Research on analytical methods for defect measurement systems in gas polyethylene pipe welded

Chenjia Zong[1], *Juan* Zhou[1*], *Qiang* Wang[2], *Zhilong* Yu[1], *Haiting* Zhou[1], *Yun* Song[1], and *Jiayan* Chen[1]

[1]College of Quality and Standardization, China Jiliang University, Hangzhou 310018, China.
[2]College of Energy Environment and Safety Engineering and College of Carbon Metrology, China Jiliang University, Hangzhou 310018, China.

Abstract. Welded defects in gas polyethylene pipelines exhibit low occurrence frequency and high variability, and the traditional attribute measurement analysis method fails to assess their performance comprehensively. A two-tiered progressive attribute measurement analysis model is proposed. Initially, key statistical measures for assessment agreement (P_a, Cohen's Kappa, and Fleiss' Kappa) were systematically evaluated to identify those suitable for various analysis tiers. Subsequently, the application of Cohen's Kappa and Fleiss' Kappa across varying levels of agreement was examined, and thresholds for sample size selection were established. Lastly, an attribute measurement analysis model for gas polyethylene pipeline welded defects is developed: the first tier is based on preliminary assessment and data filtering regarding the presence of welded defects; the second tier employs weighted kappa analysis to quantitatively assess agreement on specific defect types, This results in a comprehensive evaluation of the performance of the measurement system.

1 Introduction

The integrity of welded joints in polyethene (PE) gas pipelines constitutes a critical determinant of system reliability[1]. In contrast, welded defects, including Cold Weld (CW), Porosity (POR), and Under Plug Insertion (UPI), can potentially result in severe safety incidents[2-4]. Consequently, ensuring the reliability of weld quality measurement systems becomes imperative. Measurement System Analysis (MSA) is vital for evaluating measurement system reliability[5]. While conventional metrological MSA applies to continuous variables, its effectiveness in analyzing attribute data of weld defects remains constrained, necessitating the exploration of attribute-type MSA methods for weld defect classification and assessment[6].

Typical defects in welded joints of gas polyethene pipelines, including Cold Weld (CW), Porosity (POR), Inclusion in Fusion Face (IFC), Structural Distortion (SD), Over Weld (OW), Short Weld Length (SWL), and Under Plug Insertion (UPI), can markedly diminish the mechanical properties and sealing integrity of the joints[7]. The industry predominantly employs Visual Inspection (VT), Ultrasonic Testing (UT), and X-ray Radiographic Testing (XRT) for defect identification[8-10]. Visual Inspection (VT) exhibits a low detection rate[11], Ultrasonic Testing (UT) necessitates highly skilled operators[12], and X-ray Radiographic Testing (XRT) equipment is expensive and requires stringent radiation protection measures[13]. However, subjectivity in operator experience can result in inconsistencies within inspection results[14]; for instance, Xiang Rong et al. reported that the Kappa coefficient of determination agreement for closed tomato identification by various operators within the same machine vision image was merely 0.37 (subthreshold at 0.4), highlighting significant inter-operator variability within the inspection system[15]. This phenomenon is notably pronounced in gas pipeline engineering, where the random error and bias within the measurement system are exacerbated by ambiguous determination criteria for welded defects and the absence of standardization in the inspection process. Consequently, while the aforementioned inspection methods can identify physical defects, their outcomes heavily rely on the operator's subjective experience, necessitating an urgent enhancement of the agreement and reliability of inspection results through the execution of attribute measurement system analysis[16].

Although a foundational framework for attribute-based Measurement System Analysis (MSA) has been established, it still encounters challenges in conducting hierarchical analyses of multi-category defects in gas pipeline welded joints and statistical modelling under special operating conditions. The attribute measurement analysis process delineated in the AIAG MSA Handbook (4th edition), featuring key metrics like the assessment agreement rate P_a, predominantly caters to binary determination and is challenging to apply directly to complex multi-category defect scenarios[17]. Cohen's and Fleiss' Kappa coefficients are extensively employed to measure assessment agreement[18-19]. Yet, research predominantly centers on the numerical computation of individual statistics, neglecting the integrated visual analysis of multiple metrics, which impacts the rational

* Corresponding author: zhoujuan@cjlu.edu.cn

selection of statistical measures. Although the attribute-based MSA has demonstrated success in automotive manufacturing[20] and medical diagnostics[21], the model lacks universal applicability and fails to address the statistical bias issues stemming from the uneven distribution of small samples and multi-category defects in gas pipeline inspections. More critically, the current MSA methodology employs a single-tier data recording format, which cannot accommodate the multi-tiered judgment logic required for interpreting welded defects in gas pipelines, resulting in data collection redundancy and diminished analysis precision.

While conventional attribute-type measurement system analysis (MSA) exhibits three critical limitations—inadequate statistical characterization, absence of systematic sample size selection criteria, and insufficient analytical capability for hierarchical defects—this study introduces a two-tier progressive framework specifically designed for evaluating defect detection systems in gas polyethylene pipeline welds. The methodological advancements comprise the following key innovations:

(1) Given the ambiguous trends observed in Cohen's Kappa, Fleiss' Kappa, and the assessment agreement rate (P_a), we develop an interactive visualization platform to systematically quantify the dynamic responses of these metrics under two critical scenarios: agreement-level enhancement (via operator training optimization) and defect category granularity refinement. This framework establishes a principled foundation for selecting optimal statistical measures in gas polyethylene pipeline weld defect evaluation systems.

(2) Addressing the issue of sample size selection based on empirical thresholds, a sample size decision-making framework that leverages the decay law of confidence intervals is proposed. This framework features comparative multi-statistic analysis and trend visualization modeling. The convergence curve of the sample size-confidence interval is constructed by comparing the decay trajectories of Cohen's Kappa and Fleiss' Kappa confidence interval widths, thereby clarifying the critical threshold for sample size selection.

(3) Considering the two-tiered identification of welded defects in gas polyethene pipelines, a hierarchical data recording and two-stage analysis framework is suggested. The data recording employs a two-tiered structure: the first tier involves a binary assessment of "defect presence", and the second tier encompasses a combination of "defect type" and "risk level" categorization. The analysis process is bifurcated into two stages: Stage 1 calculates the intra-operator and inter-operator assessment agreement and filters the compliant data; Stage 2 constructs a hierarchical, weighted Kappa model and calculates the assessment agreement ratios for the four categories. The framework enhances the entire process, from data collection to analysis and decision-making, and offers a systematic methodology to facilitate the assessment of gas pipeline welded quality.

2 Comparative analysis of agreement evaluation indicators

2.1 Basic agreement statistics

2.1.1 Assessment agreement rate P_a

Assessment agreement rate quantifies the alignment level of multiple assessment outcomes within a single measurement system (intra-system) or between distinct systems (inter-system). This metric is defined as the ratio of the sample size exhibiting complete concordance across all measurement results to the total sample size evaluated.

$$P_a = \frac{m}{n} \tag{1}$$

In this formulation, n denotes the total sample size, while m represents the number of samples demonstrating consensus measurement outcomes.

Employing an adapted Clopper-Pearson exact approach, the 95% confidence bounds for the assessment agreement rate (P_a) are determined through the following computational expressions:

$$CI_L = \left[1 + \frac{n-m+1}{m \cdot F_{2m,(n-m+1)}(0.025)}\right]^{-1} \tag{2}$$

$$CI_U = \left[1 + \frac{n-m}{(m+1) \cdot F_{2(m+1),2(n-m)}(0.975)}\right]^{-1} \tag{3}$$

Under boundary conditions: When P_a = 0, the lower confidence bound CI_L is set to 0; when P_a = 1, the upper confidence bound CI_U defaults to 1. The F-distribution quantiles are adjusted for these extremities: the 0.025 quantile is substituted with 0.05 when P_a = 1, and the 0.975 quantile is replaced with 0.95 when P_a = 0. The notation $F_{v_1,v_2}(\alpha)$ denotes the α level critical value of the F-distribution with v_1 and v_2 degrees of freedom.

2.1.2 Cohen's Kappa statistic

Cohen's Kapp[22-23] is a statistical metric that evaluates inter-rater agreement between two independent measurement systems (or repeated measurements within a single system) for categorical assignments while accounting for chance agreement. The metric is computed using an $a \times a$ contingency matrix, where p_{ij} denotes the joint probability that measurement A is classified into category i. In contrast, measurement B is classified into category j, with $p_{i.}$ and $p_{.i}$ represent the row-wise and column-wise marginal probabilities, respectively. Both the global Kappa coefficient and category-specific variants, along with their respective standard errors (SE), are computed via the following computational framework:

$$K_{Overall}^{Cohen} = \frac{P_a - P_c}{1-P_c} = \frac{\sum_{i=1}^{a} P_{ii} - \sum_{i=1}^{a} P_{i.}P_{.i}}{1-\sum_{i=1}^{a} P_{i.}P_{.i}} \tag{4}$$

$$SE(K_{Overall}^{Cohen}) = \frac{\sqrt{P_c + {P_c}^2 - \sum_{i=1}^{a} P_{i.}P_{.i}(P_{i.}+P_{.i})}}{1-P_c\sqrt{n}} \tag{5}$$

$$K_k^{Cohen} = \frac{P_a - P_c}{1-P_c} = \frac{p_{11}+p_{22}-p_{1.}\times p_{.1}-p_{2.}\times p_{.2}}{1-p_{1.}\times p_{.1}-p_{2.}\times p_{.2}} \tag{6}$$

$$SE(K_k^{Cohen}) = \frac{\sqrt{P_c + P_c^2 - \sum_{i=1}^{2} p_{i.}p_{.i}(p_{i.}+p_{.i})}}{(1-P_c)\sqrt{n}} \tag{7}$$

2.1.3 Fleiss' kappa statistic

Fleiss' Kappa[24-25] serves as a statistical framework for quantifying inter-annotator agreement in multi-rater assessments. It is applicable to cross-system evaluations and repeated measurement trials involving identical sample sets. For the overall and each subcategory Kappa statistics and their standard errors, see the formula:

$$K_{Overall}^{Fleiss} = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{k=1}^{a} x_{ik}^2}{nm(m-1)\sum_{k=1}^{a}\bar{p}_k\bar{q}_k} \quad (8)$$

$$SE\left(K_{Overall}^{Fleiss}\right) = \frac{\sqrt{2}\sqrt{\left(\sum_{k=1}^{a}\bar{p}_k\bar{q}_k\right)^2 - \sum_{k=1}^{a}\bar{p}_k\bar{q}_k(\bar{q}_k - \bar{p}_k)}}{\sum_{k=1}^{a}\bar{p}_k\bar{q}_k\sqrt{nm(m-1)}} \quad (9)$$

$$K_k^{Fleiss} = 1 - \frac{\sum_{i=1}^{n} x_{ik}(m - x_{ik})}{nm(m-1)\bar{p}_k\bar{q}_k} \quad (10)$$

$$SE\left(K_k^{Fleiss}\right) = \sqrt{\frac{2}{nm(m-1)}} \quad (11)$$

The statistical community has established quantitative interpretations of Kappa coefficients to benchmark inter-rater reliability levels, forming an established benchmarking framework across empirical research domains. Our investigation operationalizes the Fleiss-recommended reliability classification system for agreement analysis, with categorical thresholds formally specified in Table 1.

Table 1 Kappa statistic evaluation criteria.

Kappa statistic	Agreement strength
＜0.40	Very poor
[0.40,0.75]	Fair
＞0.75	Very good

2.2 weighted kappa statistic

The weighted kappa statistic[26-27] becomes methodologically imperative for ordinal classification schemes in attribute measurement analysis. This methodological advancement introduces differential weighting matrices to quantify inter-rater disagreement patterns, enhancing sensitivity to ordinal scale gradations. The calculation formula is as follows:

$$k_\omega = \frac{p_o^* - p_e^*}{1 - p_e^*} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{a} w_{ij}p_{ij} - \sum_{i=1}^{a}\sum_{j=1}^{a} w_{ij}p_{i.}p_{.j}}{1 - \sum_{i=1}^{a}\sum_{j=1}^{a} w_{ij}p_{i.}p_{.j}} \quad (12)$$

The ordinal disagreement weight ω_{ij} maintains computational congruence with conventional Kappa calculations while differentially scaling inter-rater discordances. The specific calculation method of the weight parameter is as follows: There are mainly the following two methods:

(1) Linear weights proposed by Cicchetti and Allison:

$$\omega_{ij} = 1 - \frac{|C_i - C_j|}{C_c - C_1} \quad (13)$$

(2) Squared weights as proposed by Fleiss and Cohen:

$$\omega_{ij} = 1 - \frac{(C_i - C_j)^2}{(C_c - C_1)^2} \quad (14)$$

where C_c is the maximum number of categories, C_1 is the minimum number of categories, and C_i, C_j are the numbers of categories for the two researches, respectively.

2.3 Comparative study of agreement evaluation indicators

Cohen's Kappa, Fleiss' Kappa, and the assessment agreement rate (P_a) are commonly used indicators to evaluate the agreement of classification. However, the relationship between their changing trends and data distribution characteristics (such as the number of distinguishable categories and the frequency of agreement/disagreement) has not been systematically studied. This may lead to a lack of theoretical basis when selecting statistics for practical applications. Through numerical simulation experiments, the response trends of the three types of statistics under the three scenarios of the number of distinguishable categories, the increase in the frequency of agreement, and the increase in the frequency of disagreement were systematically compared to reveal differences in their mathematical properties and provide suggestions for the selection of statistics for different research objectives.

(1) Fix the number of distinguishable categories (a=2, a=3), and gradually increase the number of agreement frequencies m_{11} or disagreement frequencies m_{21}

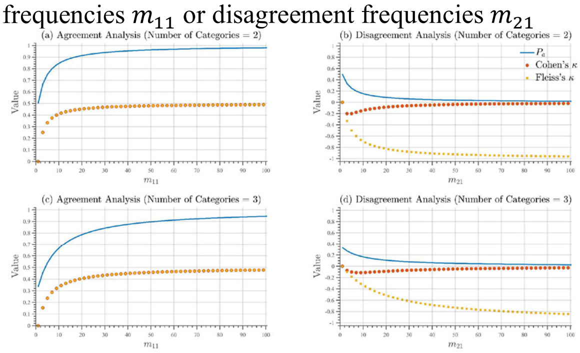

Fig. 1 The response potential of the statistic to changes in the frequency of agreement and disagreement events in two- and three-category scenarios.

The effect of increasing the agreement frequency m_{11} (and the disagreement frequency m_{21}) from 1 to 101 in increments of 2 on Cohen's Kappa, Fleiss' Kappa and P_a was analyzed for fixed category numbers of 2 and 3. The results show that P_a is very sensitive to changes in frequency in the dichotomous and trichotomous scenarios but may overestimate or underestimate the level of agreement. Cohen's Kappa is sensitive at low levels of agreement but tends to stabilize at high levels. In contrast, Fleiss' Kappa shows greater robustness in the multiclassification scenario and is more suitable for handling data from multiple raters. The results of the relevant visual analysis are shown in Fig. 1.

(2) Synergistic adjustment can increase the number of distinguishable categories a and the number of agreement frequencies (disagreement frequencies).

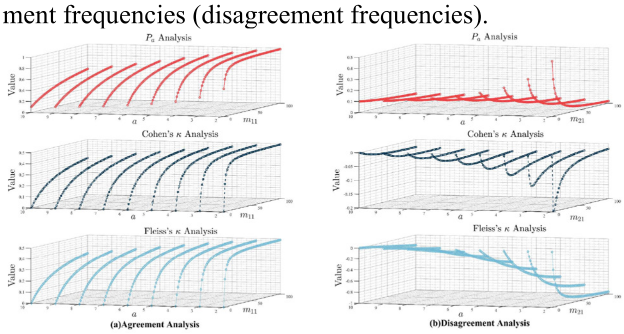

Fig. 2 Three-dimensional visualization of the response of a statistic to changes in the frequency of agreement and disagreement frequencies in a multi-category scenario.

As depicted in Fig. 2, the sensitivity of the assessment agreement rate (P_a) to fluctuations in agreement and welding tends to diminish as the number of distinguishable categories increases. Cohen's Kappa exhibits marked sensitivity to frequency alterations under conditions of low category dimensionality, yet this sensitivity diminishes rapidly and levels off as category dimensionality augments. In contrast, Fleiss' Kappa manifests a agreement pattern of variation across both low and high-category dimensionalities. This attribute suggests that Fleiss' Kappa can more accurately reflect the proper level of inter-rater agreement when managing multi-rater assessments.

A two-tier gas polyethene pipe welded defect identification analysis proposes a statistical measure optimisation strategy. The first level deals with binary classification problems (pass/fail). Cohen's Kappa was chosen as the primary metric due to its sensitivity to frequency changes with a small number of categories, and it can effectively avoid the risk of overestimation when there is extreme agreement. The second level involves multi-classification tasks (holes, cracks, lack of fusion, etc.), for which Fleiss' Kappa is employed as the primary metric because of its robustness in multi-category scenarios and low dependence on the number of raters, making it suitable for quantifying the reliability of multi-rater measurement systems.

3 Model building

3.1 Analysis model of a two-tier progressive at tribute measurement system

Gas polyethylene pipe welded joints are susceptible to process parameters during the forming process, resulting in seven typical defects: Cold Weld (CW), Porosity (POR), and Inclusion in Fusion Face (IFC). A two-tier progressive measurement system analysis model was developed. The model encompasses three principal components: data acquisition and pre-processing, binary defect assessment agreement analysis, and weighted agreement analysis for multiple defect types. Phased array ultrasonic scanning data is standardized and processed to construct a two-tier data structure comprising defect presence (binary variable) and type attributes (multi-class variable). Then, Cohen's Kappa is utilized at the primary tier to assess the intra- and inter-operator binary discrimination agreement. When $k < 0.6$, analysis and enhancement are conducted; ultimately, the samples that pass the initial screening undergo secondary analysis, and a weighted Kappa coefficient is computed based on the risk level. Conducting stratified diagnosis of the agreement in detecting seven defects and devising targeted measurement system enhancement plans based on failure modes. The overall model flowchart is depicted in Fig. 3.

(1) Data acquisition and pre-processing

This module is responsible for collecting and pre-processing the raw data to detect welded defects in gas polyethene pipes based on the requirements of the measurement system analysis. First, the measurement plan is determined based on the requirements of the measurement system analysis. Then, the operator employs ultrasonic phased array technology to detect internal defects and collect data. The collected data needs to be scrutinized for sample representativeness and randomness to ensure the validity and reliability of the data. For incomplete samples, re-collection is necessary to ensure comprehensiveness. Eligible samples will be subjected to data cleaning and preliminary analysis to enhance data quality and be recorded according to a two-tiered standard: the first tier records the presence of defects, and the second tier records the type of defect.

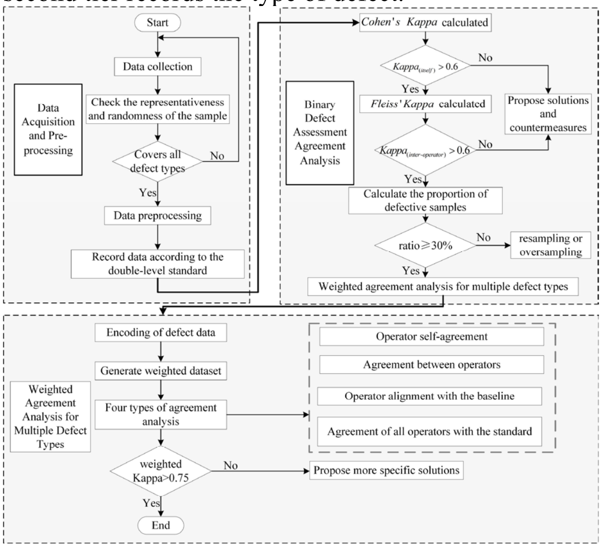

Fig. 3 Two-tier progressive attribute measurement system analysis flowchart.

(2) Binary defect assessment agreement analysis

The task of this module is to analyze the agreement of the presence of defects and to evaluate the performance of the optimized measurement system preliminarily. First, the agreement of each operator is evaluated using Cohen's Kappa statistic (refer to Equations (4) and (6)), and a Kappa value greater than 0.6 is considered acceptable. Improvement measures need to be proposed if this value is below the threshold. Second, Fleiss' Kappa is used to assess agreement among different operators (refer to Equations (8) and (10)), again with a threshold of 0.6. The analysis results will guide data filtering to ensure the quality of the data set. Following filtering, the proportion of defect types is calculated. If the proportion is below 30%, resampling or optimization strategies are considered; the analysis continues if it is 30% or more. These analyses provide a preliminary assessment of the measurement system's performance and propose optimization measures to improve accuracy and reliability.

(3) Weighted agreement analysis for multiple defect types

This module is based on a comprehensive data analysis of specific defect types. First, the defect types and risk levels of gas polyethene pipe welding are categorized and quantified. Then, the dataset is assigned weights using the squared weight weighting method (see Equation (14)) to reflect the impact of different defect types on the measurement system's performance. Based on the weighted dataset, a weighted Fleiss' Kappa concordance analysis is performed, including operator self-concordance, operator-operator concordance, operator-benchmark concordance, and overall concordance. If the

 https://doi.org/10.1051/matecconf/202541301003

concordance is below the 0.75 threshold, specific remedial actions, such as training and process adjustments, are developed based on the concordance category and defect type. This detailed analysis provides a scientific basis for quality control and process improvement.

3.2 Sample size selection for attribute measure ment systems

Minitab and the AIAG MSA reference manual specify four agreement ratios when analyzing attribute measurement systems. However, these definitions are notably stringent regarding agreement and do not offer explicit recommendations for sample size determination. In contemporary practice, most researchers adhere to conventional empirical thresholds, such as $n \geqslant 30$, when determining sample sizes. To a certain degree, this approach can impact the precision of agreement analysis and the dependability of resulting judgments.

To tackle this challenge, the focus is on applying Cohen's Kappa statistic and Fleiss's Kappa statistic—particularly when involving three operators—at varying levels of agreement. By examining the trend in the width of the confidence interval with increasing sample size, the goal is to establish a cost-effective sample size threshold.

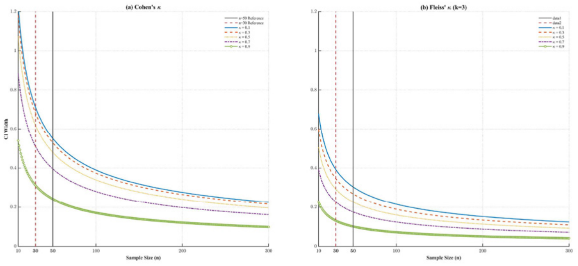

Fig. 4 The trend of the width of the confidence interval for the Cohen Kappa and Fleiss Kappa statistics for different sample sizes.

Fig. 4 shows the trends of the widths of the confidence intervals for Cohen’s Kappa and Fleiss’ Kappa statistics decreasing with increasing sample size for different Kappa values, which indicates that a large sample size can improve the accuracy of agreement assessment. Fleiss’ Kappa trend is more stable than Cohen’s, with a minor decrease, which suggests that it is more robust when assessing agreement in a multi-rater setting.

To further quantify the effect of varying sample sizes on the width of the confidence interval, a detailed examination of the changes in confidence interval width (CI Width) and standard error (SE) across different sample sizes will be conducted to ascertain an appropriate sample size threshold. A sample size of 50 will be chosen for comparative analysis against the conventional empirical threshold of 30.

Table 2 Width of confidence interval and standard error for different Kappa values and sample sizes.

Kappa	Sample size	Cohen’s Kappa (CI Width, SE)	Fleiss’ Kappa (CI Width, SE)
0.1	30	0.712, 0.1817	0.392,0.1000
	50	0.552, 0.1407	0.304,0.0775
0.3	30	0.683, 0.1742	0.346,0.0882
	50	0.529, 0.1349	0.268,0.0683
0.5	30	0.620, 0.1581	0.292,0.0745
	50	0.480, 0.1225	0.226,0.0577
0.7	30	0.511, 0.1304	0.226,0.0577
	50	0.396, 0.1010	0.175,0.0447
0.9	30	0.312, 0.0796	0.131,0.0333
	50	0.242, 0.0616	0.101,0.0258

The quantitative analysis in Table 2 shows that when the sample size is increased from 30 to 50, the widths of the confidence intervals for Cohen’s Kappa and Fleiss’ Kappa are reduced by an average of 22.51% to 22.64%. The standard errors are also reduced by an average of 22.56% to 22.64%, significantly improving the accuracy of the estimates. In particular, the reductions in the widths of the confidence intervals were most significant at a moderate level of agreement (Kappa values 0.3 to 0.5), reaching 25.4%. In conclusion, a sample size of 50 was selected.

4 Experimental verification

4.1 Experimental design

In this experiment, 50 gas polyethylene pipes from a single welding batch were utilized as the test subjects, employing random sampling to ensure a representative distribution of the samples. To detect defects within the welded joints, a high-precision ultrasonic phased array detector, characterized by a resolution of ±0.1 mm and a frequency range of 5-10 MHz (refer to Fig. 5), was employed to conduct axial and circumferential scans of each joint in two dimensions. The internal structure was subsequently analyzed using Full Matrix Capture (FMC) and Synthetic Aperture Focusing Technique (SAFT) to ascertain the nature of the defects.

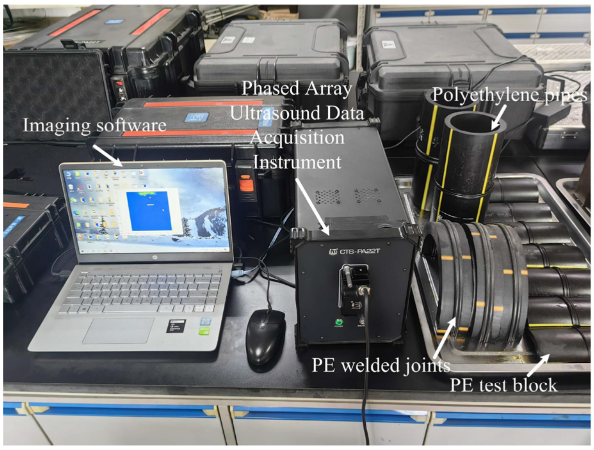

Fig. 5 Ultrasonic phased array data acquisition and detection device.

To ascertain the safety and reliability of the welded joints, a thermal-oxidative aging test was performed on the gas polyethylene pipes, as depicted in Fig. 6. The test simulates the thermal and oxidative stresses encountered during extended use, assessing the effects of these stresses on the pipes' performance. Utilizing the outcomes of the thermal-oxidative aging test and adhering to industry standards, a systematic summary of the predominant types of defects in the welded joints of gas polyethylene pipes and their associated hazards is presented in Table 3 Types of defects in gas polyethylene pipe welded joints and their potential hazards. The safety risks are quantified according to hazard levels ranging from 1 to 7.

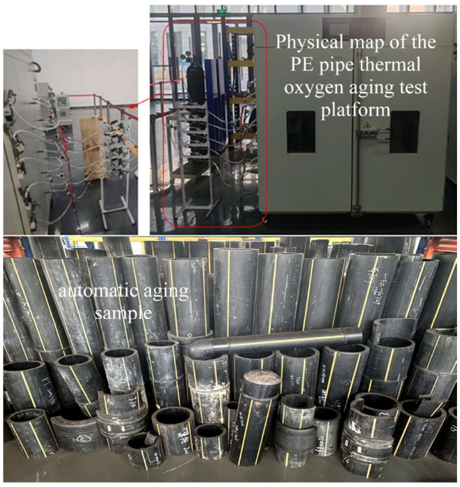

Fig. 6 PE pipe thermal-oxidative aging test.

In the experimental design, a sample of 50 pipes (p=50) was selected, and three operators (o=3) were tasked with defect detection on the welded joints. Each pipe underwent two measurements (r=2) to ensure the acquisition of representative data. Uniform environmental conditions were meticulously maintained throughout the measurement process to mitigate the impact of external factors on the data.

Table 3 Types of defects in gas polyethylene pipe welded joints and their potential hazards.

Defect name	Hazardous effect	Risk level
Under Plug Insertion(UPI)	Reduced tightness and structural integrity	1
Inclusion in Fusion Face(IFC)	Reduced joint strength and pressure resistance	2
Porosity (POR)	Reduced material strength and corrosion resistance	3
Cold Weld (CW)	Insufficient joint strength, prone to breaking	4
Structural Distortion(SD)	Geometric deformation, stress concentration	5
Over Weld (OW)	Degraded material properties, increased risk of cracking	6
Short Weld Length (SWL)	Reduced load-bearing and tensile strength	7

Table note: Hazard level 1 indicates the most dangerous, while hazard level 7 indicates the least dangerous.

4.2 Analysis based on a two-tier progressive attribute measurement system

A two-tiered data structure derived from phased array ultrasound scanning data was established. This structure integrates defects' presence (binary variable) and type (multiclass variable), ensuring standardization. At the initial analysis stage, the Kappa coefficient is utilized to assess the agreement of binary judgments among and within operators. Should the Kappa coefficient fall below 0.6, the analysis method is subjected to optimization. In the subsequent analysis phase, a weighted Kappa coefficient is computed for the initial screening sample, followed by a hierarchical diagnosis to assess the agreement of defect detection across seven distinct types of defects. A measurement system improvement plan is formulated based on the identified failure modes.

Module 1 Data acquisition and pre-processing

Initially, the integrity of the dataset was verified to ensure that it encompassed all defect types. Following the confirmation of the dataset's integrity, a subset of the data was documented using a two-tiered recording system proposed by the multi-category, small-sample defect data from gas polyethylene pipe welded joints, as detailed in Table 4.

Module 2 Binary defect assessment agreement analysis

This stage focuses on the agreement of binary defect detection for gas polyethylene pipe welded joints, including intra-operator and inter-operator agreements. The intra-operator agreement was assessed using Cohen's Kappa statistic (Table 5), while the inter-operator agreement was analyzed using Fleiss' Kappa statistic (Table 6). For data with a Kappa coefficient below 0.6, a root cause analysis was conducted to identify the cause of the error, and targeted calibration training was provided to the operator to enhance inspection standardization. After excluding outliers, the proportion of defect samples was recalculated to meet the data balance requirements for multi-category analysis.

Cohen's Kappa agreement analysis results in Table 5 show that Operators A, B, and C all met the preset agreement threshold (Kappa>0.6) for binary defect judgments. Specifically, Operator A had the highest Kappa value, 0.757 (standard error SE=0.141, Z value=5.370, p value<0.001), significantly higher than that of Operator B (Kappa=0.677, SE=0.140, Z=4.854, p<0.001) and Operator C (Kappa=0.642, SE=0.139, Z=4.634, p<0.001). The Fleiss Kappa agreement analysis results in Table 6 show that there is a high degree of agreement among operators in the determination of binary defects (Kappa=0.696, SE=0.037), significantly exceeding the model threshold requirement of 0.6, indicating a systematic consensus among different inspectors on the existence or absence of defects.

Table 4 Dataset for defect detection of gas polyethylene pipe welded joints (partial data).

Sample number	Standard	Measured by A				Measured by B				Measured by C			
		1st round		2st round		1st round		2st round		1st round		2st round	
1	UPI	1	UPI	1	OV	1	UPI	0	Pass	1	UPI	1	SWL
2	IFC	1	IFC	1	IFC	1	IFC	1	IFC	1	IFC	1	CW
3	POR	1	POR	1	POR	1	POR	1	POR	1	POR	0	Pass
4	CW	1	CW	1	CW	1	CW	1	CW	1	CW	1	CW
5	SD	1	SD	1	CW	1	SWL	1	SD	1	SD	1	CW
6	OV	1	OV	0	Pass	1	OV	1	OV	1	OV	1	SWL
7	SWL	1	SWL	1	SWL	1	SWL	1	SWL	1	SWL	1	SWL
8	Pass	1	POR	1	OV	0	Pass	0	Pass	0	Pass	1	SD
9	Pass	0	Pass	1	IFC	1	POR	0	Pass	0	Pass	0	Pass
10	Pass	0	Pass	0	Pass	1	SWL	0	Pass	0	Pass	1	SD

Table note: Pass indicates that the welded joint has no detectable defects or flaws.

Table 5 Cohen's Kappa statistic within the appraiser.

Appraiser	Response	Kappa	SE Kappa	Z	P
A	0	0.75688	0.140956	5.3696	0.0000
	1	0.75688	0.140956	5.3696	0.0000
B	0	0.67741	0.139570	4.8536	0.0000
	1	0.67741	0.139570	4.8536	0.0000
C	0	0.64228	0.138601	4.6341	0.0000
	1	0.64228	0.138601	4.6341	0.0000

Table 6 Fleiss Kappa statistics between appraisers.

Response	Kappa	SE Kappa	Z	P
0	0.695623	0.0365148	19.0504	0.0000
1	0.695623	0.0365148	19.0504	0.0000

Module 3 Weighted agreement analysis for multiple defect types

Following the initial agreement analysis, the weighted Kappa and squared weight weighting methods assign weights to the original defect type data, thereby creating a weighted dataset. Subsequently, the weighted dataset undergoes analysis across four dimensions of agreement to comprehensively evaluate the performance of the gas polyethylene pipe welded joint defect detection system at a more profound and nuanced level.

Table 7 Weighted Kappa statistic within the appraisers.

Appraiser	Response	Kappa	SE Kappa	Z	P
A	CW	0.72826	0.141421	5.1495	0.0000
	IFC	0.36842	0.141421	2.6051	0.0046
	OV	0.40476	0.141421	2.8621	0.0021
	Pass	0.75649	0.141421	5.3492	0.0000
	POR	0.45652	0.141421	3.2281	0.0006
	SD	0.45652	0.141421	3.2281	0.0006
	SWL	0.72826	0.141421	5.1495	0.0000
	UPI	-0.0309	0.141421	-0.2186	0.5866
	Overall	0.57435	0.066996	8.5730	0.0000
B	CW	0.55555	0.141421	3.9283	0.0000
	IFC	0.47916	0.141421	3.3882	0.0004
	OV	0.73474	0.141421	5.1954	0.0000
	Pass	0.67532	0.141421	4.7752	0.0000
	POR	0.72826	0.141421	5.1495	0.0000
	SD	0.55555	0.141421	3.9283	0.0000
	SWL	0.55555	0.141421	3.9283	0.0000
	UPI	-0.0101	0.141421	-0.0714	0.5285
	Overall	0.62736	0.067464	9.2992	0.0000
C	CW	0.63370	0.141421	4.4809	0.0000
	IFC	-0.0416	0.141421	-0.2946	0.6159
	OV	0.38107	0.141421	2.6946	0.0035
	Pass	0.63869	0.141421	4.5162	0.0000
	POR	0.36842	0.141421	2.6051	0.0046
	SD	0.28498	0.141421	2.0151	0.0219
	SWL	0.55555	0.141421	3.9283	0.0000
	UPI	-0.0101	0.141421	-0.0714	0.5285
	Overall	0.47787	0.069453	6.8806	0.0000

Table 8 Weighted Kappa statistic between appraisers.

Response	Kappa	SE Kappa	Z	P(
CW	0.666260	0.0365148	18.2463	0.0000
IFC	0.372822	0.0365148	10.2102	0.0000
OV	0.545958	0.0365148	14.9517	0.0000
Pass	0.695623	0.0365148	19.0504	0.0000
POR	0.621096	0.0365148	17.0094	0.0000
SD	0.503754	0.0365148	13.7959	0.0000
SWL	0.606092	0.0365148	16.5985	0.0000
UPI	0.227119	0.0365148	6.2199	0.0000
Overall	0.600728	0.0174747	34.3770	0.0000

The agreement analysis presented in Table 7 and Table 8 revealed significant variability in the discriminant agreement across specific weld defect types. Measurer A demonstrated a high level of agreement in discriminating cold weld (CW) defects, whereas Measurer B exhibited high agreement in discriminating over weld (OV) defects. Conversely, for the Under Plug Insertion (UPI) defect type, all three measurers recorded negative Kappa values, signifying an extremely low level of agreement in identifying this defect type and highlighting the urgent need for improvement measures.

The agreement analysis, utilizing the benchmark values from Table 9 and Table 10, revealed that the Kappa value notably exceeded the benchmark in comparisons across different measurers and defect types. Nonetheless, the analysis identified deficiencies in the repeatability among measurers across multiple measurements and the agreement between different measures.

Table 9 Weighted Kappa statistic between the appraiser and the reference value.

Appraiser	Response	Kappa	SE Kappa	Z	P
A	CW	0.86413	0.100000	8.6413	0.0000
	IFC	0.63432	0.100000	6.3432	0.0000
	OV	0.66777	0.100000	6.6777	0.0000
	Pass	0.79947	0.100000	7.9947	0.0000
	POR	0.70853	0.100000	7.0854	0.0000
	SD	0.75580	0.100000	7.5580	0.0000
	SWL	0.84639	0.100000	8.4639	0.0000
	UPI	0.31797	0.100000	3.1797	0.0007
	Overall	0.75174	0.048552	15.483	0.0000
B	CW	0.72826	0.100000	7.2826	0.0000
	IFC	0.73958	0.100000	7.3958	0.0000
	OV	0.86109	0.100000	8.6109	0.0000
	Pass	0.75896	0.100000	7.5897	0.0000
	POR	0.86214	0.100000	8.6215	0.0000
	SD	0.77777	0.100000	7.7778	0.0000
	SWL	0.74004	0.100000	7.4004	0.0000
	UPI	0.49494	0.100000	4.9495	0.0000
	Overall	0.78032	0.048696	16.024	0.0000
C	CW	0.81685	0.100000	8.1685	0.0000
	IFC	0.47916	0.100000	4.7917	0.0000
	OV	0.69975	0.100000	6.9976	0.0000
	Pass	0.77936	0.100000	7.7937	0.0000
	POR	0.68421	0.100000	6.8421	0.0000
	SD	0.64308	0.100000	6.4309	0.0000
	SWL	0.77777	0.100000	7.7778	0.0000
	UPI	0.49494	0.100000	4.9495	0.0000
	Overall	0.71963	0.049690	14.482	0.0000

Table 10 Weighted Kappa statistic between the appraiser and the reference value.

Response	Kappa	SE Kappa	Z	P
CW	0.803080	0.0577350	13.9098	0.0000
IFC	0.617690	0.0577350	10.6987	0.0000
OV	0.742875	0.0577350	12.8670	0.0000
Pass	0.779269	0.0577350	13.4973	0.0000
POR	0.751630	0.0577350	13.0186	0.0000
SD	0.725555	0.0577350	12.5670	0.0000
SWL	0.788071	0.0577350	13.6498	0.0000
UPI	0.435958	0.0577350	7.5510	0.0000
Overall	0.750570	0.0282798	26.5409	0.0000

5 Conclusion

A two-tiered progressive measurement system analysis model has been developed, with an in-depth study of the agreement evaluation index and the selection of sample sizes. The findings have been closely aligned with the measurement system analysis for defects in gas polyethylene pipe welded joints, aiming to offer innovative approaches for future quality inspections of defects in gas polyethylene pipe welds.

(1) A systematic investigation into the trends of key statistical measures of agreement—P_a, Cohen's kappa, and Fleiss's kappa—and their correlation with data distribution characteristics, including the number of distinguishable categories and the frequency of agreement/disagreement, has been conducted to identify suitable statistical measures for analyzing defects in gas polyethylene pipe welded joints at various levels of analysis.

(2) This study focuses on applying Cohen Kappa and Fleiss Kappa across varying agreement ratios and analyzes how the width of the confidence interval changes as the sample size increases. This analysis establishes a threshold for sample size selection, offering guidance for determining appropriate sample sizes in the measurement system analysis of welded defects in gas polyethylene pipelines, thereby enhancing the accuracy and reliability of statistical analyses.

(3) A two-tiered progressive attribute measurement system analysis model has been developed for defects in gas polyethylene pipe welded joints. At the initial tier, the measurement system is assessed to determine the presence of defects in the welded joint. A remedial strategy is proposed for non-conforming parts, and unqualified data is excluded. A weighted Kappa analysis is conducted at the subsequent tier to evaluate specific types of welded joint defects, calculating the agreement across four categories. It is concluded that while the measurement personnel demonstrated high accuracy in defect identification, there remains scope for enhancing the repeatability and agreement of measurements.

However, this approach is not without its limitations. Specifically, the method's capacity to quantify the impact of local characteristics on systematic errors is less robust than that of a metrological measurement system analysis, potentially compromising the accuracy of an exhaustive assessment of the measurement system's performance. Future research endeavors could focus on developing multivariate error decomposition techniques and exploring statistical methodologies capable of addressing multiple error sources concurrently, such as structural equation modeling (SEM) or variance component analysis (VCA), to ascertain the precise contribution of each local characteristic to the overall systematic error.

Fundings: This research was supported by the National Key Research and Development Program of China (No. 2023YFF0611600)

References

1. H. Lai, D. Fan, K. Liu, The Effect of Welding Defects on the Long-Term Performance of HDPE Pipes. Polymers **14**, 3936 (2022).
2. X. L. Meng, Study on Safety Status Assessment System of Polyethylene Gas Pipes in Urban Areas. Master Thesis, China University of Petroleum (Beijing), China, 2017.
3. H. Nie et al. Preliminary study on terahertz non-destructive testing for defect detection in hot melt joints of polyethylene pipes. Infrared Phys. Technol. **139**, 105300 (2024).
4. H. S. Lai, N. N. Tun, K. B. Yoon, S. H. Kil, Effects of defects on failure of butt fusion welded

polyethylene pipe. Int. J. Pressure Vessels Pip. **139**, 117–122 (2016).

5. R. K. Burdick, C. M. Borror, D. C. Montgomery, A Review of Methods for Measurement Systems Capability Analysis. J. Qual. Technol. **35** (4), 342–354 (2003).
6. Z. He, Y. Zhao, F. Zou, Research and Application of Methods for Attribute Measurement System Analysis. Industrial Engineering Journal. 82–85+121 (2008).
7. M. S. Alavijeh, R. Scott, F. Seviaryn, R. G. Maev, Using machine learning to automate ultrasound-based classification of butt-fused joints in medium-density polyethylene gas pipes. J. Acoust. Soc. Am. **150**, 561–572 (2021).
8. R. Kafieh, T. Lotfi, R. Amirfattahi, Automatic detection of defects on polyethylene pipe welding using thermal infrared imaging. Infrared Phys. Technol. **54**, 317–325 (2011).
9. S. Park et al., Early detection of steel tube welded joint failure using SPC-I nonlinear ultrasonic technique. Struct. Health Monit. **24**, 148–163 (2025).
10. F. Zuo, J. Liu, X. Zhao, L. Chen, L. Wang, An X-Ray-Based Automatic Welding Defect Detection Method for Special Equipment System. IEEE-ASME Trans. Mechatron. **29**, 2241–2252 (2024).
11. Y. Li, M. Hu, T. Wang, Visual inspection of weld surface quality. J. Intell. Fuzzy Syst. **39**, 5075–5084 (2020).
12. E. Mirmahdi, D. Afshari, M. A. Karimi Ivanaki, A Review of Ultrasonic Testing Applications in Spot Welding: Defect Evaluation in Experimental and Simulation Results. Trans. Indian Inst. Met. **76**, 1381–1392 (2023).
13. W. Qian, S. Dong, L. Chen, Q. Ren, Image enhancement method for low-light pipeline weld X-ray radiographs based on weakly supervised deep learning. NDT E Int. **143**, 103049 (2024).
14. Y. Zhang, Y.-M. Cheung, Graph-Based Dissimilarity Measurement for Cluster Analysis of Any-Type-Attributed Data. IEEE Trans. Neural Netw. Learn. Syst. **34**, 6530–6544 (2023).
15. R. Xiang, Y. Chen, J. W. Shen, S. Hu, Method for assessing the quality of data used in evaluating the performance of recognition algorithms for fruits and vegetables. Biosyst. Eng. **156**, 27–37 (2017).
16. K. Zhu et al., Robotic MAG welding defects and quality assessment with a defect threshold decision model-driven method. Mech. Syst. Signal Proc. **224**, 112056 (2025).
17. J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977).
18. S. Senn, Review of Fleiss, statistical methods for rates and proportions. Res. Synth. Methods **2**, 221–222 (2011).
19. S. Burt, L. Punnett, Evaluation of interrater reliability for posture observations in a field study. Appl. Ergon. **30**, 121–135 (1999).
20. Automotive Industry Action Group (AIAG), Measurement Systems Analysis Reference Manual , 4th edition, Troy, Mich: Chrysler, Ford, General Motors Supplier Quality Requirements Task Force, 2010.
21. J. R. Baldwin, A. Reuben, J. B. Newbury, A. Danese, Agreement Between Prospective and Retrospective Measures of Childhood Maltreatment: A Systematic Review and Meta-analysis. JAMA Psychiatry **76**, 584–593 (2019).
22. J. Yang, V. M. Chinchilli, Fixed-effects modeling of Cohen's weighted kappa for bivariate multinomial data. Comput. Stat. Data Anal. **55**, 1061–1070 (2011).
23. T. Yu, Z. Chen, H. Yin, N. Yi, M. Zhao, Statistical issues on evaluating agreement between THC and 11-OH-THC analysis in hair samples by Cohen's kappa. Forensic Sci. Int. **343**, 111496 (2023).
24. M. H. Lerchbaumer et al., Point-of-care lung ultrasound in COVID-19 patients: inter- and intra-observer agreement in a prospective observational study. Sci. Rep. **11**, 10678 (2021).
25. K. L. Gwet, Large-Sample Variance of Fleiss Generalized Kappa. Educ. Psychol. Meas. **81**, 781–790 (2021).
26. H. Brenner, U. Kliebsch, Dependence of weighted kappa coefficients on the number of categories. Epidemiology **7**, 199–202 (1996).
27. N. Moradzadeh, M. Ganjali, T. Baghfalaki, Weighted kappa as a function of unweighted kappas. Commun. Stat.-Simul. Comput. **46**, 3769–3780 (2017).
28. Q. Wang, H. Zhou, J. Xie, X. Xu, Nonlinear ultrasonic evaluation of high-density polyethylene natural gas pipe thermal butt fusion joint aging behavior. Int. J. Pressure Vessels Pip. **189**, 104272 (2021).
29. Q. Bao, B. Wang, M. Li, C. Li, J. Gao, The failure mechanism of field polyethylene gas pipeline with gas leakage at electrofusion joint. Anti-Corros. Methods Mater. **71**, 676–685 (2024).
30. S. Deveci, N. Antony, S. Nugroho, B. Eryigit, Effect of carbon black distribution on the properties of polyethylene pipes part 2: Degradation of butt fusion joint integrity. Polym. Degrad. Stabil. **162**, 138–147 (2019).