

Original Article

Open Access

Metrological Implications of Statistical Estimator Selection on Laboratory Performance Classification: Evidence from an Electrical Materials Proficiency Testing Program

Hoda Aleali ^{1*}, Elham Hashemi Shad ²

1. Zobe Ahan blvd., Standard Sq., Iranian National Standard Organization Standard Research Institute

2. Metrology department, Standard Research Institute

Abstract

In proficiency testing (PT), laboratory performance classifications are used to support conformity assessment and accreditation-related decisions. This study performed an Electrical Materials PT Program and investigated the effect of four commonly used estimators on laboratory performance classifications using the data related to the tensile strength of cable sheathing. To evaluate classification stability, performance categories obtained from the complete dataset were compared with those obtained after removal of two outlying results. Laboratories were classified as satisfactory, warning, or action signal according to their z-scores, and changes in classification were examined for each estimator. The results showed that estimator selection can influence conformity assessment outcomes even when the underlying measurement results remain unchanged. The arithmetic mean exhibited the greatest classification instability, resulting in changes in the performance status of 2 out of 50 laboratories (4%) following outlier removal. Hampel and median estimators each produced one classification change, whereas Algorithm A maintained identical classifications for all laboratories before and after outlier removal, demonstrating the highest classification stability. Furthermore, Hampel showed a slightly more conservative behavior near decision limits, increasing the likelihood of assigning borderline laboratories to the action category. These findings demonstrate that the choice of statistical estimator is not merely a computational consideration but a factor that can affect laboratory classifications and introduce decision uncertainty in proficiency testing. The results highlight classification stability as an important criterion in the selection of statistical methods for PT evaluation and support the use of robust estimators when participant results are close to decision thresholds.

Keywords Proficiency Testing, z-score, Statistical Estimator, Performance Classification, the tensile strength.

1. Introduction

Proficiency Testing is one of the most important tools for quality assurance and the evaluation of laboratory technical competence in measurement, testing, and calibration activities. Participation in proficiency testing schemes enables laboratories to assess their performance in com-

parison with peer laboratories and provides objective evidence of technical competence, measurement reliability, and compliance with quality standards. Consequently, PT occupies a central role in both ISO/IEC 17025 and ISO/IEC 17043 and is widely recognized as a key component of ensuring the reliability of laboratory results [1-4].



* Correspondence:
hoda.aleali@gmail.com

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Because PT datasets frequently contain extreme observations, skewed distributions, or heterogeneous participant populations, robust statistical methods have been widely adopted to reduce the influence of outliers. ISO 13528 recommends robust procedures such as Algorithm A, while alternative approaches including Hampel estimators, median-based methods, and other robust estimators have also been proposed and evaluated in the literature [5].

In recent years, direct comparisons of robust statistical methods used in proficiency testing have received increasing attention. Cofino et al., through extensive simulations involving more than 33,000 proficiency testing datasets, evaluated the performance of Algorithm A, Q/Hampel, and other robust estimators with respect to outlier resistance [6-7]. Their results showed that increased robustness is generally accompanied by reduced statistical efficiency and that the choice of estimator should depend on the distributional characteristics of the data. They also reported that differences among robust estimators are relatively small for nearly normal datasets but may become substantial in the presence of skewness and outliers [6].

Beyond studies focusing on the statistical properties of estimators, some researchers have examined the practical implications of using z-scores for laboratory performance assessment. Cordeiro et al. demonstrated that reliance solely on z-scores may, under certain circumstances, lead to misleading interpretations of laboratory performance, with biased results occasionally being classified as satisfactory [8]. These findings emphasize the importance of selecting appropriate parameters for z-score calculation. Despite significant advances in robust statistics and proficiency testing, most published studies have focused primarily on theoretical estimator properties such as bias, variance, efficiency, robustness, breakdown point, and influence function. Furthermore, many investigations have relied on simulated or synthetic datasets. In contrast, the practical impact of estimator selection on the entire proficiency testing decision chain- including the determination of the assigned value, estimation of the standard deviation for proficiency assessment, calculation of z-scores, and ultimately laboratory performance classification -has received considerably less attention in real PT datasets. Although numerous studies have compared the statistical properties of conventional and robust estimators, limited attention has been paid to the influence of estimator selection on laboratory classification outcomes in real PT schemes [6-8].

From a metrological perspective, the choice of statistical estimator should be considered part of the measurement and conformity assessment process, since changes in the

estimation method may lead to different assigned values, different standard deviations for proficiency assessment, different z-scores, and ultimately different laboratory classifications. Therefore, evaluating the impact of estimator selection on PT outcomes is of considerable practical and metrological importance.

Accordingly, the objective of the present study is to investigate the effect of statistical estimator selection on the assigned value, standard deviation for proficiency assessment, z-scores, and laboratory performance classification using a proficiency testing dataset of scheme on "Wires and Cables. In the present study, the data related to the tensile strength of cable sheathing were selected as a case study. Four methods—the arithmetic mean, median, Algorithm A, and the Hampel estimator—were compared, and the influence of outlier removal on the results obtained from each method was evaluated involving 50 participating laboratories. The primary focus of this work is not merely the comparison of estimator properties but rather the examination of the practical consequences of estimator selection on proficiency testing decision-making and the stability of laboratory performance classifications. Particular attention is given to determining how many laboratories change classification category when either the statistical estimator is changed or outliers are removed.

2.Theoretical Background

2.1. Proficiency Testing and the Role of Statistical Estimators

In proficiency testing, laboratory performance is assessed by comparing reported results with a reference value or assigned value. The most widely used performance statistic in many proficiency testing schemes is the z-score, which is defined as follows [5]:

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

Where:

x_i is the result reported by laboratory ,

x_{pt} is the assigned value, and

σ_{pt} is the standard deviation for proficiency assessment.

Therefore, the accurate determination of these parameters is essential for the reliability of proficiency testing results. Any change in the method used to estimate these parameters may directly affect the calculated z-scores and, consequently, the final decision regarding the acceptance or rejection of laboratory performance. According to ISO 13528, the selection of an appropriate statistical method for the analysis of PT data is one of the most critical steps in statistical evaluation, since interlaboratory datasets frequently contain outliers, skewness, or

non-normal distributions in practice [9].

Participants' z-scores are interpreted according to the following criteria:

$|z| \leq 2$: Indicates satisfactory performance.

$2 < |z| < 3$: Indicates a questionable (warning) performance and results in the issuance of a warning signal.

$|z| \geq 3$: Indicates unsatisfactory performance and results in the issuance of an action signal.

: Indicates satisfactory performance.

: Indicates a questionable (warning) performance and results in the issuance of a warning signal.

: Indicates unsatisfactory performance and results in the issuance of an action signal.

Classical Estimators

In classical statistical methods, the location and dispersion parameters of the data are estimated using the arithmetic mean and the sample standard deviation, respectively. If the participant results are given as:

$$x_1, x_2, \dots, x_n \quad (2)$$

Then, the arithmetic mean is defined as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3)$$

Similarly, the sample standard deviation is calculated as:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad (4)$$

When the data follow a normal distribution, the arithmetic mean possesses the highest statistical efficiency and the minimum possible variance among unbiased estimators [10].

Robust Estimators

To reduce the influence of outliers, robust estimators are employed. These methods are designed to maintain their statistical stability in the presence of a limited number of outlying observations.

Median Estimator

The median of the data is defined as follows:

$$\tilde{x} = \text{Median}(x_i) \quad (5)$$

Unlike the arithmetic mean, the median has a break point equal to 50%; That is, as long as less than half of the data are contaminated, the median value does not undergo statistical collapse [4]. To estimate the dispersion of the data, MAD is used:

$$\text{MAD} = \text{Median}(|x_i - \tilde{x}|) \quad (6)$$

In order to adapt to the standard deviation of the normal distribution, the following equation is used:

$$\text{MAD}_e = 1.483 \times \text{MAD} \quad (7)$$

The coefficient of 1.483 is derived from the properties of the normal distribution and makes MAD_e a consistent estimate for σ [10].

Algorithm A Estimator

Algorithm A is the most important robust method proposed in ISO 13528 for analyzing proficiency test data [5]. The goal of this algorithm is to balance high statistical efficiency with robustness against outliers. In this method, an initial estimate of the location and dispersion is first calculated through the median and MAD:

$$x^* = \text{Median}(x_i) \quad (8)$$

$$s = 1.483 \times \text{Median}(|x_i - x^*|) \quad (9)$$

Then the cutting limits are defined as follows:

$$\delta = 1.5s^* \quad (10)$$

For each x_i the following relationship must be satisfied:

$$w_i = \begin{cases} x^* - \delta & x_i < x^* - \delta \\ x^* + \delta & x_i > x^* + \delta \\ x_i & \text{otherwise} \end{cases} \quad (11)$$

After restricting the data, the new location and dispersion values are calculated from the following relationships:

$$x^* = \sum \frac{x_i^*}{p} \quad (12)$$

$$s = 1.134 \sqrt{\frac{\sum (x_i^* - x^*)^2}{p-1}} \quad (13)$$

This process is repeated until convergence. Unlike the outlier data removal method, Algorithm A does not completely remove the outlier data, but only limits their statistical effect. This feature is of great importance from a metrology point of view, because in a proficiency test, outlying data may indicate the actual poor performance of a laboratory and not just statistical error. Various studies have shown that Algorithm A establishes a suitable balance between Robustness and Efficiency in many practical situation [5, 11].

Hampel Estimator

One of the most important robust statistical methods is the Hampel Estimator, which is developed based on the concept of the influence function [9]. The Hampel method is used when the distribution is heavily contaminated, or has multiple clusters, or the number of outliers is large.

In this approach, outlying data are gradually given less weight, and after a certain limit, their effect becomes almost constant or negligible. The Hampel weight function is usually defined as follows:

$$w_i = \begin{cases} 0 & |q| > 4.5 \\ (4.5 - |q|)/q & 3 < |q| \leq 4.5 \\ 1.5/q & 1.5 < |q| \leq 3.0 \\ 1 & |q| \leq 1.5 \end{cases} \quad (14)$$

$$q_i = \left| \frac{x_i - x^*}{s^*} \right| \quad (15)$$

where x^* is the mean and s^* is a robust estimate of the standard deviation. This implementation of the Hampel estimator has an efficiency of approximately 96% for normally distributed data. The Hampel estimator can be tuned for higher efficiency or greater robustness to outliers by changing the weight function [14]. In this structure, data close to the center of the distribution participate almost entirely in the estimation, relatively distant data receive limited weight, and very distant data are effectively eliminated from the estimation process. Unlike direct elimination of outliers, the Hampel method prevents sudden changes in the data structure and provides greater robustness in interlaboratory data analysis [10, 12].

3. Research Method

A proficiency testing scheme on “Wires and Cables – Mechanical Properties of Sheathing” was conducted to evaluate the performance of laboratories operating in the electrical and electronics industry. In the present study, the data related to the tensile strength of cable sheathing were selected as a case study. Given the critical role of the assigned value and the standard deviation for proficiency assessment in the calculation of laboratory performance indicators, the impact of applying different statistical estimators on evaluation outcomes and participant performance classification was investigated. The objective of this study was to analyze the metrological implications of selecting different statistical estimation methods and to assess the sensitivity of proficiency testing results to these choices. The dataset comprised results from 50 participating laboratories and contained several outlying observations.

To evaluate the influence of outliers, the analyses were performed under two scenarios. In the first scenario, all reported results were included in the calculations. In the second scenario, two identified outlying observations were removed from the dataset. For both scenarios, the assigned value, standard deviation for proficiency assessment, and corresponding z-scores were calculated using four statistical methods: the arithmetic mean, median, Algorithm A, and the Hampel estimator. Subsequently,

the impact of estimator selection and outlier removal on laboratory performance classification was evaluated. Laboratory performance was assessed using the z-score defined in ISO 13528. In addition to the conventional classification categories, special attention was given to laboratories located near the decision limits ($|z| = 2$) and ($|z| = 3$)), since these results are the most sensitive to changes in statistical estimators and may lead to different conformity assessment decisions. The robustness of the estimators was evaluated by comparing changes in the assigned value, standard deviation for proficiency assessment, z-scores, and laboratory classifications before and after outlier removal. The stability of laboratory classifications across different estimators was also examined to assess the potential impact of estimator selection on decision-making in proficiency testing.

4. Results and discussion

4.1. Estimated values of assigned value and standard deviation

Figure 1 shows the histogram of the results reported by the laboratories before and after the outliers were removed. In the original form, the presence of a few outliers at the two ends of the distribution, especially on the right side of the graph, increased the range of variation in the data. Also, the distribution of the data shows asymmetry with respect to the ideal normal distribution. After the outliers were removed, the mean of the data and the standard deviation changed. In this case, the concentration of observations around the center of the distribution is greater and the shape of the distribution appears closer to the normal curve. In addition, the effect of outliers on the overall structure of the data is reduced and the dispersion of the data is mainly determined by the central results. Prior to the statistical analyses, the data were screened for outliers using Grubbs' test at a significance level of 0.05. The results of the test indicated that two observations were identified as outliers ($p < 0.05$); therefore, these values were removed from the dataset. Subsequent statistical analyses were conducted using the cleaned dataset.

Comparing the two histograms shows that a limited number of outliers can affect the location of the center of distribution and the overall shape of the data dispersion. This observation highlights the importance of using robust statistical methods in the analysis of proficiency test data, since such methods can control for the effect of outliers without directly removing information.

The classical mean and standard deviation were calculated using the arithmetic mean method without outliers. This method is valid only when the data are normal, have no outliers, and are unimodal and symmetric in distribution.

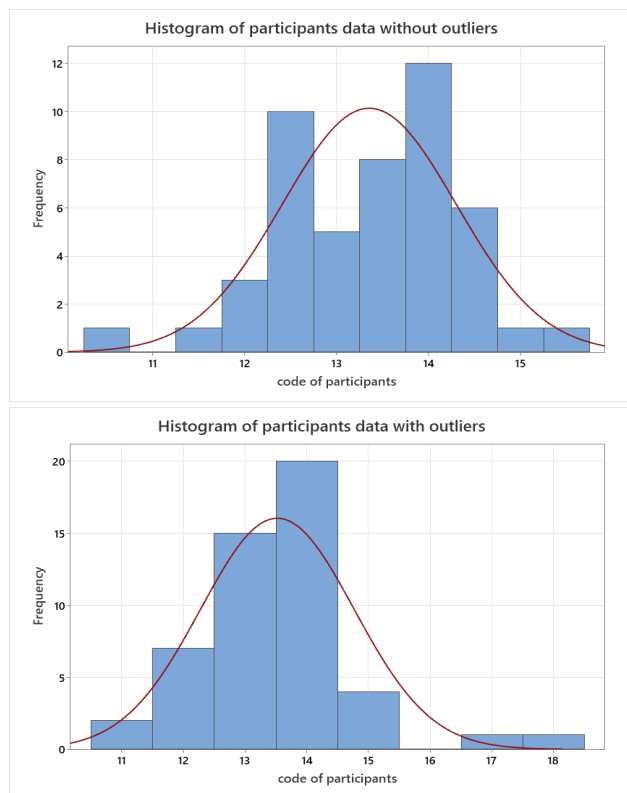


Fig 1. Histogram of participants data with and without outliers

Given that we have outliers in the data, we use different estimators to see their effect on the evaluation of laboratory performance. The results of calculating the assigned value and standard deviation of proficiency assessment for four statistical estimators, including the arithmetic mean, median, A-algorithm, and Hampel's estimator, are presented in Table 1 and 2. Table 1 (left side) shows the results from the entire data set of participants, while Table 1 (right side) shows the same calculations after removing two outlier observations.

Comparing the results of the two tables shows that the presence of only two outliers has a significant impact on some estimators. The largest change in the assigned value is related to the arithmetic mean method. The Assigned Value in this method has decreased from 13.5192 to 13.3531, which is equivalent to a shift of 0.166 units. This change is almost three times larger than the change observed in Algorithm A and more than seven times the change observed in the Hampel estimator. Such behavior shows that the arithmetic mean is very sensitive to the presence of extreme data and even a limited number of outlier observations can significantly shift the estimated center of the data.

Table 1. Estimated values of assigned value and standard deviation of participants' result without removing outliers and after removing two outliers

Method	Assigned Value	Standard Deviation
Mean	13.5192	1.187515
Algorithm A	13.4446	0.98579
Hampel	13.4021134	1.014996
Median	13.565	1.060

Method	Assigned Value	Standard Deviation
Mean	13.35308333	0.948411
Algorithm A	13.379	0.93205
Hampel	13.37904545	0.920919
Median	13.43	1.104835

In contrast, the Hampel estimator showed the least change in the assigned value. The difference in the Assigned Value before and after removing the outliers was only about 0.023 units. This result indicates that the limited effect function used in the Hampel method was able to effectively control the effect of the outliers. The median also showed a relatively stable behavior and the observed change in it was equal to 0.05 units, which is consistent with the theoretical expectation resulting from the breakpoint above the median.

Algorithm A was placed in a position between the median and the Hampel in terms of stability of the assigned value. The Assigned Value obtained by this method changed by about 0.066 units after removing the outliers. Although this value is greater than the change observed in the Hampel and median, it is much smaller compared to the arithmetic mean, indicating that the outlier restriction mechanism in Algorithm A has been able to significantly reduce the effect of extreme observations. The behavior of the estimators in estimating the dispersion also shows important differences. In the mean and standard deviation method, removing the two outliers reduced the standard deviation from 1.1875 to 0.948411. This change is equivalent to about a 20 percent decrease in the standard deviation value. Such a large difference indicates that a major part of the estimated dispersion in Table 1 was due to the presence of the same two outliers. From a statistical perspective, this result is an example of the phenomenon of variance inflation, in which a limited number of outliers artificially increase the dispersion of the entire data set. In contrast, Algorithm A showed only about 5.5% change in standard deviation. The standard deviation value de-

creased from 0.98579 to 0.93205, which is much smaller than the classical method. This result indicates that Algorithm A was able to control the effect of data on dispersion estimation without removing them. The Hampel estimator also showed relatively stable behavior, and the standard deviation reduction was about 10%. The behavior of the median in dispersion estimation was different from the other methods. Unlike the other three methods, removing outliers increased the standard deviation from 1.104835 to 1.060. This result indicates that in the data set under study, outliers are not necessarily located at the farthest points of the distribution, and their removal has changed the rank structure of the data in such a way that

the median-based dispersion estimate has increased. This observation indicates that the performance of robust estimators depends not only on the number of outliers, but also on their position in the distribution.

4.2. Comparing z-score values from different estimators

After examining the effect of outliers on the assigned value and standard deviation of the proficiency assessment, the next step is to analyze the effect of these changes on the main indicator of laboratory performance assessment, the z-score.

Table 3. Z-score values obtained from the entire data in statistical analyses

Mean	Median	Algorithm A	Hample	Code of lab	Mean	Median	Algorithm A	Hample
-0.580	-0.693	-0.6235	-0.564	1	-0.55154	-0.5431	-0.5890	-0.59619
-0.985	-1.146	-1.1104	-1.037	2	-1.05765	-0.9775	-1.1040	-1.11741
1.643	1.797	2.0546	2.037	3	2.232068	1.8464	2.2434	2.270508
0.060	0.024	0.1475	0.185	4	0.249804	0.1448	0.2264	0.22907
0.599	0.627	0.7967	0.816	5	0.924617	0.7241	0.9130	0.924027
0.253	0.240	0.3808	0.412	6	0.492315	0.3530	0.4732	0.47882
-0.875	-1.023	-0.9785	-0.908	7	-0.92058	-0.8599	-0.9645	-0.97625
-0.867	-1.014	-0.9684	-0.899	8	-0.91003	-0.8508	-0.9538	-0.96539
-0.723	-0.853	-0.7959	-0.731	9	-0.73078	-0.6969	-0.7714	-0.78079
0.119	0.090	0.2185	0.254	10	0.323612	0.2082	0.3015	0.305081
-0.336	-0.420	-0.3293	-0.278	11	-0.24576	-0.2806	-0.2779	-0.28129
0.573	0.599	0.7663	0.786	12	0.892985	0.6969	0.8809	0.891451
-0.218	-0.288	-0.1873	-0.140	13	-0.09815	-0.1539	-0.1277	-0.12927
1.003	1.080	1.2836	1.289	14	1.430727	1.1585	1.4280	1.445246
-1.094	-1.268	-1.2422	-1.165	15	-1.19472	-1.0952	-1.2435	-1.25857
-0.429	-0.523	-0.4409	-0.386	16	-0.36175	-0.3801	-0.3959	-0.40074
0.674	0.712	0.888	0.904	17	1.019513	0.8056	1.0096	1.021756
0.018	-0.024	0.0968	0.136	18	0.197084	0.0996	0.1727	0.174776
-0.353	-0.439	-0.3496	-0.298	19	-0.26685	-0.2987	-0.2993	-0.30301
-0.606	-0.721	-0.6539	-0.593	20	-0.58317	-0.5702	-0.6212	-0.62877
-1.237	-1.429	-1.4147	-1.332	21	-1.37397	-1.2491	-1.4259	-1.44317
-0.707	-0.835	-0.7756	-0.711	22	-0.7097	-0.6788	-0.7500	-0.75907
0.380	0.382	0.533	0.559	23	0.650474	0.4888	0.6341	0.641701
0.548	0.571	0.7359	0.757	24	0.861353	0.6698	0.8487	0.858875
-0.681	-0.806	-0.7452	-0.682	25	-0.67806	-0.6517	-0.7178	-0.7265
3.883	4.305	4.7529	4.658	26	5.03676	4.2540	5.0974	5.158926
-0.816	-0.957	-0.9075	-0.840	27	-0.84677	-0.7965	-0.8894	-0.90024
0.700	0.740	0.9185	0.934	28	1.051145	0.8327	1.0418	1.054332

Table 4. Z-score values after removing two outliers in statistical analyses

0.346	0.344	0.4924	0.520	29	0.608298	0.4526	0.5912	0.598266
0.287	0.278	0.4214	0.451	30	0.534491	0.3892	0.5161	0.522255
-1.136	-1.316	-1.293	-1.214	31	-1.24744	-1.1404	-1.2971	-1.31287
-0.976	-1.136	-1.1002	-1.027	32	-1.0471	-0.9685	-1.0933	-1.10655
0.346	0.344	0.4924	0.520	33	0.608298	0.4526	0.5912	0.598266
-0.791	-0.929	-0.8771	-0.810	34	-0.81514	-0.7693	-0.8573	-0.86766
0.455	0.467	0.6243	0.648	35	0.74537	0.5702	0.7306	0.739429
0.413	0.420	0.5736	0.599	36	0.69265	0.5250	0.6770	0.685136
-0.749	-0.882	-0.8263	-0.761	37	-0.76242	-0.7241	-0.8036	-0.81337
-0.195	-0.261	-0.1588	-0.112	38	-0.06862	-0.1285	-0.0976	-0.09886
0.127	0.099	0.2287	0.264	39	0.334156	0.2172	0.3122	0.315939
-1.827	-2.089	-2.1248	-2.022	40	-2.11204	-1.8826	-2.1769	-2.20328
-2.458	-2.796	-2.8856	-2.761	41	-2.90284	-2.5615	-2.9816	-3.01769
-0.176	-0.240	-0.1365	-0.091	42	-0.04543	-0.1086	-0.0740	-0.07497
0.287	0.278	0.4214	0.451	43	0.534491	0.3892	0.5161	0.522255
1.121	1.212	1.4257	1.426	44	1.578342	1.2853	1.5782	1.597268
0.860	0.920	1.1112	1.121	45	1.25148	1.0047	1.2456	1.260647
0.220	0.203	0.3402	0.372	46	0.450139	0.3168	0.4302	0.435385
-0.168	-0.231	-0.1264	-0.081	47	-0.03488	-0.0996	-0.0633	-0.06412
2.830	3.126	3.4849	3.427	48	3.718765	3.1226	3.7562	3.801587
0.615	0.646	0.817	0.835	49	0.945705	0.7422	0.9345	0.945745
0.624	0.655	0.8272	0.845	50	0.956249	0.7512	0.9452	0.956603

A comparison of the two tables (3 and 4) indicates that outliers have a substantial impact on z-scores. This occurs because the z-score is simultaneously a function of two quantities: the assigned value and the standard deviation of participants' results. Any change in either of these parameters is directly reflected in the z-score.

In the classical method, the presence of two outliers caused a considerable increase in the standard deviation of proficiency assessment results. This increase enlarged the denominator of the z-score equation, thereby reducing the absolute values of the z-scores. From a statistical perspective, this phenomenon implies a reduction in the sensitivity of the evaluation system, because results that deviate substantially from the center of the data are less accurately represented due to the inflated value of σ_{pt} . After removal of the outliers, σ_{pt} in the mean-based method decreased by approximately 20%. The direct consequence of this reduction was an increase in the absolute values of the z-scores.

In robust statistics, this phenomenon is known as the masking effect. In such situations, a few extreme observations artificially inflate the dispersion, causing other abnormal results to be concealed within the enlarged variability. In contrast, the z-scores obtained using Algorithm

A exhibited only minor changes. The reason is that the influence of outliers is controlled from the outset through an iterative weighting procedure. Consequently, the physical removal of outliers has little effect on the final results, and the z-scores before and after outlier removal remain largely unchanged.

The Hampel estimator showed behavior similar to that of Algorithm A, although slightly greater changes were observed for some laboratories. Nevertheless, both robust methods demonstrated substantially greater stability than the classical approach when outliers were removed. The median-based method exhibited the smallest changes in z-scores. This finding is consistent with the theoretical foundations of robust statistics, since the median has a breakdown point of 50%, and the presence or removal of a limited number of outliers cannot substantially alter the center of the distribution.

4.3. Effect of Estimator Selection on Decision-Making and Laboratory Performance Classification

Table 5 presents the laboratory performance evaluation results according to the z-score acceptance limits defined previously. The results indicate that the four estimators differ not only in terms of the Assigned Value and Stan-

dard Deviation for Proficiency Assessment, but also in their ultimate impact on decision-making outcomes.

Table 5. Classification Transition Matrices for Mean, Algorithm A, Hampel, and Median Estimators Before and After Outlier Removal (S=Satisfactory, W=Warning, A=Action)

Arithmetic Mean			
From / To	S	W	A
S	45	2	0
W	0	2	1
A	0	0	1

Algorithm A			
From / To	S	W	A
S	45	0	0
W	0	3	0
A	0	0	2

Median			
From / To	S	W	A
S	46	1	0
W	1	1	0
A	0	0	0

Hampel			
From / To	S	W	A
S	45	0	0
W	0	2	1
A	0	0	2

In the presence of outliers, the Mean method produced the highest number of satisfactory results (47 cases) and the lowest number of action results (1 case). This finding is consistent with the observations presented in the previous tables, where outliers inflated the classical standard deviation and consequently reduced the absolute values of the z-scores. As a result, some borderline results that were classified as warning or action by other methods remained within the satisfactory category when the Mean estimator was used. Algorithm A and the Hampel estimator produced identical classifications for the raw dataset (45 satisfactory, 3 warning, and 2 action results). However, their behavior diverged after the removal of outliers. The arithmetic mean exhibited the greatest classification instability, resulting in changes in the performance status of 2 out of 50 laboratories (4%) following outlier removal.

The most notable finding in Table 5 concerns the differ-

ence between Algorithm A and the Hampel estimator. Following outlier removal, Algorithm A produced no change in performance classification, whereas Hampel shifted one laboratory from the warning category to the action category. This behavior suggests that Hampel is more sensitive to borderline observations and adopts a slightly more conservative approach near the decision threshold ($|z| = 3$). From the perspective of robust statistics, this result is consistent with the theoretical properties of the Hampel estimator. Its bounded influence function and higher resistance to extreme observations cause the location and dispersion estimates to be less affected by outlying values. Consequently, for some borderline observations, larger z-scores are obtained, increasing the likelihood of exceeding the action limit. Therefore, within this dataset, Hampel can be considered more stringent than Algorithm A. This behavior indicates that Algorithm A is capable of managing information from extreme observations without requiring their exclusion.

Although outlier removal is generally expected to increase the number of unsatisfactory results, the opposite trend was observed for the Median method. The number of satisfactory results increased from 46 to 47, while the number of warning results decreased from 2 to 1, and the number of action results remained unchanged. This finding indicates that, in this dataset, the removal of outliers altered the median and dispersion estimates in such a way that one borderline result moved from the warning category into the satisfactory category. This behavior illustrates that high robustness does not necessarily imply greater stringency. In some situations, the changes induced by outlier removal may act in the opposite direction and lead to a more lenient evaluation.

From a metrological perspective, the most important finding is that the measurement data themselves remain unchanged, while the final decisions differ and the numbers of warning and action classifications vary among estimators. This demonstrates that the choice of statistical estimator forms part of the conformity assessment process and can directly influence the final evaluation outcome. These differences become particularly evident near the ISO 13528 decision limits ($|z| = 2$ and $|z| = 3$). In the present study, the Mean estimator tended to provide more optimistic classifications, Hampel adopted a slightly more conservative classification strategy, Algorithm A exhibited the greatest stability against outlier removal, and the Median estimator demonstrated an intermediate yet distinct behavior compared with the other approaches. Overall, the results show that differences among statistical estimators are not limited to Assigned Values or Standard Deviations. Rather, they can lead to changes in the number of laboratories classified within the warning and

action categories. In this study, Algorithm A demonstrated the highest stability with respect to outlier removal, whereas Hampel adopted a slightly more conservative approach near the action threshold, and the Median estimator occasionally resulted in more lenient classifications. These findings indicate that the selection of a statistical estimator may constitute an important component of decision uncertainty in proficiency testing.

5. Conclusion

This study investigated the influence of statistical estimator selection on laboratory performance classifications using a proficiency testing dataset of scheme on “Wires and Cables. In the present study, the data related to the tensile strength of cable sheathing were selected as a case study. Four approaches - arithmetic mean, median, Algorithm A, and the Hampel estimator - were compared, and the effect of outlier removal on the results obtained from each method was evaluated. The findings showed that the arithmetic mean was the most sensitive estimator to outliers. The presence of only a small number of outlying observations resulted in substantial changes in the statistical parameters. These changes ultimately reduced the absolute values of certain z-scores and increased the number of laboratories classified within the satisfactory category. The arithmetic mean exhibited the greatest classification instability, resulting in changes in the performance status of 2 out of 50 laboratories (4%) following outlier removal. Comparison of the results before and after outlier removal demonstrated that Algorithm A provided the highest level of stability in laboratory performance classification. The numbers of satisfactory, warning, and action results remained unchanged under both scenarios, indicating that Algorithm A was able to effectively control the influence of outliers without requiring their explicit removal. In contrast, although the Hampel estimator exhibited strong resistance to outliers, it showed a slightly more conservative behavior near decision thresholds. Following outlier removal, one laboratory was reclassified from the warning category to the action category. The Median estimator, on the other hand, produced more lenient classifications in certain cases and placed a larger number of laboratories within the satisfactory category.

A key outcome of this study is that differences among estimators extend beyond statistical parameter estimates and can lead to different decisions regarding laboratory performance. When laboratory results lie close to decision thresholds, the choice of statistical estimator may determine whether a laboratory is classified as satisfactory, warning, or action. Consequently, estimator selection should be regarded as a contributor to decision uncertainty in proficiency testing. From a metrological perspective,

the results demonstrate that the selected statistical estimator constitutes an integral part of the measurement chain and conformity assessment process. Therefore, the choice of estimator may have significant technical and managerial implications for the interpretation of proficiency testing outcomes.

Conflict of Interest

The authors declare that there are no conflicts of interest associated with this study.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: Jun. 2026 Accepted: Aug. 2026

Published online: Aug. 2026

DOI: 10.22034/ASAS.2026.585565.1109

References

- [1] ISO/IEC 17043:2023, Conformity assessment — General requirements for the competence of proficiency testing providers, Published (Edition 2).
- [2] Thompson, M., Ellison, SLR., Wood, R., (2006) The international harmonized protocol for the proficiency testing of analytical chemistry laboratories. *Pure Appl Chem* 78(1):145-196.
- [3] Iglesias-Gonzalez et al (2021) Proficiency test in mt-respiration: A necessary tool for reliable and reproducible results, *MitoFit Preprints*.
- [4] M Thompson, SLR Ellison, 1.02 - Proficiency Testing in Analytical Chemistry, Editor(s): Steven Brown, Romà Tauler, Beata Walczak, *Comprehensive Chemometrics* (Second Edition), Elsevier, 2020, Pages 53-69.
- [5] ISO 13528:2022, Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparison, Published (Edition 3).
- [6] Cofino, W.P., Crum, S., van Vark, W. et al. Robustness comparison of three statistical methods commonly used in Proficiency Tests. *Accred Qual Assur* 30, 507–519 (2025). <https://doi.org/10.1007/s00769-025-01669-3>
- [7] Tsamatsoulis, D., (2022). Comparing the Robustness of Statistical Estimators of Proficiency Testing Schemes for a Limited Number of Participants. *Computation*, 10(3), 44.

- [8] Cordeiro, F., Emons, H. & Robouch, P. Is the z score sufficient to assess participants' performance in proficiency testing? The hidden corrective action. *Accred Qual Assur* 27, 145–153 (2022). <https://doi.org/10.1007/s00769-022-01496-w>
- [9] Hogg, R. V., McKean, J., & Craig, A. T., (2019), *Introduction to Mathematical Statistics*, 8th edition, Pearson.
- [10] Rousseeuw, P. J., & Leroy, A. M., (1987) *Robust Regression and Outlier Detection.*, John Wiley & Sons.
- [11] Hubert, M. and Debruyne, M. (2010), Minimum covariance determinant. *WIREs Comp Stat*, 2: 36-43.
- [12] Hampel, F., Ronchetti, E., Rousseeuw, P. & Stahel, W., (1986), *Robust Statistics: The Approach Based on Influence Functions.* John Wiley & Sons.

Submit your manuscript to *Advances in the standards and applied sciences* journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open Access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

**Submit your next manuscript at:
journal.standards.ac.ir**