Ingeniería Investigación y Tecnología
Volume XXVII (Issue 1), January-March 2026 1-11
ISSN 2594-0732 FI-UNAM Peer-reviewed article
Article information: Received: July 9, 2025, accepted: October 31, 2025.
Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license
https://doi.org/10.22201/fi.25940732e.2026.27.1.006

Análisis de un sistema Gage R&R por atributos en la inspección de suelas de EVA mediante el coeficiente Kappa de Fleiss

Analysis of a Gage R-R system by attributes in the inspection of EVA soles using the Fleiss Kappa coefficient

Mares-Castro, Armando (Corresponding author)
Tecnológico Nacional de México, Guanajuato
E-mail: armando.mc@purisima.tecnm.mx
https://orcid.org/0000-0003-3884-958X

Aceves-Romero, Omar Alejandro
Tecnológico Nacional de México, Guanajuato
E-mail: RS20110469@purisima.tecnm.mx
https://orcid.org/0009-0000-9373-756X

Resumen

El presente estudio aborda la problemática inherente a la subjetividad en los sistemas de inspección visual por atributos, un factor que tradicionalmente eleva el riesgo de decisiones erróneas, como la aceptación de productos no conformes o el descarte injustificado de elementos válidos. La finalidad primordial de esta investigación es establecer una metodología cuantitativa y sistemática que permita valorar la capacidad de los inspectores para emitir juicios acertados, asegurando así una toma de decisiones consistente con la calidad real que la organización ofrece a sus clientes. Para lograr este objetivo, se propone un estudio de repetibilidad y reproducibilidad (R&R) por atributos, fundamentado en el método de análisis de riesgos y el índice Kappa de Fleiss. Este enfoque busca mitigar la variabilidad asociada al proceso de inspección mediante la estandarización de métodos, la unificación de criterios y la capacitación del personal. Mediante un estudio aplicado en un defecto cualitativo que se presenta en suelas de EVA, que consideró diversos escenarios y niveles de capacitación de los inspectores, se demostró que el procedimiento propuesto es capaz de alcanzar niveles de calidad del proceso de medición entre 95 % y 99 %. Los resultados evidencian que la reducción de la subjetividad y la mejora en la fiabilidad de las decisiones de inspección con este sistema se traducen en una disminución significativa del desperdicio y una mejora tangible de la eficiencia operativa. La metodología presentada es replicable en una amplia gama de procesos de inspección cualitativa, tanto en manufactura como en servicios.
Descriptores: Análisis de concordancia, Estudios R&R, Kappa de Fleiss, Evaluación por atributos, suela de EVA.

Abstract

This study addresses the inherent problem of subjectivity in visual inspection systems by attributes, a factor that traditionally increases the risk of erroneous decisions, such as the acceptance of non-conforming products or the unjustified discarding of valid elements. The primary purpose of this research is to establish a quantitative and systematic methodology to assess the ability of inspectors to make accurate judgments, thus ensuring decision making consistent with the actual quality that the organization offers to its customers. To achieve this objective, a repeatability and reproducibility (R&R) study by attributes is proposed, based on the risk analysis method and the Fleiss Kappa index. This approach seeks to mitigate the variability associated with the inspection process by standardizing methods, unifying criteria and training personnel. Through a study applied in a qualitative defect presented in EVA soles, which considered different scenarios and levels of training of inspectors, it was demonstrated that the proposed procedure to achieve the quality levels of the measurement process between 95 % and 99 %. The results show that the reduction of subjectivity and the improvement in the reliability of inspection decisions with this system translate into a significant decrease in waste and a tangible improvement in operational efficiency. The methodology presented is replicable in a wide range of qualitative inspection processes, both in manufacturing and services.
Keywords: Concordance analysis, R&R studies, Fleiss Kappa, Attribute evaluation, EVA soles.

Introducción

En el ámbito de la calidad y la producción, la fiabilidad de los datos de medición es fundamental para el monitoreo de la variabilidad del proceso y la implementación de iniciativas de mejora, especialmente en proyectos Six Sigma (Rodríguez *et al.,* 2020). Si bien, existen directrices y procedimientos estandarizados para evaluar la precisión de los sistemas de medición de variables continuas, el estudio Gage R&R para datos de atributos no ha sido tan exhaustivamente investigado y carece de pautas claras (Lyu & Chen, 2008). Esta situación es particularmente relevante en entornos de fabricación modernos, donde la supervisión de procesos a menudo involucra dos o más características de calidad relacionadas que son evaluadas cualitativamente (Lyu & Chen, 2008). La dependencia de la inspección visual para identificar inconformidades, comúnmente utilizada en la industria (Jiménez *et al.,* 2022), introduce una marcada subjetividad, lo que aumenta el riesgo de tomar decisiones incorrectas, ya sea aceptando productos que no cumplen con los estándares (error Tipo II) o rechazando aquellos que sí lo hacen (error Tipo I) (Morfín *et al.,* 2023).

Para abordar este desafío inherente a las evaluaciones subjetivas, los estudios Gage R&R para atributos surgen como una herramienta esencial (Murphy *et a*l., 2009; Vágõ & Kemény, 2012). Estos estudios permiten evaluar la consistencia de los criterios empleados por los inspectores al clasificar piezas o productos en categorías limitadas, como "aceptable" o "inaceptable" (Pulido & de la Vara Salazar, 2013). A diferencia de los sistemas que generan mediciones en escalas continuas, los sistemas de atributos, especialmente aquellos basados en apreciaciones sensoriales humanas como la vista, el gusto o el olfato, producen datos binarios o categóricos (Pulido & de la Vara Salazar, 2013).

Recientemente, estudios como el de Soares *et al*. (2022), con su revisión sistemática en Quality Engineering, muestran que, si bien hay numerosos trabajos de Gage R&R, la mayoría se concentra en mediciones continuas o cuantitativas, dejando relativamente pocos casos que usan atributos visuales y el coeficiente Kappa para evaluar juicios de inspección. Por ejemplo, Jiménez *et al*. (2022) presentan un análisis de R&R por atributos que cuantifica el desempeño de inspectores en inspección visual, utilizando análisis de riesgos y Kappa, pero se limita a unos pocos escenarios y no examina la variación bajo diferentes tamaños de muestra o condiciones operativas distintas. Además, trabajos sobre mediciones en superficies rugosas como los de Engler *et al*. (2024), resaltan qué características físicas del objeto inspeccionado afectan fuertemente la repetibilidad, lo que sugiere que los sistemas visuales requieran condiciones controladas para lograr niveles altos de fiabilidad.

Dentro de los estudios Gage R&R para atributos, se distinguen varios métodos de análisis, como el análisis de riesgos, la teoría de señales y el método analítico (Pulido & de la Vara Salazar, 2013). El método de análisis de riesgos, por ejemplo, es particularmente útil cuando no es factible obtener suficientes piezas con valores de referencia conocidos (Jiménez *et al*., 2022; Pulido & de la Vara Salazar, 2013). Este enfoque se basa en el análisis de hipótesis y tablas de contingencia para evaluar la consistencia de los criterios, aunque no cuantifica directamente la variabilidad del sistema de medición (Pulido & de la Vara Salazar, 2013). Por otro lado, el método de teoría de señales requiere una referencia en escala continua para cada pieza, buscando estimar el ancho promedio de las zonas de incertidumbre ((AIAG), 2010; Pulido & de la Vara Salazar, 2013).

Un componente crucial en la evaluación de la capacidad de los inspectores para tomar decisiones acertadas es el análisis de concordancia, que a menudo se apoya en tablas de contingencia y el índice Kappa de Cohen (Morfín *et al*., 2023). Este índice, propuesto por Cohen en 1960 (De Mast & Van Wieringen, 2007; Jiménez *et al*., 2022), mide el acuerdo entre dos observadores que califican el mismo objeto, ajustándose para la concordancia que podría ocurrir por azar ((AIAG), 2010; De Mast & Van Wieringen, 2007). Su interpretación clara y facilidad de cálculo han favorecido su adopción, especialmente en literatura médica y, más recientemente, en aplicaciones industriales, tras publicaciones clave como la del manual de la Automotive Industry Action Group ((AIAG), 2010). Para el caso de más de dos evaluadores, Fleiss generalizó este método, dando origen al coeficiente Kappa de Fleiss (Conger, 2017; Falotico & Quatto, 2015; Morfín *et al*., 2023). Este coeficiente es fundamental para determinar el grado de acuerdo entre evaluadores con un estándar, y se considera una medida más precisa que el simple porcentaje de acuerdo, ya que elimina el efecto del azar (Jiménez *et al*., 2022). Las directrices de AIAG sugieren que valores de Kappa superiores a 0.80 indican una concordancia de muy buena a excelente, mientras que valores inferiores a 0.40 señalan poca concordancia y oportunidades de mejora (Morfín *et al*., 2023).

Los estudios de Gage R&R para atributos también se enfocan en dos aspectos principales de la variabilidad: La repetibilidad y la reproducibilidad (Frenz *et al.,* 2012; Marques *et al.,* 2018; Pulido & de la Vara Salazar, 2013). La repetibilidad se refiere a la variabilidad de las mediciones obtenidas por el mismo operador bajo condiciones similares, es decir, la capacidad del instrumen-

to para obtener medidas con poca variabilidad (Frenz *et al.,* 2012; Jiménez *et al.,* 2022; Marques *et al.,* 2018; Pulido & de la Vara Salazar, 2013). Por otro lado, la reproducibilidad se asocia con la variabilidad de las mediciones del mismo objeto realizadas por diferentes operadores, reflejando la consistencia en el desempeño del observador (Frenz *et al.,* 2012; Marques *et al.,* 2018; Pulido & de la Vara Salazar, 2013). Si bien, tradicionalmente el análisis Gage R&R se realizaba con el método de rangos o promedios y rangos, el método ANOVA es preferido por su capacidad para identificar la interacción entre el operador y las piezas, proporcionando una estimación más exacta de las varianzas ((AIAG), 2010). La importancia de estos componentes radica en que, si la repetibilidad es alta en comparación con la reproducibilidad, esto puede indicar problemas con el instrumento o el procedimiento, mientras que lo opuesto puede señalar diferencias sistemáticas entre los operadores ((AIAG), 2010; Pulido & de la Vara Salazar, 2013).

La implementación de estudios de Gage R&R por atributos, especialmente mediante el método de análisis de riesgos y el índice Kappa, ha demostrado ser una herramienta efectiva para evaluar y mejorar los sistemas de inspección visual (Jiménez *et al.,* 2022). Estos estudios pueden medir el impacto de variaciones en el nivel de capacitación, experiencia, unificación de criterios, características físicas, métodos y estándares disponibles, lo que permite a las organizaciones optimizar sus sistemas de medición y tomar decisiones más consistentes sobre la calidad de sus productos (Rodríguez *et al.,* 2020). La estandarización de métodos y la capacitación del personal son factores clave para reducir la variabilidad y mejorar el nivel de calidad en los procesos de inspección (Jiménez *et al.,* 2022; Morfín *et al.,* 2023). Este enfoque se ha aplicado exitosamente en diversos sectores, desde el automotriz hasta el alimentario, como la inspección de zarzamoras, donde se ha evidenciado la importancia de la capacitación y la unificación de criterios para lograr altos niveles de concordancia (Morfín *et al.,* 2023).

En conclusión, la evaluación de sistemas de medición por atributos es un campo crítico para garantizar la calidad en la producción. Aunque históricamente menos estudiado que sus contrapartes de variables continuas, el desarrollo de metodologías basadas en el análisis de concordancia y el uso de coeficientes como el Kappa de Fleiss, respaldados por organizaciones como AIAG, ofrece un camino robusto para cuantificar y mejorar la fiabilidad de las decisiones de inspección, especialmente en contextos de evaluación subjetiva ((AIAG), 2010; Lyu & Chen, 2008; Morfín *et al.,* 2023). Esto permite a las empresas reducir el desperdicio, minimizar los errores y asegurar que la calidad percibida por el cliente sea consistente con la calidad real del producto (Rodríguez *et al.,* 2020). Nuestro trabajo aporta en ese contexto al proponer un estudio práctico con suelas de EVA, donde se implementa un diseño de inspección visual bajo criterios estandarizados, con capacitación previa y pruebas piloto, y se analiza el coeficiente Kappa de Fleiss combinado con un análisis de riesgos. De esta forma buscamos llenar la brecha identificada: Poca evidencia aplicada reciente que combine atributos, condiciones operativas reales, distintos evaluadores, y evaluación cuantitativa robusta.

La finalidad primordial de esta investigación se centra en la concepción de una metodología sistemática para cuantificar con precisión el desempeño de los sistemas de inspección visual por atributos aplicada en un proceso de inspección de suelas de EVA. Esta iniciativa surge de la imperante necesidad de abordar la subjetividad inherente a la evaluación cualitativa, un factor que históricamente incrementa la probabilidad de incurrir en errores significativos, como la aceptación de productos no conformes o el rechazo indebido de aquellos que sí lo están. Al proveer este enfoque, nuestro propósito es facilitar a las organizaciones la toma de decisiones coherentes con la calidad genuina de sus productos, lo cual se traduce en una reducción sustancial del desperdicio y una mejora tangible en la fiabilidad operativa, satisfaciendo así las expectativas del cliente (Rodríguez *et al.,* 2020).

La principal contribución de este trabajo consiste en la integración del método de análisis de riesgos con el coeficiente Kappa de Fleiss para evaluar la fiabilidad de sistemas de inspección visual. A diferencia de los métodos basados en teoría de señales o en análisis ANOVA, este enfoque no requiere valores de referencia continuos, resultando más adecuado para entornos industriales donde las decisiones dependen de juicios binarios. Su aplicabilidad práctica permite a las organizaciones validar de manera objetiva la competencia de los inspectores y la consistencia del proceso de inspección.

Metodología

En una empresa manufacturera de suelas, se ha identificado una problemática significativa relacionada con la aparición de burbujas en el proceso de inyección de suelas de EVA, la cual se manifiesta en un porcentaje de desperdicio que, lamentablemente, excede el 6 % de la producción total. La ausencia de un criterio unificado y estandarizado para clasificar los productos como "aceptables" o "inaceptables" ha provocado que material en buenas condiciones sea erróneamente desechado (Rodríguez *et al.,* 2020). Ante este escenario, la validación

del sistema de medición se vuelve imperativa para asegurar que las piezas clasificadas como defectuosas sean verdaderamente así y que las aceptadas cumplan con los estándares de calidad. Para abordar esta necesidad apremiante, se llevó a cabo un estudio enfocado en la aplicación del Gage R&R para atributos, con el objetivo de evaluar la fiabilidad del sistema de inspección visual y la concordancia entre los evaluadores.

Como parte de la validación del sistema de medición, se implementaron acciones de estandarización encaminadas a disminuir la variabilidad en los juicios de los inspectores. En primer término, se elaboraron instructivos visuales que incluyeron fotografías y descripciones claras de los distintos tipos y grados de defectos observados en las suelas. Dichos materiales se desarrollaron en conjunto con el personal de inspección y con apoyo del área de calidad, con el objetivo de consensuar una terminología y un conjunto de criterios comunes.

Posteriormente, se diseñó una plantilla de inspección que delimitó tres zonas críticas en la suela (puntera, arco y talón), con el propósito de guiar la observación visual y reducir la ambigüedad durante la clasificación. Esta herramienta se complementó con la calibración de condiciones de iluminación y la estandarización de la distancia y el ángulo de observación.

En cuanto a la capacitación, se realizaron talleres teórico-prácticos. En la primera sesión se abordaron los fundamentos del análisis R&R por atributos y la interpretación del índice Kappa; en la segunda se efectuaron ejercicios prácticos con piezas controladas, donde los participantes compararon sus resultados frente a un estándar de referencia. A partir de estos ejercicios, se identificaron discrepancias y se acordaron criterios finales de aceptación o rechazo. Los resultados de estas actividades mostraron un incremento en la concordancia promedio entre inspectores, pasando del 22 % inicial a 32 % tras la capacitación previa al estudio formal, lo que evidenció la efectividad de las acciones implementadas.

Diseño experimental

El presente estudio se centró en el análisis del sistema de medición, dado que la inspección visual de las suelas de EVA se basa en atributos (Rodríguez *et al.*, 2020). A continuación, se mencionan las etapas de importancia para el estudio:

Los sistemas de medición por atributos, a diferencia de aquellos que generan valores continuos, clasifican las mediciones en un número finito de categorías, como "pasa" o "no pasa". Dada la naturaleza cualitativa de la inspección de las suelas, donde la detección de burbujas es el defecto de interés, en la suela existen tres oportunidades de defecto, como se muestra en la Figura 1.

El estudio se desarrolló en la etapa final del proceso de moldeo por inyección, previo al área de empaque, en una línea de producción de suelas de EVA estilo "BR" de color blanco. La inspección visual se llevó a cabo en un área independiente del taller, acondicionada con iluminación controlada, aislamiento parcial de ruido y acceso restringido durante las sesiones experimentales.

Se implementó la estación de inspección, equipada con los siguientes elementos:

- Superficie de trabajo nivelada de 700 × 600 mm con fondo neutro gris 18 %.
- Fuente de iluminación LED calibrada entre 500 y 600 lux, con temperatura de color de 4000 a 4500 K, medida con luxómetro digital.
- Regla de referencia para fijar la distancia de observación en 40 cm y soporte ajustable que mantuvo el ángulo de visión a 90° respecto al plano de la suela.
- Plantilla física con delimitación de tres zonas críticas (puntera, arco y talón) y ejemplos impresos de defectos tipo "burbuja" para referencia inmediata.

Antes de iniciar las inspecciones se verificó el cumplimiento de las condiciones de iluminación y se aplicó una prueba de calibración con piezas patrón (una conforme y una no conforme). Las muestras se identificaron mediante códigos aleatorios alfanuméricos y fueron distribuidas por un encargado distinto a los evaluadores, manteniendo un esquema de ensayo ciego.

Cada inspector realizó dos evaluaciones consecutivas separadas por un intervalo mínimo de 15 minutos para minimizar el efecto de recuerdo visual. El tiempo promedio de evaluación por pieza fue de 8 a 10 segundos. Para evitar sesgos visuales, las piezas se manipularon con guantes de nitrilo y se colocaron sobre el fondo neutro sin exposición directa a luz natural.

Durante cada sesión se monitorearon temperaturas (22–26 °C) y humedad relativa (35–60 %), registrándose los valores iniciales y finales. Los incidentes o desviaciones de procedimiento fueron anotados en un registro de control. Las observaciones fuera de especificación (p. ej., caída de lux por debajo de 480 o interrupción de energía) se descartaron y se repitieron las evaluaciones.

Este protocolo permitió asegurar condiciones homogéneas de inspección, garantizando que la variabilidad observada se debiera únicamente al desempeño de los evaluadores y no a factores ambientales o ergonómicos.

Se empleó un criterio binario: "Aceptar" para suelas sin burbujas y "Rechazar" para aquellas con este tipo de imperfección. Para la recolección de datos, se utilizó

una hoja de muestreo específica, similar a la que se ilustra en la Tabla 1, donde dos inspectores, denominados "Persona 1" y "Persona 2", realizaron dos ensayos cada uno para evaluar la concordancia en sus clasificaciones.

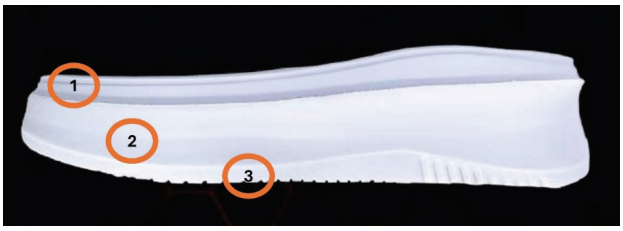

Figura 1. Oportunidades de defecto (zonas) para el defecto de la burbuja en la suela de EVA. Fuente: Elaboración propia.

Tabla 1. Ejemplo de formato de datos para un análisis de datos para atributos.

Evaluador	Muestra	Clasificación	Estándar
Cox	1	2	2
Montgomery	1	2	2
Taguchi	1	2	2
Duncan	1	1	2
Box	1	2	2

Fuente: Elaboración propia.

Tabla 2. Hoja de muestreo para evaluar la concordancia entre los evaluadores.

Ensayo	Atributo	Evaluador A-1	Evaluador B-1	Evaluador A-2	Evaluador B-2
1	_Rechazar_	_Aceptar_	_Rechazar_	_Rechazar_	_Rechazar_
2	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
3	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
4	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
5	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Aceptar_
6	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
7	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
8	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
9	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
10	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
11	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
12	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
13	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
14	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
15	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
16	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
17	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
18	_Aceptar_	_Aceptar_	_Aceptar_	_Aceptar_	_Aceptar_
19	_Aceptar_	_Aceptar_	_Aceptar_	_Aceptar_	_Aceptar_
20	_Rechazar_	_Rechazar_	_Aceptar_	_Rechazar_	_Aceptar_
21	_Aceptar_	_Aceptar_	_Aceptar_	_Aceptar_	_Aceptar_
22	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
23	_Aceptar_	_Aceptar_	_Aceptar_	_Aceptar_	_Aceptar_
24	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
25	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
26	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
27	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
28	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_	_Rechazar_
29	_Aceptar_	_Aceptar_	_Aceptar_	_Aceptar_	_Aceptar_
30	_Aceptar_	_Aceptar_	_Aceptar_	_Aceptar_	_Aceptar_

Fuente: Elaboración propia.

Para ello, se seleccionó una muestra de 30 piezas, de las cuales 24 se clasificaron como "malas" (rechazar) y 6 como "buenas" (aceptar), buscando representar la diversidad de calidades en el proceso. El defecto por evaluar es la burbuja, se utilizó una máquina tipo Strong, el producto es la suela de EVA estilo "BR" color blanco. Se utilizaron 2 evaluadores y se realizaron 2 ensayos por evaluador. Las suelas se asignaron a los inspectores de forma aleatoria y se registró la clasificación realizada en cada caso, cabe señalar que, en el momento de la evaluación, solamente el encargado del muestreo conoce cuál es cada una de las piezas proporcionadas a los evaluadores. Los datos obtenidos para el estudio se muestran en la Tabla 2.

Antes de este estudio, se realizaron dos evaluaciones previas del Gage R&R en las condiciones operativas existentes, seguidas de una fase de capacitación del personal y el diseño de una plantilla para la identificación de defectos, con el propósito de mejorar la congruencia en la inspección (Rodríguez *et al.,* 2020).

Análisis de concordancia

El análisis de la concordancia entre las evaluaciones es un componente crítico de este tipo de estudios. Una medida fundamental para cuantificar este acuerdo es el porcentaje de concordancia, el cual se calcula mediante la relación entre el número de clasificaciones coincidentes (m) y el número total de piezas inspeccionadas (N). Esta proporción, expresada en porcentaje, se determina utilizando la Ec. 1:

$$\textit{Porcentaje de concordancia} = 100 * (m)/N \tag{1}$$

Además del porcentaje simple de acuerdo, se recurrió al coeficiente Kappa, una métrica robusta que ajusta la concordancia observada eliminando el acuerdo que podría ocurrir simplemente por azar (Jiménez *et al.,* 2022). Inicialmente propuesto por Cohen para evaluar el acuerdo entre dos observadores, este método fue posteriormente generalizado por Fleiss para el caso de más de dos evaluadores, dando origen al coeficiente Kappa de Fleiss (Jiménez *et al.,* 2022). Su interpretación es clara: Valores superiores a 0.80 denotan una concordancia de muy buena a excelente, mientras que aquellos inferiores a 0.40 sugieren una concordancia deficiente y la necesidad de mejoras en el sistema de medición.

El estudio empleó una formulación para el cálculo de Kappa de Fleiss que se basa en el número de clasificaciones por muestra y categoría, así como en el número de ensayos y muestras. Para este cálculo, se considera k como el número total de categorías, m el número de ensayos, n el número de muestras, y x_{ij} el número de clasificaciones de la muestra i que corresponden a la categoría j. La fórmula empleada es la siguiente (Ec. 2):

$$Kappa = \frac{\left[\dfrac{\left(\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^2 - nm\right)}{nm(m-1)}\right] - \left[\sum_{i=1}^{k} p_j^2\right]}{1-\left[\sum_{i=1}^{k} p_j^2\right]} \tag{2}$$

Donde p_j representa la proporción general de clasificaciones en la categoría j.
Otra forma de simplificar la Ec. 2 se muestra en Ec. 3:

$$K = 1 - \frac{nm^2 - \sum_{j=1}^{n}\sum_{j=1}^{k} x_{ij}^2}{nm(m-1)\sum_{j=1}^{k} pj(1-P_j)} \tag{3}$$

La Varianza de K se calcula como Ec.4:

$$Var(K) = \frac{2}{nm(m-1)\left(\sum_{j=1}^{K} P_j(1-P_j)\right)} \tag{4}$$

Posteriormente, la significancia de este coeficiente se determina con el estadístico de prueba. Para determinar la significancia de Kappa > 0 se utiliza la Ec. 5 como estadístico de prueba:

$$Z = \frac{K}{\sqrt{Var(K)}} \tag{5}$$

Los resultados del análisis de Kappa, son cruciales para comprender la consistencia del sistema de inspección. La implementación de este procedimiento permitió no solo cuantificar la discrepancia en las evaluaciones iniciales, sino también medir el impacto de las mejoras introducidas, como la capacitación y las ayudas visuales estandarizadas.

Resultados y discusión

La presente investigación se propuso valorar cuantitativamente la aptitud de los inspectores en sistemas de inspección visual por atributos, buscando asegurar una toma de decisiones coherente con los estándares de calidad de la empresa. Para ello, se aplicó un estudio de Repetibilidad y Reproducibilidad (R&R) por atributos utilizando el método de análisis de riesgos y el índice Kappa de Fleiss (Jiménez *et al.,* 2022). Los resultados obtenidos, producto de un proceso de mejora continua que incluyó capacitación y el diseño de una plantilla para la identificación de defectos (Rodríguez *et al.,* 2020), demuestran una notable congruencia en el sistema de inspección visual.

Análisis del acuerdo individual de los evaluadores (Repetibilidad)

La Tabla 3, titulada "Acuerdo por Evaluador", presenta la consistencia de cada inspector consigo mismo en la clasificación de las piezas. Se observa que el Evaluador A alcanzó un 96.67 % de coincidencias con el estándar conocido en 30 inspecciones, mientras que el Evaluador B logró un 96.67 % de acuerdo con el mismo número de inspecciones. Estos porcentajes reflejan una alta repetibilidad individual, lo que indica que ambos evaluadores son consistentemente capaces de clasificar las piezas de manera similar en diferentes ensayos. Es importante destacar que una concordancia del evaluador con su propio criterio es fundamental para la fiabilidad del sistema de medición.

La Tabla 4 proporciona una medida crucial de la concordancia entre las evaluaciones repetidas realizadas por cada inspector. A diferencia de un simple porcentaje de acuerdo, el índice Kappa ajusta la concordancia por el acuerdo que podría ocurrir únicamente por azar, ofreciendo una valoración más robusta de la fiabilidad. Los valores de Kappa obtenidos para ambos evaluadores, tanto para decisiones de «ACEPTAR» como de «RECHAZAR», son consistentemente superiores a 0.80. Según las directrices generales para la interpretación del índice Kappa, un valor por encima de 0.80 se considera una concordancia de muy buena a excelente. Esto indica que la consistencia interna de cada evaluador (su repetibilidad) es sumamente alta, lo que significa que cada inspector tiende a clasificar la misma pieza de la misma manera en múltiples ocasiones.

El hecho de que los valores Kappa sean tan elevados y que el valor P (vs > 0) sea 0.0000 para todos los casos, refuerza la significancia estadística de estos hallazgos. Un valor p tan bajo significa que la probabilidad de que el acuerdo observado se deba simplemente al azar, es insignificante. Por lo tanto, el alto nivel de concordancia es genuino y refleja una capacidad sólida de los inspectores para aplicar sus criterios de manera consistente.

Evaluación de discrepancias (Errores Tipo I y II)

La Tabla 5, "Discrepancia en la evaluación", detalla los errores de clasificación incurridos por los inspectores. Para el Evaluador A, no se registraron errores de "RECHAZAR/ACEPTAR" (es decir, clasificar una pieza buena como mala, conocido como error Tipo I o riesgo del productor) ni de "ACEPTAR/RECHAZAR" (aceptar una pieza mala como buena, error Tipo II o riesgo del consumidor). Se observa una discrepancia combinada de 3.33 %. De manera similar, el Evaluador B mostró un 0.00 % de errores de Tipo I, un 4.17 % de errores de Tipo II y una discrepancia combinada del 3.33 %. Estos porcentajes extremadamente bajos de errores críticos son un indicador clave de la efectividad del sistema de inspección, ya que minimizar estos riesgos es esencial para la calidad del producto y la satisfacción del cliente.

Tabla 3. Acuerdo por Evaluador.

Evaluador	Núm. de inspeccionados	Núm. de coincidencias	Porcentaje	IC de 95 %
A	30	29	96.67	(82.78, 99.92)
B	30	29	96.67	(82.78, 99.92)

Nota: El número de coincidencias significa que el evaluador coincide consigo mismo a través de las mediciones.
Fuente: Elaboración propia.

Tabla 4. Estadísticos Kappa de Fleiss.

Evaluador	Respuesta	Kappa	Error estándar de Kappa	Z	P (vs > 0)
A	ACEPTAR	0.901800	0.182574	4.93936	0.0000
	RECHAZAR	0.901800	0.182574	4.93936	0.0000
B	ACEPTAR	0.911111	0.182574	4.99036	0.0000
	RECHAZAR	0.911111	0.182574	4.99036	0.0000

Fuente: Elaboración propia.

Tabla 5. Discrepancia en la evaluación.

Evaluador	RECHAZAR/ACEPTAR	Porcentaje	ACEPTAR/RECHAZAR	Porcentaje	Núm. de combinados	Porcentaje
A	0	0.00	0	0.00	1	3.33
B	0	0.00	1	4.17	1	3.33

Fuente: Elaboración propia.

Con el objetivo de profundizar en el estudio del comportamiento de los errores Tipo I y Tipo II, se realizó un análisis de sensibilidad en el que se simuló la clasificación de los evaluadores variando el tamaño de la muestra. Se consideraron escenarios con 20, 30 (tamaño original del estudio) y 40 unidades, manteniendo la proporción inicial de piezas conformes y no conformes. Los resultados, resumidos en la Tabla 6, muestran la estabilidad del sistema de medición y proporcionan una guía práctica para la selección del tamaño muestral en estudios futuros.

Como se puede observar en la Tabla 8, el tamaño de muestra de 30 unidades, utilizado en el estudio original, demostró ser un punto óptimo donde los errores Tipo I y II se mantienen en niveles mínimos y estables. La reducción a 20 unidades introduce un incremento en el Error Tipo II, lo que conlleva un mayor riesgo de aceptar productos no conformes. Por otro lado, el aumento a 40 unidades no representa una mejora significativa en la confiabilidad del sistema que justifique el incremento en el costo y tiempo de inspección. Este hallazgo refuerza la validez del diseño experimental inicial y ofrece un criterio cuantitativo para la planificación de estudios Gage R&R por atributos en contextos similares.

La Tabla 7 se enfoca en la repetibilidad (o consistencia intra-evaluador), es decir, la capacidad de un mismo evaluador para dar resultados consistentes cuando inspecciona la misma pieza en múltiples ocasiones. Un alto valor de Kappa en este contexto indica que el inspector es fiable en sus propias decisiones a lo largo del tiempo. Para el Evaluador A, un Kappa de 0.950900 tanto para "ACEPTAR" como para "RECHAZAR" indica una concordancia excelente consigo mismo. Esto significa que el Evaluador A es extremadamente consistente en sus propias clasificaciones a lo largo de las pruebas. Para el Evaluador B, un Kappa de 0.857732 para ambas respuestas también representa una concordancia de muy buena a excelente. Aunque ligeramente inferior al Evaluador A, sigue siendo un resultado muy robusto y positivo. Significancia Estadística (P-value = 0.0000): Para todos los valores de Kappa presentados, el valor p asociado (P (vs > 0)) es 0.0000. Esto es fundamental, ya que indica que la probabilidad de que la concordancia observada se deba únicamente al azar es prácticamente nula. Por lo tanto, los altos niveles de consistencia interna de ambos evaluadores son estadísticamente significativos y no aleatorios.

Análisis del acuerdo entre evaluadores (Reproducibilidad)

El estudio también se enfocó en la reproducibilidad, es decir, la consistencia entre los diferentes inspectores. La Tabla 8 revela que, en 30 inspecciones, todas las estimaciones de los evaluadores coincidieron entre sí en un 90.00 %. Complementariamente, la Tabla 9 proporciona una medida más robusta de esta concordancia. El índice Kappa obtenido para ambos casos (ACEPTAR y RECHAZAR) fue de 0.844720. Un valor de Kappa superior a 0.80 se considera como una concordancia de muy buena a excelente. Esto significa que, más allá de la coincidencia esperada por azar, existe un alto grado de acuerdo real entre los inspectores, lo que es crucial para la unificación de criterios en la inspección visual.

De manera complementaria, se efectuó una simulación para observar el comportamiento del coeficiente Kappa al modificar el número de piezas y de evaluadores. Los resultados mostraron que al reducir la muestra a 20 unidades, el valor de Kappa disminuyó en promedio 0.07 puntos, evidenciando una menor estabilidad

Tabla 7. Estadísticos Kappa de Fleiss Individual por evaluador.

Evaluador	Respuesta	Kappa	Error estándar de Kappa	Z	P (vs > 0)
A	ACEPTAR	0.950900	0.129099	7.36564	0.0000
	RECHAZAR	0.950900	0.129099	7.36564	0.0000
B	ACEPTAR	0.857732	0.129099	6.64397	0.0000
	RECHAZAR	0.857732	0.129099	6.64397	0.0000

Fuente: Elaboración propia.

Tabla 8. Acuerdo de evaluación, entre evaluadores.

Núm. de inspeccionados	Núm. de coincidencias	Porcentaje	IC de 95 %
30	27	90.00	(73.47, 97.89)

Nota: Núm. de coincidencias: Todas las estimaciones de los evaluadores coinciden entre sí. Fuente: Elaboración propia.

Tabla 9. Estadísticos Kappa de Fleiss entre evaluadores.

Respuesta	Kappa	Error estándar de Kappa	Z	P (vs > 0)
ACEPTAR	0.844720	0.0745356	11.3331	0.0000
RECHAZAR	0.844720	0.0745356	11.3331	0.0000

Fuente: Elaboración propia.

estadística. En cambio, al aumentar a tres evaluadores, la variación del coeficiente fue inferior a 3 %, lo que confirma la robustez del método frente a pequeñas modificaciones en el tamaño del grupo de evaluación.

Acuerdo general de los evaluadores con el estándar

Finalmente, se evaluó la capacidad del sistema en su conjunto para clasificar correctamente las piezas respecto a un estándar conocido. La Tabla 10 muestra que todas las estimaciones de los evaluadores coincidieron con el estándar conocido en un 90.00 % de las 30 piezas inspeccionadas. La medida de concordancia corregida por azar, el índice Kappa de Fleiss, se presenta en la Tabla 11. Para ambas respuestas (ACEPTAR y RECHAZAR), el valor Kappa fue de 0.904316. Este resultado supera el umbral de 0.80, indicando una excelente capacidad del sistema de inspección visual para tomar decisiones acertadas en línea con la calidad real que la empresa busca entregar a sus clientes. La Figura 2, aunque no presenta datos gráficos en el extracto, es referenciada como una representación visual del acuerdo de evaluación. Típicamente, estas figuras complementan la información numérica, permitiendo una rápida identificación de patrones y la confirmación de la consistencia del sistema de medición.

Los resultados presentados demuestran la efectividad del sistema de inspección visual por atributos en la empresa. La alta repetibilidad individual de los inspectores, combinada con una excelente reproducibilidad entre ellos y un elevado grado de concordancia con el estándar, valida la fiabilidad de las decisiones de inspección. Estos logros no son triviales, especialmente considerando la naturaleza subjetiva inherente a la inspección visual de atributos como la apariencia (Morfín *et al.*, 2023; Pulido & de la Vara Salazar, 2013). La reducción de la variabilidad en los resultados de inspección se ha conseguido mediante la estandarización de métodos, la unificación de criterios y la capacitación del personal (Pulido & de la Vara Salazar, 2013).

La metodología implementada, que incluye el uso del índice Kappa de Fleiss para múltiples evaluadores, ha permitido cuantificar el desempeño del sistema de manera precisa, eliminando el efecto de la concordancia por azar. Antes de estas mejoras, el sistema de medición presentaba una concordancia mucho menor (22 % inicialmente, luego 32 % tras capacitación básica), lo que subraya la importancia de las intervenciones realizadas, como la capacitación exhaustiva y el desarrollo de ayudas visuales como plantillas para la identificación de defectos ((AIAG), 2010).

Tabla 10. Acuerdo de evaluación, todos los evaluadores *vs* el estándar.

Núm. de inspeccionados	Núm. de coincidencias	Porcentaje	IC de 95 %
30	27	90.00	(73.47, 97.89)

Nota: Número de coincidencias significa que todas las estimaciones de los evaluadores coinciden con el estándar utilizado. Fuente: Elaboración propia.

Tabla 11. Estadísticos Kappa de Fleiss, todos los evaluadores *vs* el estándar.

Respuesta	Kappa	Error estándar de Kappa	Z	P (vs > 0)
ACEPTAR	0.904316	0.0912871	9.90629	0.0000
RECHAZAR	0.904316	0.0912871	9.90629	0.0000

Fuente: Elaboración propia.

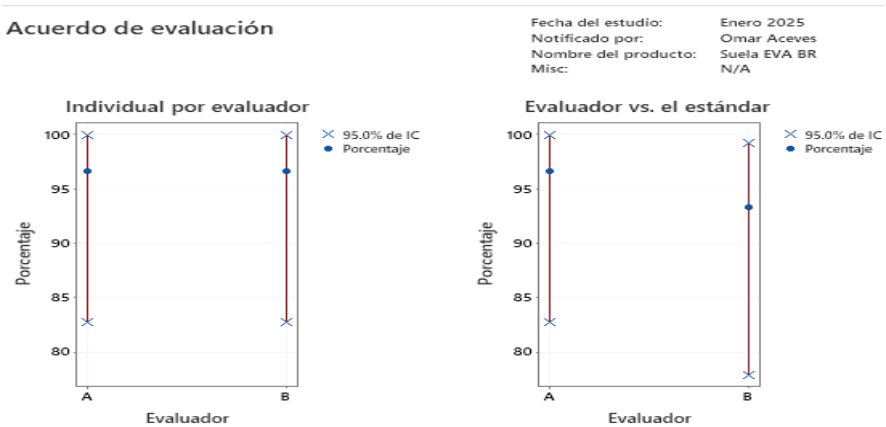

Figura 2. Acuerdo de evaluación.
Fuente: Elaboración propia.

Los resultados obtenidos en este estudio se alinean y fortalecen los hallazgos reportados en la literatura reciente sobre sistemas de medición por atributos. En particular, Jiménez *et al.* (2022) reportaron valores del coeficiente Kappa superiores a 0.85 tras la implementación de un proceso de estandarización en un sistema de inspección visual para la industria automotriz. Nuestro trabajo, al alcanzar valores de Kappa consistentemente superiores a 0.90 en la evaluación de suelas de EVA, no solo corrobora la efectividad de la metodología, sino que sugiere su robustez para ser aplicada en contextos donde la definición del defecto puede ser inherentemente más subjetiva. Un factor común decisivo en ambos estudios fue la implementación de una capacitación especializada y el uso de ayudas visuales para la unificación de criterios.

Al contrastar nuestros hallazgos con el trabajo de Morfin *et al.* (2023) en la inspección de zarzamoras, se pueden discernir los matices de la aplicabilidad del método. Mientras que en su estudio la variabilidad ambiental y la perecibilidad del producto introdujeron factores de ruido adicionales, nuestro entorno de manufactura controlada para suelas de EVA permitió una mayor estandarización de las condiciones de inspección, lo que probablemente contribuyó a los niveles excepcionalmente altos de repetibilidad intra-evaluador. Esta comparación subraya la versatilidad del estudio Gage R&R por atributos, al mismo tiempo que resalta la importancia de contextualizar sus resultados en función de las particularidades del proceso y el producto bajo análisis.

Si bien el presente estudio demuestra la efectividad del coeficiente Kappa de Fleiss y el método de análisis de riesgos, es importante reconocer sus limitaciones. En primer lugar, el tamaño de la muestra (n=30), si bien se demostró suficiente y eficiente para los objetivos de este caso de estudio, podría ser incrementado en investigaciones futuras que busquen generalizar los hallazgos a una gama más amplia de defectos o configuraciones de producción. En segundo lugar, el análisis de reproducibilidad se limitó a dos evaluadores; la inclusión de un mayor número de inspectores en estudios subsecuentes permitiría un análisis más robusto de la variabilidad inter-evaluador. Finalmente, esta investigación se centró en un único defecto cualitativo (burbujas); la aplicación y validación de esta metodología para productos con múltiples defectos concurrentes o con características cualitativas más complejas representan una fructífera línea de investigación futura.

Conclusiones

La investigación y los estudios analizados en las fuentes demuestran la imperativa necesidad y la efectividad de implementar metodologías robustas para evaluar los sistemas de medición por atributos, especialmente en contextos donde la valoración es intrínsecamente subjetiva, como la inspección visual (Morfín *et al.,* 2023; Pulido & de la Vara Salazar, 2013).

Se ha evidenciado que el Análisis de Repetibilidad y Reproducibilidad (R&R) por atributos, particularmente a través del uso del coeficiente Kappa de Fleiss, constituye una herramienta de valor incalculable para cuantificar el desempeño de los sistemas de inspección visual. Este estadístico es fundamental porque ajusta la concordancia observada, eliminando aquella que podría ocurrir simplemente por azar (Morfín *et al.,* 2023; Pulido & de la Vara Salazar, 2013), proporcionando así una medida más fiel de la fiabilidad del proceso de medición ((AIAG), 2010).

En última instancia, la aplicación de estos estudios Gage R&R por atributos no solo permite validar la capacidad de un sistema de medición para distinguir entre productos conformes y no conformes, sino que también se convierte en una herramienta estratégica para la mejora continua (Pulido & de la Vara Salazar, 2013). Al asegurar la confiabilidad de los datos de inspección, las empresas pueden reducir significativamente el desperdicio, optimizar la eficiencia operativa y garantizar que la calidad entregada a los clientes sea consistente y alineada con las expectativas del mercado ((AIAG), 2010; Rodríguez *et al.,* 2020).

Referencias

(AIAG). (2010). *Measurement systems analysis reference manual,* 4th ed. Chrysler, Ford.

Conger, A. J. (2017). Kappa and rater accuracy: Paradigms and parameters. *Educational and Psychological Measurement,* 77(6), 1019-1047. https://doi.org/10.1177/0013164416663277

De Mast, J., & Van-Wieringen, W. N. (2007). Measurement system analysis for categorical measurements: Agreement and kappa-type indices. *Journal of Quality Technology*, 39(3), 191-202. https://doi.org/10.1080/00224065.2007.11917688

Engler, C., Georgiadis, A., Lange, D., & Meier, N. (2024). New method for assessing the repeatability of the measuring system for roughness measurements. *International Journal of Metrology and Quality Engineering*, 15, 10. https://doi.org/10.1051/IJMQE/2024008

Falotico, R., & Quatto, P. (2015). Fleiss' kappa statistic without paradoxes. *Quality and Quantity,* 49(2), 463-470. https://doi.org/10.1007/S11135-014-0003-1/METRICS

Frenz, E. R., Smith, G. M., Rivera, R., & Morales, J. (2012). Validating NVH EOL measurements using GR&R. SAE Technical

Papers. On 7th International Styrian Noise, Vibration & Harshness Congress: The European Automotive Noise Conference. Retrieved on https://doi.org/10.4271/2012-01-1536

Jiménez-Jiménez, L. C., Patiño-Rodríguez, C. E., Guevara-Carazas, F. J. (2022). Análisis de R&R por atributos que valora cuantitativamente el desempeño de un sistema de inspección visual. *Ingeniare. Revista Chilena de Ingeniería*, 30(3), 466-481. https://doi.org/10.4067/S0718-33052022000300466

Lyu, J., & Chen, M. (2008). A bivariate attribute measurement model for six sigma project. On Proceedings of the 4th IEEE International Conference on Management of Innovation and Technology, ICMIT, 1129-1134. Retrieved on https://doi.org/10.1109/ICMIT.2008.4654528

Marques, C., Lopes, N., Santos, G., Delgado, I., & Delgado, P. (2018). Improving operator evaluation skills for defect classification using training strategy supported by attribute agreement analysis. *Measurement: Journal of the International Measurement Confederation*, 119, 129-141. https://doi.org/10.1016/j.measurement.2018.01.034

Morfín-Magaña, R., Pulido-Toro, E., Corona-Soto, M. de J., Mendoza-Ballines, L. B., & Topete-Betancourt, K. C. (2023). Análisis del sistema de medición por atributos en la inspección de calidad para zarzamoras. *Ingeniería Investigación y Tecnología*, 24(3), 1-7. https://doi.org/10.22201/FI.25940732E.2023.24.3.018

Murphy, S. A., Moeller, S. E., Page, J. R., Cerqua, J., & Boarman, M. (2009). Leveraging measurement system analysis (MSA) to improve library assessment: The attribute gage R&R. *College and Research Libraries*, 70(6), 568-574. https://doi.org/10.5860/0700568

Pulido, H. G., & de la Vara-Salazar, R. (2013). *Control estadístico de la calidad y Seis Sigma*. 3rd. ed. M.-H. Interamericana. Retrieved on https://books.google.com.mx/books?id=PCYKkAEACAAJ

Rodríguez-Medina, M. A., Terrones-Lucero, J. G., Poblano-Ojinaga, E. R., & Terrazas-Mata, L. E. (2020). Reducción de la variación en un proceso de moldeo de partes a través de la metodología 6 Sigma. *DYNA MANAGEMENT*, 8(1), 18. https://doi.org/10.6036/MN9831

Soares, W. de O. S., Peruchi, R. S., Silva, R. A. V., & Rotella-Junior, P. (2022). Gage R&R studies in measurement system analysis: A systematic literature review. *Quality Engineering*, 34(3), 382-403. https://doi.org/10.1080/08982112.2022.2069505

Vágõ, E., & Kemény, S. (2012). Random effects model for attribute gauge R&R. *Quality and Reliability Engineering International*, 28(8), 807-823. https://doi.org/10.1002/qre.1269

How to cite:

Mares-Castro, A., & Aceves-Romero, O. A. (2026). Análisis de un sistema Gage R&R por atributos en la inspección de suelas de EVA mediante el coeficiente Kappa de Fleiss. *Ingeniería Investigación y Tecnología,* 27(01), 1-11. https://doi.org/10.22201/fi.25940732e.2026.27.1.006