the classical standard deviation. Despite this relatively low efficiency, MADe is highly resistant to outliers and is widely used in proficiency testing for the robust estimation of values such as the assigned standard deviation (σ_{pt}). In particular, MADe is a reliable, robust measure of dispersion when the dataset contains many marginal deviations but a limited number of extreme outliers.

2.3. Algorithm A (Robust Mean and Standard Deviation) Method

I

Algorithm A is an iterative method used to simultaneously calculate the robust mean (x*) and the robust standard deviation (s*) of a dataset.

Initially, the robust mean estimate (x*) is taken as the median of all results, and the initial robust standard deviation (s*) is calculated as the scaled median absolute deviation (i.e., MADe). In other words, in the first step:
$x^*=\text{median}(x_i)$ $s^* =1.483\times\text{median}(|x_i-x^*|)$

After these initial values are determined, Algorithm A iteratively updates the estimates of the mean and standard deviation by restricting (downweighting) the influence of extreme values, thereby reducing the impact of outliers on the final estimates.

The calculation steps of this method can be expressed as follows:

a) Determine the initial robust mean (x*) as the median of all laboratory results. Then, calculate the absolute deviations of each result from x*, take the median of these deviations, and multiply it by 1.483 to obtain the initial robust standard deviation (s*):
If $s^* =0$ at this step, it indicates that more than half of the data have identical values; in such cases, the classical standard deviation may be used as an initial value, and if obvious outliers exist, they should be excluded before continuing the iteration.
b) Threshold Definition (δ): A threshold value is defined based on the current s* value. In ISO 13528, this threshold is specified as: $\delta=1.5\times s^*$
This corresponds to 1.5 times the current robust standard deviation and is used as the "outlier" cut-off limit.
c) Winsorization (Data Limitation): For each laboratory result xi, the following steps are performed by comparing it with x*

- If $x_i> x^*+\delta$, set $x_i= x^*+ \delta$ (very large values are trimmed to the upper limit).
- If $x_i< x^*-\delta$, set $x_i= x^*+-\delta$ (very small values are raised to the lower limit).
- If $x^*-\delta \leq x_i \leq x^*+ \delta$ leave xi unchanged.

This step limits the influence of extreme outliers (Winsorization) by allowing the data to vary only within a specified range. For example, any value deviating more than 1.5 s* from the current x* is restricted so that its deviation does not exceed 1.5 s*
d) Updated Mean: The arithmetic mean of the trimmed (Winsorized) values is calculated. This value represents the updated estimate of the robust mean (x*). (Using the mean of Winsorized data instead of the median increases the efficiency of the method.)
e) Updated Standard Deviation: The dispersion of the trimmed values around the new x* is calculated. For this purpose, the mean of the squared differences between the trimmed values and the updated x* is computed, and its square root is taken. The obtained value is then multiplied by a correction factor. In Algorithm A, this correction factor is approximately 1.134. (The constant 1.134 compensates for the slight reduction in variance caused by the trimming process, ensuring that s* is as accurate as an untrimmed estimate under a normal distribution.)
f) Convergence Check: The newly obtained **x*** and **s*** values are compared with those from the previous iteration. If there is a change in the robust mean and/or robust standard deviation exceeding the desired precision (e.g., at the third decimal place, as recommended by ISO), the threshold δ is updated (using the new **s*** as in Step 2), and the procedure is repeated starting from Step 3. The iteration is terminated when the changes become

negligible (e.g., when x* and s* remain identical up to the third decimal place for two consecutive iterations)

g) Reporting of Results: The x* value obtained in the final iteration is reported as the robust mean of the dataset, and the s* value is reported as the robust standard deviation.

Conceptually, Algorithm A is a robust approach defined in ISO 13528 and ISO 5725-5, and it can also be regarded as an iterative Huber-type M-estimator [8]. The algorithm begins with the median and MADe as initial estimates and then iteratively improves both the robust mean and standard deviation by "slightly trimming" the data (keeping values within the 1.5 s* threshold).

The core idea of Algorithm A is to limit the influence of values far from the median while increasing the influence of the majority of the data, thereby producing an outlier-resistant mean and standard deviation. With a breakdown point of approximately 25%, Algorithm A may fail when more than one-quarter of the data are extreme outliers. This demonstrates that Algorithm A shares the same breakdown resistance (~25%) as nIQR and is less robust against extreme contamination than median/MAD-based methods.

However, when the proportion of outliers remains reasonable (below 25%), Algorithm A is highly effective: under a normal distribution, the efficiency of the robust mean estimate reaches approximately 97%, while the efficiency of the robust standard deviation estimate is around 73–74%. These values indicate substantially higher statistical efficiency (lower variance) compared to simple robust location estimators such as the median. In other words, if the dataset is approximately normally distributed, the x* and s* values calculated by Algorithm A are very close to the classical mean and standard deviation, yet they remain unaffected by occasional outliers.
Today, Algorithm A is widely preferred by many proficiency testing providers for robust statistical calculations and provides a reliable method for determining assigned values and standard deviations, particularly in situations where outliers are present but limited. Nevertheless, if the proportion of outliers is expected to be very high (greater than ~20%) or if the initial s* estimate becomes meaningless due to extreme contamination, ISO 13528 recommends using certain variations of this algorithm or alternative robust methods, such as the Q/Hampel method.

2.4. Q Method (Robust Standard Deviation Estimation)

The Q method is a robust technique developed for estimating the standard deviation of a dataset, characterized by a high breakdown point and high efficiency. Unlike methods that depend on a location measure (such as the mean or median), this method derives a measure of dispersion by considering all pairwise differences within the dataset.

In the Q method, all possible pairwise differences in the dataset are first calculated: for each laboratory result xi and every other result xj , the absolute differences (|xi−xj|) are listed.

The total number of such differences for a dataset of p observations is (p–1)/2. These differences are then sorted in ascending order, and a statistic around the lower quartile (25th percentile) is selected. Specifically, the value corresponding approximately to the first 25% of the ordered differences is identified.

A practical way to determine this position, as proposed by Rousseeuw and Croux [7], is to select the k-th element from the ordered sequence of differences, where k is chosen as a function of the number of observations. Here is h = p/2 + 1 which roughly corresponds to the position where about one-quarter of all pairwise differences are smaller than or equal to this value. Finally, a scale factor (Cn), depending on the sample size, is applied to this selected difference value to ensure that the estimate is a consistent (unbiased) estimator of the standard deviation under a normal distribution. For large sample sizes, the constant Cn converges to approximately 2,2219[7]. Thus, the Q method can be summarized as Q= Cn × d(k) where d(k) is the k-th ordered absolute pairwise difference (|xi−xj|)

The calculation steps of this method can be expressed as follows:

a) Calculation of Pairwise Differences: Compute the absolute differences between all laboratory results. For each pair, i<i dij=|xi−xj|. The total number of such differences is p(p−1)/2
b) Ordering of Differences (Ranking) : All dij values are ranked in ascending order to form an ordered sequence of pairwise absolute differences d(1) ≤ d(2) ≤...≤d(n) where n = p(p−1)/2
c) Calculation of the $H_I(x)$ Function: At this stage, the proportion of differences up to a certain threshold is computed. The indicator function 1(·) takes the value 1 if the condition is true and 0 otherwise. This calculation produces an empirical cumulative distribution function (ECDF) of all pairwise differences
d) Calculation of the $G_I(x)$ Function: A new distribution function is defined based on the normalized pairwise differences. Here, sss denotes the standard deviation estimate, which may be iteratively updated. However, in most applications, MADn, IQRn, or an initial Q estimate can be used at this stage.
e) Determination of $G_I^{-1}(z)$ by Interpolation: For the target trimming proportion (e.g., z=0.25 or z=0.75), the inverse function value $G_I^{-1}(z)$ is obtained through linear

interpolation based on the previously constructed $G_1(x)$ function.
f) Calculation of the Robust Standard Deviation (s*): The interpolated $G_1^{-1}(z)$ value is then transformed into a standard deviation estimate according to the following formula: $s^* = Cn \times G_1^{-1}(z)$ where Cn is the sample size–dependent scaling factor that ensures consistency under a normal distribution.

Conceptually, the Q method is a derivative of the innovative scale estimators developed by Rousseeuw and Croux in the field of robust statistics and has been widely adopted in proficiency testing. The most important characteristic of this method is that it focuses on the pairwise distances between measurement results rather than the tails of the distribution.

While the classical standard deviation measures deviations from the mean, the Q method evaluates the relationships among data points, allowing it to estimate dispersion independently of any location measure (mean or median). This feature makes the method applicable even when the data distribution is multimodal (contains several groups) or when results are clustered at certain rounded values. Even if many values are identical, their pairwise differences are zero; the method accounts for these zero differences appropriately.

The breakdown point of the Q method is approximately 50%, meaning that even if half of the data are contaminated by extreme outliers, the dispersion calculated from the remaining half will not increase without bound. This demonstrates that the Q method is as robust as the median/MAD approach. Moreover, under a normal distribution, its efficiency reaches approximately 82%, which is considered very high for a robust method and indicates strong performance compared to the classical standard deviation.

The Q method is also referred to in the literature as Qn (when used for a single dataset, as described in the next section) and is recommended in ISO 13528 in combination with the Hampel estimator (Q/Hampel method) for robust mean calculation. In situations where a considerable number of outliers exist (e.g., when some laboratories systematically report significantly different results), the 25% tolerance limit of Algorithm A may be insufficient; in such cases, the Q/Hampel combination provides a solution that can tolerate up to 50% contamination.

Although computationally more intensive than simpler methods, this is not a limitation in practice with modern computing resources. In conclusion, the Q method is a powerful statistical tool for robust standard deviation estimation, offering both high robustness and high accuracy when these properties are required.

2.5. Hampel Method (Robust Mean Estimation)

The Hampel method aims to calculate a reliable mean value, referred to as the robust assigned value, from a dataset. This robust assigned value (commonly denoted as x*) represents the central tendency of the data in a stable manner by reducing the influence of outliers (extreme results) on the mean.
The calculation steps of this method can be expressed as follows:
a) Median-Based Initial Estimate: The calculation process begins by taking the median of the dataset as the initial estimate. The median of all results is denoted as $X_{initial}$. Since the median represents the central value of the data and is insensitive to extreme values, it provides a robust location estimate and serves as an appropriate starting point for the Hampel method.
b) Calculation of Deviations ($x_i - x^*$): For each result, the deviation from the initial median estimate is calculated. Specifically, for each observation i, $\Delta i = x_i - x^*$. These deviations indicate how far each laboratory result is from the current estimated mean value.
c) Normalization with MAD: To evaluate and compare the magnitude of deviations, the MAD (Median Absolute Deviation) is used. First, a robust measure of dispersion around the current x* is obtained by calculating the median of the absolute deviations $s^* = 1.483 \times \text{median}(|x_i - x^*|)$. The resulting s* value serves as a robust (outlier-resistant) estimate of the standard deviation. The constant 1.483 is used to convert the median absolute deviation into an equivalent standard deviation under the assumption of normality. The normalized deviation for each result is then calculated as
$q_i = x_i - x^* / s^*$. In the calculations, the absolute values of qi ($|q_i|$) are taken into consideration.
d) Definition of the Weighting Function: In the Hampel method, the contribution of each result to the mean is limited by a weighting function defined based on the magnitude of the normalized deviation. According to ISO 13528:2022, the Hampel weighting function has a three-part structure, and the parameters are typically chosen as a = 1.5, b = 3.0, and c = 4.5.
The function is defined for each result depending on its qiq, iqi value as follows:

- Results very close to the centre are assigned a full weight (1).
- As the deviation increases, the weights are gradually reduced.
- Extremely distant results (e.g., $|q_i| > c$) are assigned a weight of 0, meaning they are completely excluded from the calculation.

In this way, the influence of outliers on the mean estimation is effectively limited.
e) Calculation of the Weighted Mean (Hampel x*): After the weights defined above are applied to each result, the Hampel weighted mean is calculated to obtain a new value. This robust assigned value is determined by taking the weighted average of all results.

The obtained value is referred to as the robust assigned value calculated by the Hampel method (in the literature, this value is also denoted by the symbols Hampel x or Hx*). This value represents a reliable consensus value that is free from the influence of outliers in the data.
f) Iteration Process (Convergence)**:** In the Hampel method, the calculation steps described above can, if desired, be repeated iteratively. In the first iteration, the initial x* value starts with the median and is then updated based on the weighted mean. Subsequently, using this new x*, the deviations and corresponding qi weights are recalculated, and x* updated again. This process continues until the change in x* between successive iterations becomes negligible (i.e., convergence is achieved). Once convergence is reached, the final x* and the corresponding s* values are accepted as the robust mean and the robust measure of dispersion of the dataset.
Consequently, the Hampel method provides a stable assigned value that is effectively resistant to outliers. The Hampel method is known in the robust statistics literature as a robust location estimator with a high breakdown point (~50%) and high efficiency [8]. With a breakdown point of 50%, the method can theoretically maintain a stable mean estimate even if half of the dataset consists of extreme outliers. Under a normal distribution, the efficiency of the Hampel mean reaches approximately 95–96%, which is considerably higher compared to simpler robust location estimators such as the median (~64%).

This implies that, in the absence of outliers, the variance of the mean estimated by the Hampel method is very close to that of the classical mean. The method achieves these advantages by using a complex weighting strategy:

- Small deviations are almost fully utilized,
- Moderate deviations are down-weighted, and
- Very large deviations are completely discarded.

Thus, the Hampel method focuses primarily on the main body of the distribution, while still allowing limited influence from slightly distant but potentially meaningful values. This redescending weighting approach prevents excessive shifting of the mean even when the dataset contains secondary modes (minor peaks) or asymmetries, effectively ignoring these side modes beyond a certain point.

Although computationally more demanding than simpler methods, ISO 13528 generally recommends applying the Q/Hampel method (Hampel mean+Q standard deviation) with computer assistance. Modern computational tools make this feasible, and some proficiency testing software (e.g., PROLab Plus) already integrates the Q/Hampel approach [13]. Alternative redescending estimators such as biweight have also been proposed [10].

The Hampel method is particularly preferred when the proportion of outliers is expected to be high (>20%) or when the data distribution contains small secondary clusters, as it can provide a reasonable mean estimate even under these extreme conditions—beyond the capability of simpler methods such as Algorithm A.

In summary, the Hampel robust mean estimator demonstrates an outstanding combination of robustness (resistance to up to 50% contamination) and efficiency (up to 96% under normality). ISO 13528:2022 recommends this method in cases where the data structure is complex or significantly contaminated.

2.6. Qn Method (Rousseeuw–Croux Robust Scale Estimator)

The Qn method is a specific formulation of the Q method described in the previous section and is used for robust scale (standard deviation) estimation. Introduced in the literature by Rousseeuw and Croux (1993) [7].Qn is a statistic that remains unaffected by outliers even if up to 50% of the data are contaminated (50% breakdown point) and provides high efficiency (~82%) under a normal distribution.

In ISO 13528:2022, the Qn method is recommended as a highly reliable scale estimator **,** particularly for laboratories reporting a single test result. Mathematically, Qn is defined as follow: All pairwise differences are calculated as $d_{ij}=|x_i-x|$,for $i<j$. These differences are sorted in ascending order, and the k-th smallest difference is selected (the definition of k is the same as in the Q method; it approximately corresponds to the first quartile of all pairwise differences). The Qn value is then calculated as $Q_n=c_n \times d(k)$ where c_n is a consistency factor determined according to the sample size. For example, tables of c_n are available for datasets containing 10–20 laboratories, and for large p, c_n converges to approximately 2.2219.

ISO 13528 notes that when calculating Qn, numerical correction factors should be used to adjust for bias in small sample sizes, and, similar to the Q method, laboratory replicates may be included if necessary.

The calculation steps of this method can be expressed as follows:
a) Calculation of Pairwise Differences: Create a single list of laboratory results (one result per laboratory). Compute the absolute differences between all pairs of results (dij values).
b) Sorting the Differences and Determining k: Sort all dij values in ascending order. For p observations, determine the position of the value to be selected using: $h = \lfloor p/2 \rfloor + 1$, $k = h(h-1)/2$ (This step corresponds to Step 3 of the Q method.).
c) Selecting the Relevant Difference: Take the k-th smallest difference d(k)) from the ordered list.

d) Multiplying by the Consistency Factor: Multiply the selected difference by the corresponding cn factor to obtain the Qn value Qn=cn×d(k).
e) Interpretation of the Result: The Qn value represents the robust standard deviation estimate of the dataset. If a robust mean (location estimate) is also required (e.g., Qn is typically used together with the median), the median is reported as the location estimate. ISO 13528 often refers to the Qn method in conjunction with the Hampel robust mean estimator, as Qn alone provides only the measure of dispersion.)

Conceptually, the Qn method can be regarded as a form of "probabilistic median absolute deviation" in the robust statistics literature. One of the disadvantages of the MAD (or MADe) method is that it partially relies on the assumption of symmetry in the data and has relatively low efficiency. The Qn method was designed to overcome these limitations: it does not require any symmetry assumption (as it is based on pairwise differences rather than deviations from the mean or median) and provides a relatively high efficiency of approximately 82%, making it effective even for small sample sizes.

With a 50% breakdown point, Qn is at least as robust as the median. In other words, even if half of the data are arbitrarily contaminated, the internal pairwise differences among the remaining half will prevent Qn from increasing without bound. Thanks to this property, Qn provides a reliable scale estimate even under highly challenging conditions. For example, even if a few laboratory results deviate significantly from the others, Qn still provides a general measure of dispersion by incorporating both the differences among these extreme values and their differences with the rest of the data.

In ISO 13528:2022, Qn is cited along with its bias-corrected formula for all practical sample sizes. Conceptually, the calculation of Qn is the same as that of the Q method; therefore, the term "Q method" is often used interchangeably to refer to Qn. In fact, when referring to the "Q method," Qn is generally implied, as it is the version originally defined by Rousseeuw and Croux.

In proficiency testing, particularly in the context of the Q/Hampel method, the "Q" component refers to the Qn scale estimator, whereas the "Hampel" component refers to the robust mean estimator. This combination is considered one of the most powerful pairs of robust statistical methods available today, recommended for datasets with a high proportion of outliers due to its ability to provide reliable assigned values and standard deviation estimates.
When used alone, the Qn method offers a practical solution in situations where a robust standard deviation is required, but the mean is sufficiently represented by the classical mean or median. For example, in a proficiency test, it is common practice to report the median as the assigned value and Qn as the robust standard deviation. In this way, when calculating robust z-scores, both location and dispersion are determined in a manner that is resistant to the influence of outliers.

3. Examination of Z-Score Distributions for Individual Test Parameters

In this study, the data for the parameters of 0.2% proof strength ($Rp_{0.2}$), tensile strength (Rm), elongation (A% - measured by extensometer), and modulus of elasticity (mE) were obtained from tensile tests conducted by 15 different laboratories according to ISO 6892-1:2019[5]. Table 1 below presents the raw laboratory results for each of these parameters. These data form the basis for the z-score calculations performed using robust statistical methods as defined in ISO 13528:2022..

Lab No	$Rp_{0,2}$ (MPa)	Rm (MPa)	A% (Eks)	mE (GPa)
Lab.1	354,7	607,3	23,1	184
Lab.2	321,5	562	25,7	207
Lab.3	313,5	516	24,7	198
Lab.4	315,1	553,2	25,4	204
Lab.5	276,3	522,1	24,6	190
Lab.6	320,4	564,2	25,4	260
Lab.7	318,6	526,3	25	202
Lab.8	321,7	522,4	25,9	209
Lab.9	394,9	611,9	24,2	186
Lab.10	317,4	536,8	24,9	201
Lab.11	312,3	532,6	24,9	195
Lab.12	319,7	550,2	25,3	203
Lab.13	389,4	503,6	23	182
Lab.14	325,5	554,2	26	208
Lab.15	331,6	566,3	26,3	210

Table 1. Raw Data from Participating Laboratories

3.1. Rp0,2 – 0,2% Proof Strength

All the defined robust statistical methods produced highly similar results, which can be attributed to the homogeneity of the measurement data and the absence of significant outliers. Median-based methods (nIQR, MADe) suppressed small deviations, resulting in narrower distributions. Algorithm A and Q/Hampel, on the other hand, slightly accounted for potential outliers, producing somewhat wider but still comparable distributions. These results are illustrated in Table 2.

For Rp0.2, the influence of method selection is minimal, and all methods can be used with confidence. The heatmaps below visualize the z-scores obtained by the laboratories according to the robust methods.
Color Coding:

- Green ($|z| \leq 2$): Acceptable,
- Orange ($2 < |z| < 3$): Warning
- Red ($|z| \geq 3$): Not Acceptable

Celal Bayar University Journal of Science
Volume 22, Issue 1, 2026, p 32-41
Doi: 10.18466/cbayarfbe.1747154

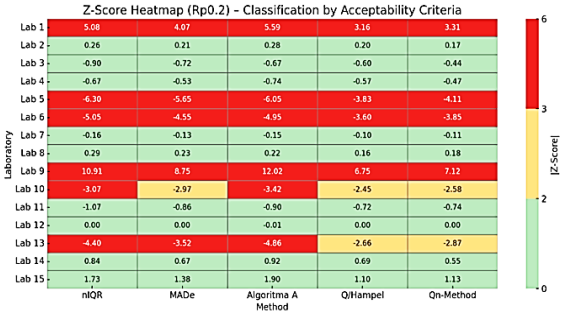

Table 2. Z score Evaluation for Rp0,2

3.2. Rm –Ultimate Tensile Strength

Like Rp0.2, all robust statistical methods produced comparable z-score ranges. MADe and nIQR yielded slightly higher |z||z||z| values due to their smaller standard deviations, whereas Algorithm A and Q/Hampel provided smoother distributions.

For Rm, the differences between methods are not significant, and the statistical confidence level is high. The heatmaps below illustrate the z-scores obtained by the laboratories according to the robust methods (Table 3).

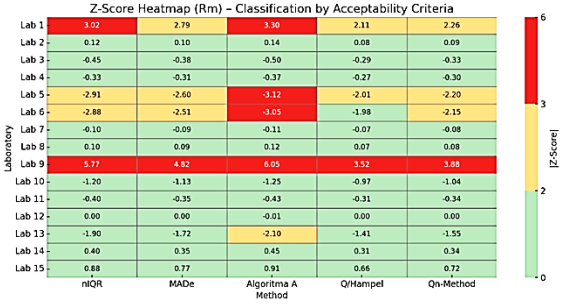

Table 3. Z score Evaluation for Rm

3.3. Elongation at Break (A%)

Due to the high variance and the presence of numerous outliers, notable differences were observed among the robust statistical methods. Median-based methods such as nIQR, MADe, and Qn strongly penalized the outliers, resulting in higher z-scores. Algorithm A, by treating outliers more gently, retained a greater concentration of values around the centre. Q/Hampel offered a balanced performance, combining both sensitivity and robustness[12].
For the A% parameter (elongation at break), highly resistant methods such as Qn and MADe are recommended.

The heatmaps in Table 4 visualize the z-scores obtained by the laboratories according to the robust methods for the A% parameter.

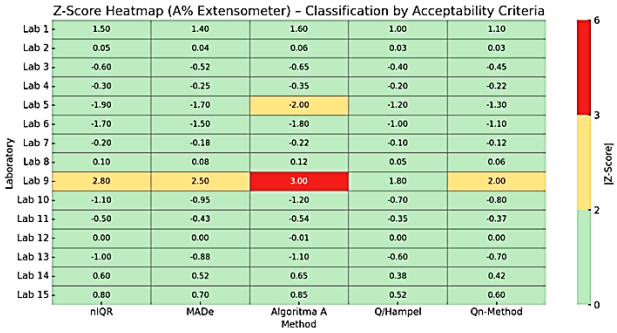

Table 4. Z score Evaluation for %A

3.4. Elastic Modulus (mE)

Among all the defined robust statistical methods, the highest variance was observed for the mE parameter (modulus of elasticity). Qn and MADe revealed outliers clearly due to their narrower standard deviations, whereas Algorithm A softened inter-laboratory differences by including some outliers. Q/Hampel maintained the data structure while providing a moderate level of discrimination.

For mE, Qn and MADe offer precise and robust results, while Q/Hampel provides a well-balanced alternative. These results are illustrated in Table 5.

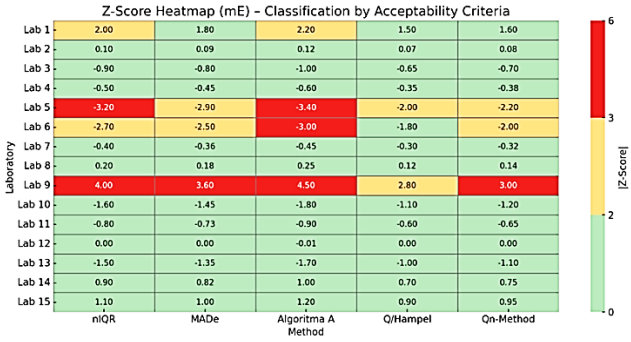

Table 5. Z score Evaluation for mE

4. Conclusion

In this study, z-score calculations were performed for different test parameters (Rp0.2, Rm, A%, and mE) based on the results of tensile tests conducted by various laboratories, using five robust statistical methods defined in Annex C of ISO 13528:2022. The methods applied were: normalized interquartile range (nIQR), Median Absolute Deviation (MADe), Algorithm A, Q/Hampel, and Qn. Each method was compared in terms of its capacity to suppress outliers, its impact on variance, and its ability to discriminate between laboratories.

For parameters with a homogeneous structure (Rp0.2 and Rm), all methods produced similar z-score distributions, and the effect of robust method selection on the results remained minimal. However, for parameters with high variance and sensitivity to outliers (particularly A% and mE), the choice of robust statistical method proved to be

critical[12]. For these parameters, both the absolute values of the z-scores and the distribution characteristics assigned to laboratories varied depending on the method used; some methods applied stricter outlier suppression, thereby providing higher inter-laboratory discrimination.

Key Technical Findings of the Study Can Be Summarized as Follows:

- Qn and MADe methods, with their high breakdown points (~50%), effectively isolated outliers and enabled sharper discrimination of laboratory performance, particularly for the A% and mE parameters. Therefore, these methods are recommended for quality-critical parameters.
- The Q/Hampel method was found to be effective in suppressing outliers while preserving the overall structure of the dataset. Especially for parameters with moderate variance (e.g., A%), it produced balanced results and offered a good compromise between statistical reliability and robustness.
- Algorithm A, which is the reference method recommended by ISO 13528, provides moderate resistance to outliers. It demonstrates high statistical efficiency and accurate standard deviation estimation for datasets with symmetrical distributions. However, it may be insufficiently suppressive in cases with a high density of extreme outliers.
- The nIQR method, although advantageous in terms of computational simplicity and speed, has limited capacity in detecting and suppressing outliers. Therefore, it is recommended only when it is certain that the dataset contains no significant outliers.

This analysis demonstrates that the selection of statistical methods in proficiency testing is not merely a technical preference but also a strategic decision that directly affects the performance evaluation of laboratories.

The inappropriate selection of a statistical method may lead to erroneous z-scores, causing both false negatives (competent laboratories incorrectly classified as non-compliant) and false positives (non-compliant laboratories appearing acceptable). Particularly for high-variance parameters, the choice of a robust statistical method should carefully balance outlier tolerance, efficiency, and computational complexity.
Based on the findings of this study, the following recommendations are provided for users:

- Algorithm A, as defined in ISO 13528, is recommended as a safe starting method for general applications.
- When the dataset contains a high number of outliers or exhibits a multimodal distribution, the use of Qn or Q/Hampel is more appropriate.
- The MADe method, while providing maximum resistance to outliers, may exhibit bias in small sample sizes; thus, its results should be interpreted with caution.
- When a balance between efficiency and accuracy is required, the Hampel method is recommended, as it produces results close to the classical mean while effectively reducing the impact of outliers. The interpretation of robust method results depends on the number of participating laboratories; small sample sizes can affect the performance of some estimators

In this context, it is concluded that the use of robust statistical methods in z-score calculations is not only a matter of meeting normative requirements but also a critical factor directly influencing decision quality in inter-laboratory comparisons.

Author's Contributions

Kemal Kuş: Drafted and wrote the manuscript, performed the experiment and result analysis.

Bulent Aydemir: Assisted in analytical analysis on the structure, supervised the experiment's progress, result interpretation and helped in manuscript preparation.

Ethics

There are no ethical issues after the publication of this manuscript.

References

[1]. ISO 13528:2022 – Statistical methods for proficiency testing by interlaboratory comparison

[2]. ISO/IEC 17043:2023 - Conformity assessment — General requirements for the competence of proficiency testing provider

[3]. ISO/IEC 17025:2017- General requirements for the competence of testing and calibration laboratories

[4]. ASTM E1301-95e1 - Standard Guide for Proficiency Testing by Interlaboratory Comparisons

[5]. ISO 6892-1:2019 - Metallic materials — Tensile testing

[6]. ILAC P9:01/2024 – ILAC Policy for Proficiency Testing and/or Interlaboratory comparisons other than Proficiency Testing

[7]. Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. Journal of the American Statistical Association, 88(424), 1273–1283.

[8]. Huber, P. J. (1981). Robust Statistics. Wiley.

[9]. Thompson, M., Ellison, S. L. R., & Wood, R. (2006). The International Harmonized Protocol for the Proficiency Testing of Analytical Chemistry Laboratories. Pure and Applied Chemistry, 78(1), 145–196.

[10]. Kafadar, K. (2003). A Biweight Approach to Robust Estimation in Interlaboratory Studies. Technometrics, 45(4), 324–330.

[11]. Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). Robust Statistics: Theory and Methods. Wiley.

[12]. Scott, D., & Thompson, M. (2005). Robust methods in interlaboratory studies. Accreditation and Quality Assurance, 10(9), 464–468.

[13]. Huberic, M. et al. (2020). Comparing the Effectiveness of Robust Statistical Estimators in Proficiency Testing. Metrology, 3(2), 10.