Acceptance sampling with multiple inspectors: Critical aspects in aggregating individual assessments

Acceptance Sampling with Multiple Inspectors: Critical Aspects in Aggregating Individual Assessments

Fiorenzo Franceschini*, Domenico A. Maisano, Luca Mastrogiacomo,

Politecnico di Torino

Corso Duca degli Abruzzi 24, 10129, Torino, Italy

*Corresponding author: fiorenzo.franceschini@polito.it

Abstract

In productions characterised by small scale, high customization and high added value – e.g., in sectors like nuclear energy, aerospace, and *haute couture* – it is imperative to strive for zero risk of delivering defective products to customers. In order to achieve this, the production conformity verification can be redundant and involve multiple inspectors. In addition, since inspections are often performed manually by human inspectors, whose assessments may be somewhat subjective, a precise definition of inspection procedures and thorough training of inspectors are essential to achieve consistent results. Unfortunately, combining the assessments by multiple inspectors with the traditional acceptance-sampling schemes used for conformity verification (e.g., *single*, *double*, *sequential* sampling plans) is not straightforward, as these schemes generally assume that inspections are unique, non-redundant, and carried out by a single inspector.

This paper focuses on the aggregation of multiple-inspector conformity assessments within traditional acceptance-sampling schemes, to reach a final *lot-disposition decision* (LDD, typically *pass*/*fail* of the whole lot). Various aggregation approaches can be applied, often markedly different from one another; the analysis of these approaches constitutes an innovative aspect of this study. With the help of a plurality of realistic examples in the *haute-couture* sector, the paper highlights how, starting from individual conformity assessments by each inspector, different global decision-making scenarios can be generated, often contradictory, potentially resulting in a state of total undecidability regarding the acceptance of a supply. This may happen especially in situations characterized by a certain degree of disagreement between inspectors. Finally, it is proposed the use of an inspection-agreement indicator (i.e., Gwet's κ_G) aimed at highlighting potentially controversial situations and triggering appropriate actions to make inspectors' assessments more robust and consistent.

Keywords: Acceptance sampling, Multiple inspectors, Conformity assessment, ISO 2859-1, *Haute couture*, Inter-inspector agreement, Gwet's κ_G, Inspector training.

List of acronyms and variables

AQL: Acceptance quality level;

ISO: International Organization for Standardization;

LDD: Lot-disposition decision;

PTN: Fictitious name of the company described in the case study;

SSP: Single sampling plan;

$\boldsymbol{c}$: Acceptance number of a single sampling;

$\boldsymbol{d}$: Number of defective units found within the sample units;

$\boldsymbol{k}$: Number of categories used for assessments;

$\boldsymbol{\kappa_G}$: Gwet's kappa indicator of inter-inspector agreement;

$\boldsymbol{n}$: Sample size of a single sampling plan;

$\boldsymbol{N}$: Lot size;

$\boldsymbol{s}$: Number of inspectors involved in the sample inspection;

$\boldsymbol{s_1}$, $\boldsymbol{s_2}$, …: Generic inspector involved in the sample inspection;

$\boldsymbol{u_1}$, $\boldsymbol{u_2}$, …: Generic product unit of the inspected sample.

1. Introduction

In manufacturing contexts characterized by small-lot productions with high added value, the conformity assessment of outgoing products often involves multiple inspectors. This is particularly common in inspections of critical components for specialized industrial applications (Odakura et al. 2015), software-application testing (Kniaght and Mayer, 1991; Aurum et al. 2002), *haute couture* (Ngan et al., 2011; Yuen et al, 2009; Keist 2015), and general situations where (i) inspections are manual and may involve *subjective* conformity assessments, and (ii) the risk of delivering defective products to customers must be (nearly) eliminated (Chandra and Schall, 1998; Chun 2016; Genta et al. 2018; Mandroli et al. 2006; Verna et al. 2021; Ferrari et al., 2024). Subjective conformity assessments can arise from various factors, including differences in attention, visual sensitivity, work experience, and mental or physical fatigue (Chun, 2009; Franceschini and Maisano, 2018; Genta et al., 2020). The work environment itself, influenced by factors such as lighting, temperature and noise levels, can also impact on manual conformity assessments (Duffuaa and Khan, 2005). Additionally, training and experience of inspectors in quality-control practices play a crucial role in reducing the inherent subjectivity of assessments.

Conformity assessments for a given supply, or production *lot*[1], are surely more complex to manage when involving multiple inspectors, compared to a single-inspector scenario. In the latter case, an individual inspector assesses the conformity of a certain portion of the entire supply (e.g., the so-called *sample*) and then makes a final decision (*pass* or *fail*) for the entire supply, which is based on the application of a conventional *lot-disposition decision* (LDD) rule. However, in the case of

[1] A *production lot* or more simply *lot* is defined as a "*group of units of the same product, produced under homogeneous conditions, e.g., with the same machines, operators, materials, and roughly at the same time*" (Montgomery, 2019).

assessments carried out by multiple inspectors, there is the added challenge of aggregating their individual assessments into a global LDD that is robust and unambiguous, mitigating issues of low robustness (i.e., cases where the overall outcome tends to vary according to the different aggregation criteria or parameters used) and *undecidability* (i.e., when a *pass/fail* decision cannot be uniquely determined).

The scientific literature on sampling plans is extensive and includes many different schemes for selecting the samples to be inspected (e.g., by *variables*, by *attributes*, *single*, *double*, *sequential*, *skip-lot*, etc.). However, it almost exclusively refers to cases in which each product unit is inspected only once, by a single inspector, and where the assessment is assumed to be free from errors or subjective interpretations in identifying potential (non)conformities (Duffuaa and El-Ga'aly, 2015). The guidelines for designing and determining the parameters of sampling plans (e.g., sample size, acceptance/rejection number(s), etc.) are well established from both a scientific and regulatory perspective (Schilling and Neubauer, 2009; Montgomery, 2019; ISO 2859-1:1999, 1999; BS 6001-0:2006, 2006).

On the other hand, the scientific literature lacks a thorough analysis of multiple-inspector conformity assessments. This gap serves as the primary motivation for this research, which aims to highlight the potential challenges characterizing multiple-inspector assessments, both from a conceptual and practical point of view, especially when the agreement between inspectors is weak. Through various realistic examples within the *haute-couture* sector, this study will first investigate the characteristics of typical aggregation approaches. It will then show that any aggregation technique can, under certain circumstances, produce contradictory or doubtful results and that making a global LDD (*pass*/*fail*) in uncertain cases would be risky and imprudent. Instead, it would likely be wiser to conduct further investigations and improve the level of homogeneity and agreement among inspectors.

In summary, the objectives of this study can be condensed into the following three research questions:

RQ#1: *What are the characteristic features of the aggregation mechanisms adopted in multiple-inspector acceptance sampling?*

RQ#2: *Under which contextual conditions do these aggregation mechanisms become critical?*

RQ#3: *What tools can be used to predict these critical conditions?*

The rest of this paper is structured into five sections. Section 2 briefly reviews the scientific literature on acceptance-sampling techniques in the presence of multiple inspectors. Section 3 introduces a case study in the *haute-couture* context, which will accompany the discussion. Section 4 illustrates the methodology of this research, (i) recalling possible approaches for adapting traditional acceptance-sampling schemes to conformity assessments by multiple inspectors, (ii) introducing several

application examples (or practical situations), which may give rise to doubtful/controversial results, and (iii) suggesting the introduction of an indicator of inter-inspector agreement (i.e., Gwet's κ_G), which can be a useful diagnostic tool to identify potentially critical situations. The concluding section summarizes the most significant contributions of this research, its practical implications, limitations, and suggestions for future developments. Finally, the appendix section provides additional information on various aspects: a categorization of garment defects, an example of inspector-training test in the *haute-couture* sector, and further details concerning the results presented in Sect. 4.

2. Literature review

The literature on acceptance sampling with multiple inspectors is relatively limited and reflects a significant research gap. Standards such as ISO 2859-1:1999 (1999) and BS 6001-0:2006 (2006), concern sampling plans but rarely address multiple-inspector scenarios. Although the literature includes some methodologies in which multiple inspections are performed on the same product units, they are generally carried out by a single inspector. For example, Duffuaa and Khan (2005) introduced a *sequential* multiple-inspection model targeting critical components to mitigate the risks associated with inaccurate defect detection and potential acceptance of life-threatening defective units. Despite increasing inspection costs, this model offers a strategy to minimise assessment errors (Chandra and Schall 1998). Mandroli et al. (2006) conducted a literature survey, categorizing inspections into four distinct types: (i) *simple* inspection of a single unit once; (ii) *fractional* inspection, examining a fixed portion of a lot, with the extremes being none or the entire lot; (iii) *repeated* inspection, assessing the same units(s) multiple times; and (iv) *dynamic* inspection, evaluating lot units sequentially with dynamic *pass* or *fail* decisions, as opposed to predetermined fractions. Chun (2009) explored the design of serial inspection processes, where a product undergoes repeated inspections, either by a single inspector or, sequentially, by multiple inspectors. This approach, grounded in Bayesian statistics, utilizes a negative-binomial prior model to estimate the residual number of defectives, the probability of detecting all defectives, and the likelihood of identifying new defectives in subsequent inspections.

In *parallel*[2] inspections, several inspectors examine the same product unit separately, identifying nonconformities without rectifying them. This is common in software engineering, where documents are reviewed independently by engineers, who note issues for subsequent resolution. The process

[2] The term *parallel* does not necessarily imply that inspections are simultaneous in time, but rather indicates that the same product units are evaluated by several inspectors, each operating independently. In other words, inspectors may be considered as parallel channels through which the same product units are inspected in rotation. At the end of the inspection process, each inspector will have independently assessed and made an individual conformity assessment for each product unit in the sample (Knight and Mayers, 1993).

culminates in a debriefing to consolidate findings on identified nonconformities. Knight and Mayers (1993) presented *phased* inspections for software-product assessments, promoting efficient utilization of resources by having inspectors review the product independently, before a reconciliation phase to consolidate and compare findings, thus enhancing the quality of documentation without group discussions of the product itself. Aurum et al. (2002) reviewed software inspection processes, highlighting innovative approaches, experimental studies, and outlining directions for future research in software inspections. More recently, Chun (2016) proposed an enhanced inspection model that accounts for the variability in detection probability across multiple inspection plans. Some studies, such as that by Mazza and Alvarez (2017), mention the practice of parallel inspections by multiple inspectors in the *haute-couture* sector, although they do not address the potential problems of aggregating the assessments by multiple inspectors and integrating them with traditional acceptance-sampling plans.

3. Case study

3.1 General description of the company

A small company in northern Italy, which for reasons of confidentiality is kept anonymous and conventionally referred to as "PTN", operates in the *haute-couture* sector. This company carries out tailoring operations on luxury garments, with a high level of customisation, using the highest quality materials and handcrafted tailoring techniques. Unlike the *prêt-à-porter* sector, which is characterised by the relatively quick production of standardised garments in large quantities, PTN's production consists of small and highly customised lots. Garments are made from very refined and expensive materials, with meticulous attention to detail. Every single article is made almost entirely by hand and has prices in the order of a few tens of thousands of euros (Mazza and Alvarez, 2017).

Despite its small-to-medium size with less than fifty employees, PTN enjoys a certain reputation in the field of *haute couture* and commonly receives orders from several renowned designers. Production is typically organised in lots of a few dozen homologous garments. The contractor usually provides the raw materials and the relevant instructions for making them. The maniacal attention to detail at every production stage is a guarantee of the quality of the final products, which are often "flaunted" by VIPs and public figures at events of public resonance. Hence the need for PTN to produce garments with virtually no defects, in order to preserve its image in the eyes of the customer or third parties. Consequently, in addition to great care during production, PTN carries out a careful inspection of the products in post-production, before their delivery.

3.2 Typical acceptance-sampling activities

Given that it is impractical to inspect 100% of products due to time constraints and the limited availability of specialized technical personnel, PTN usually adopts acceptance-sampling schemes according to which only a portion (i.e., sample) of each *lot* is inspected and – based on the result of the inspection – a decision (*pass* or *fail*) is made on the entire lot. Any *rejection*, meaning that the estimated quality of the lot is lower than desired, implies the extension of the inspection to the entire lot; furthermore, nonconformities found during inspections are commonly repaired through additional processing. This statistical approach, which is commonly known as *acceptance-sampling plan with rectification*, represents a good compromise between effective control of the production output and reasonable inspection cost/time (Montgomery, 2019).

PTN, like many other companies, is guided by existing standards in choosing the most appropriate acceptance-sampling schemes. Favouring simplicity, PTN opts for sampling plans for *attributes*, with *single* samples (therefore also referred to as *single sampling plans* or SSPs) and a *normal* inspection level, as stipulated in ISO 2859-1:1999[3] (1999). The logic of SSPs with rectification can be summarized as follows:

- Select a level of *defectiveness* (or *fraction nonconforming*) deemed acceptable for the supply. This level is commonly known as AQL, standing for "acceptable quality level";
- Depending on the desired lot size (N) and AQL value, determine the SSP's parameters: essentially sample size (n) and acceptance number (c);
- Inspect the n units of the sample and determine the number of defective units (d), i.e., the number of garments containing a combination of nonconformities/defects of some general severity. An example of the categorisation of nonconformities/defects according to their severity in the *haute-couture* sector can be found in Sect. A.1, in the appendix.
- Apply the traditional SSP's LDD rule:

$$\text{if } d \leq c \Rightarrow pass, \text{ if } d > c \Rightarrow fail. \tag{1}$$

- In case of rejection (*fail*), extend the inspection to the $(N - n)$ units of the lot outside the sample. Any nonconformities found on the garments during inspection are typically *rectified*, resulting in additional costs.

PTN inspections, like most inspections in the *haute-couture* context, are characterized by *multiple inspectors* who examine the same garments independently, with no exchange or mutual influence.

[3] In the sampling plans for *attributes*, the inspected units are classified as *conforming* or *nonconforming*. ISO 2859-1:1999 (1999) provides three inspection levels – *reduced*, *normal* and *tightened* – to be chosen depending on the desired level of protection/severity and the results of previous inspections.

This approach creates “blind” inspections, where assessments are based solely on the individual observations of the inspectors. There are several reasons for this redundancy. First, conformity assessments inevitably include a certain degree of subjectivity; therefore, the involvement of multiple inspectors – not infrequently with complementary skills and viewpoints – contributes to a more homogeneous assessment. Moreover, given that inspectors often carry out manufacturing operations, sharing inspection results is also a useful feedback and alignment opportunity.

The PTN’s approach might seem paradoxical at first: on the one hand, the inspection covers only a portion of the lot to limit costs and time, while on the other hand, each unit in the sample is inspected by multiple inspectors. However, this strategy is relatively common in specific contexts, such as *haute couture*, where inspections are inherently subjective and frequent changes of product type make it difficult to standardize assessments between inspectors. Consistency among inspectors is crucial, especially when they are supposed to be interchangeable with each other (Keist, 2015). Furthermore, in *haute couture*, the main defects are aesthetic rather than structural, which makes it advantageous to increase visual inspections on the same sample units, thereby maximizing the likelihood of detecting all defects before the products reach the customer (cf. Sect. A.1, in the appendix). The risk inherent in the inspection of the sample alone is mitigated by the possibility of extending the (multiple-inspector) inspection to the entire lot in the event of rejection.

Independent assessments by multiple inspectors are also useful for identifying potential misalignments among inspectors and, when necessary, for identifying "outlier" inspectors who may require targeted corrective actions or additional training (Sect. A.2, in the appendix, contains a realistic example of a training test). Furthermore, beyond their independence, the fact that inspections are *blinded* ensures that each inspector maintains full individual responsibility for the quality of his/her inspection work, unlike collective inspections where responsibility is shared and decisions are made jointly[4]. Consequently, this inspection process also provides an opportunity to evaluate the performance of each inspector individually.

Having clarified the notions of SSP with rectification and multiple inspectors, let us try to combine them together. The scheme in Figure 1 shows that, first, a sampling scheme of the initial lot must be identified (see insets (a) and (b)). Next, the inspection is performed by several inspectors acting separately to avoid mutual interference or conditioning. The result is represented by a set of

[4] *Independent blinded* assessments, where each inspector operates independently and without knowledge of the results from other inspectors performing the same inspection, help prevent the dilution of responsibility often encountered in collective inspections involving joint decision-making. As Deming (2018) observed, "*Incidentally, 200 percent inspection, as usually carried out, is less reliable than 100 percent inspection for the simple reason that each inspector depends on the other to do the job. Divided responsibility means that nobody is responsible*".

conformity ("✓") or *nonconformity* ("✗") assessments of the individual product units, which can be summarized in an *assessment matrix* (see inset (c)). A delicate aspect of this procedure is the aggregation of these multiple-inspector assessments, so as to determine an unambiguous final LDD decision for the entire supply (see inset (d)). There are many possible ways of performing this aggregation, also considering the lack of standards/guidelines in the scientific literature. Sect. 4 exemplifies several possible aggregation approaches observed in common practice within the *haute-couture* sector.

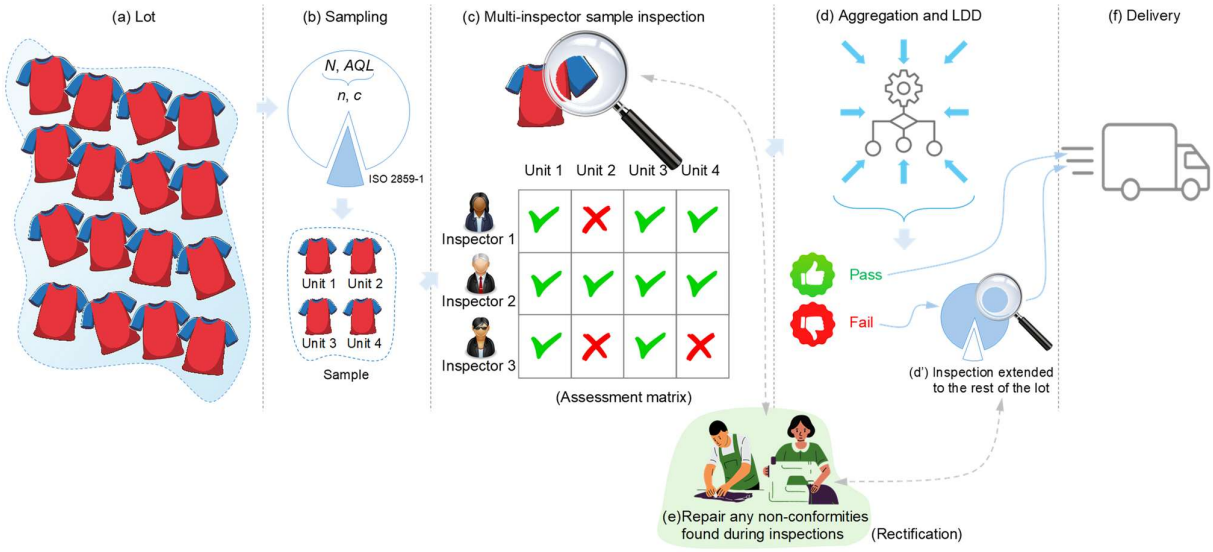

Figure 1. Scheme of an SSP with rectification, integrated with multiple-inspector conformity assessments.

4. Methodology

This section focuses on the possible approaches to combine multiple-inspector conformity assessments with SSPs. It is organised in four subsections: the first conceptualises how to aggregate the information content of the assessment matrix; the second introduces six possible aggregation approaches; the third exemplifies seven real-world situations where the application of the above aggregation approaches may lead to conflicting and doubtful results; the fourth introduces Gwet's κ_G, i.e., an indicator depicting the agreement between inspectors, which can be useful for identifying controversial situations and triggering appropriate corrective actions (Gwet, 2008; 2014; 2015).

4.1 Aggregation "by row" and "by column"

Returning to the case study (cf. Sect. 3), let us assume that an SSP for *attributes*, for *defectives* and with *rectification* is used to verify the conformity of PTN's production before delivery to the customer. After setting desired values for lot size (N) and AQL, the ISO 2859-1:1999 (1999) standard allows to determine the SSP parameters n and c. As seen in Sect. 3, conformity assessment of sample

units is commonly performed by a set of (s) inspectors. Although the number of inspectors can vary depending on their availability and the type of garments to be analysed, it is generally of the order of magnitude of a few units (e.g., 3 to 5).

The conformity assessment of the single product units is carried out independently by the inspectors, adopting appropriate guidelines. The different types of defects/imperfections can be traced back to several predefined macro categories, associated with certain levels of severity (e.g., *minor*, *major* and *critical*). Section A.1 (in the appendix) presents a detailed description of the typical categorization of garment defects in the *haute-couture* sector. When a certain unit undergoes inspection by an inspector, any detected defects are documented. Subsequently, the conformity assessment of the unit is determined by a conventional rule, considering both the quantity and severity of defects from the perspective of each inspector. For instance, garments with a few minor defects may often be quickly repaired during the inspection itself and then be classified as conforming ("✓"), while those with even one critical defect are classified as nonconforming/defective ("✗") (Keist, 2015). The results of the multiple-inspector inspections are summarised in an assessment matrix.

To integrate the information content of the assessment matrix with a traditional SSP and achieve an unequivocal LDD for the entire lot, it is necessary to introduce appropriate adaptations. Considering the conceptual diagram in Figure 2, the process begins with the assessment matrix (0), which contains the assessments by individual inspectors (s_1, s_2, … , s_s), regarding the conformity of each inspected unit (u_1, u_2, … , u_n). The transition from (0) to (2) can be achieved by means of two sequential aggregations that may follow two alternative paths: the first via the "external column" in (1.a) (see grey arrows), and the second via the "external row" (1.b) (see white arrows), as described in the caption of Figure 2.

Before detailing the aggregations, let us now indulge in a brief conceptual digression. Figure 2 includes four (coloured) stripes, two horizontal ones and two vertical ones. The upper horizontal stripe (in yellow colour) indicates *local* outputs related to individual inspectors (from s_1 to s_s), while the lower stripe (in green colour) indicates *global* outputs related to the totality of inspectors. Regarding the vertical stripes, the left one (in blue colour) shows assessments related to *single units*, while the right one (in red colour) shows assessments related to the *whole lot*, namely the LDDs. The intersection of the four stripes outlines the four regions of Figure 2: (0), (1.a), (1.b), and (2). In fact, the assessment matrix in (0) contains local conformity assessments related to single units; (1.a) contains the column of local LDDs, reflecting the viewpoints of individual inspectors; (1.b) contains the row of global conformity assessments related to single units and representative of the entire group

of inspectors; finally, (2) displays the global LDD, or more simply the *final* LDD, relating to the entire set of inspectors and the whole lot.

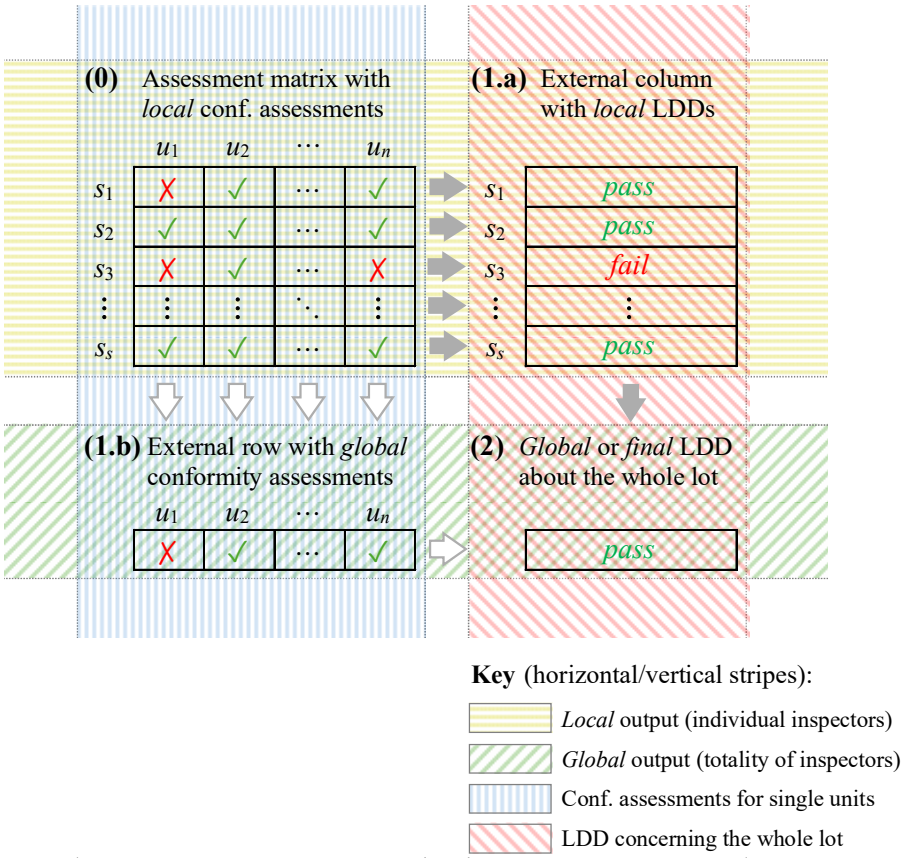

Figure 2. Conceptual scheme illustrating the sequence of aggregations, either *by row* then *by column* or *vice versa*, used to translate the assessment matrix (0) into a final LDD (2) for the whole lot. The "external column" (1.a) and "external row" (1.b) are derived from the aggregation of the elements of the assessment matrix, *by row* and *by column* respectively. The same aggregations (by column and by row) can be applied to the external column and external row respectively, in order to obtain the final LDD in (2). There are two possible ways to obtain the final LDD: **(0) → (1.a) → (2)** (represented by grey arrows), and **(0) → (1.b) → (2)** (represented by white arrows). The result may vary depending on the sequence of aggregation and the specific criteria used in these aggregations.

Below, the aggregations by row and by column are described.

The **aggregation by row**, which is applicable to the individual rows of the assessment matrix (0) or to the external row (1.b), is generally obtained through the implementation of the SSP's LDD rule (cf. Sect. 3). In less frequent cases, this aggregation may be performed using other criteria, such as:

- *Majority*. The LDD of *pass* or *fail* is determined by the *majority* of the assessments of "✓" or "✗" respectively;
- *Unanimous conformity*. The LDD of *pass* is determined by the unanimity of assessments of "✓", otherwise the LDD results in a *fail*;

- *At least r conformities (out of n)*. The LDD of *pass* is achieved if at least *r* conformity assessments of "✓" out of *n* are met, otherwise the LDD results in a *fail*; the conventional parameter *r* does not necessarily depend on *c*. This criterion can also be referred to as the "*SSP's LDD*" rule, due to its correspondence with the rule expressed in Eq. 1, where $r = n - c$. The criterion of *majority* can be considered as a special case where $r = \left\lfloor \frac{n}{2} \right\rfloor + 1$, and that of *unanimous conformity* as another special case where $r = n$.

Extending the approach already illustrated for aggregation by row, the **aggregation by column** – applicable to the individual columns of the assessment matrix (0) or to the external column (1.a) – can be obtained through various criteria, such as *majority*, *unanimous conformity*, *at least r conformities*. The global output may be expressed as "✓"/"✗" for the columns of the assessment matrix, or as *pass*/*fail* for the external column, with the total number of local assessments being *s* (as opposed to *n*, in the case of aggregation by row).

Different combinations of aggregation sequences (e.g., *by row* and then *by column*, or *vice versa*) and aggregation criteria can generate practically innumerable different aggregation approaches (Arrow and Suzumura, 2010; Felsenthal and Nurmi, 2018; Franceschini et al., 2022).

4.2 Examples of aggregation approaches

Figure 3 illustrates six possible aggregation approaches commonly used in the *haute-couture* sector, based on the same assessment matrix with *s* = 5 and *n* = 3. The first three approaches (#1, #2, and #3) apply an aggregation by row followed by an aggregation by column, each adopting different criteria, while the next three approaches (#4, #5, and #6) use the opposite sequence. Approaches #1, #2, and #4 incorporate SSP's LDDs with *c* = 1. In this particular case, all six approaches incidentally yield the same final LDD result of ***pass***. Sect. A.3 (situation *a*, in the appendix) illustrates a step-by-step application of these approaches.

In general, establishing the absolute superiority of any single approach over others is inherently complex (Arrow et al., 2010; Franceschini et al., 2022). This complexity arises from the need to consider various factors, such as ease of implementation, severity, simplicity in management, tendency toward undecidability, robustness to small input-data variations, etc. Furthermore, the suitability of each approach may vary significantly depending on the application context. For example, in less critical situations, it may be suitable to adopt more relaxed approaches, such as those that aggregate results based on *majority* rather than *unanimous-pass* criteria. Conversely, in more critical situations where greater caution is needed, more stringent approaches may be appropriate.

Traditionally, the statistical criterion used to compare sampling plans of different types consists in evaluating their *discriminatory power* (i.e., their *severity*, from the supplier's perspective, or *protection level*, from the customer's perspective) from a quantitative standpoint by constructing the corresponding *Operating Characteristic* (OC) curve[5]. However, in the case of multiple-inspector assessment aggregations, constructing the OC curve becomes significantly more complex since it must account not only for the aggregation approach used but also for the subjectivity inherent in each inspector's assessments. This subjectivity stems from their individual level of technical competence and training. In fact, inspectors may often give different assessments of the same product units, either due to (i) errors in failing to detect "hidden" defects and/or (ii) errors in identifying *false defects*. In other words, each inspector – due to the errors he/she makes – perceives an *apparent* defectiveness, which may differ from the *real* one and must be appropriately modelled (Franceschini et al., 2016). To construct an accurate OC curve, it is therefore necessary to properly estimate the apparent defectiveness rate perceived by each inspector, taking into account his/her tendency to make such evaluation errors. This aspect will be further investigated in future research.

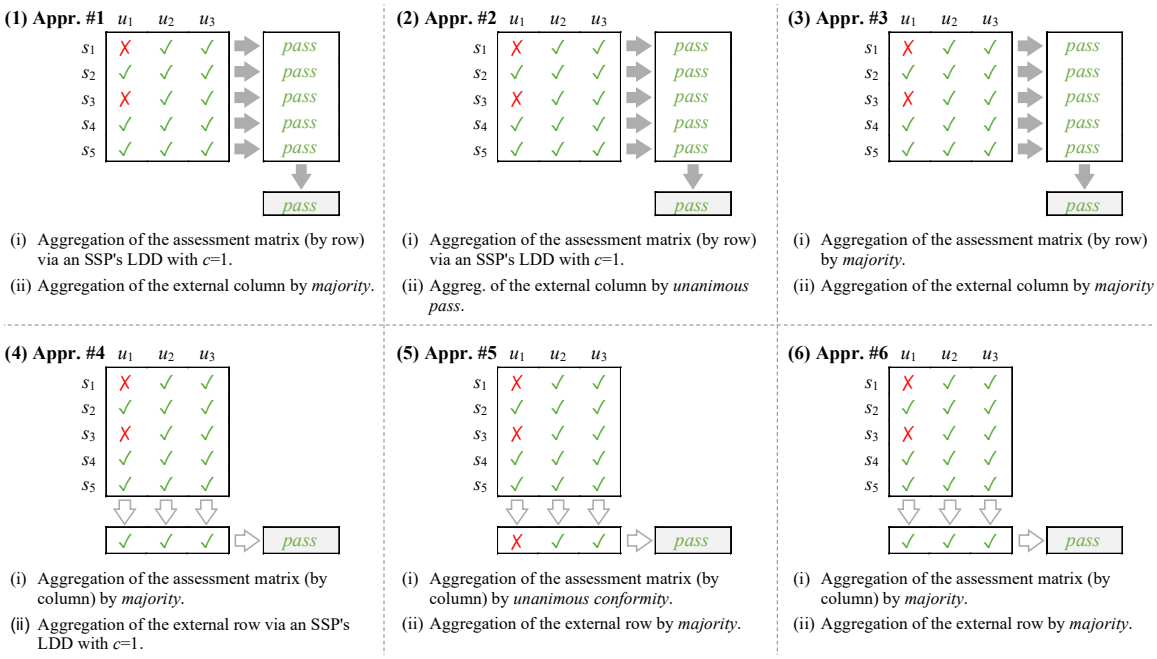

Figure 3. Aggregation of the same assessment matrix, with $s = 5$ inspectors (s_1 to s_5) and $n = 3$ product units (u_1 to u_3), using six different aggregation approaches. For each approach, the aggregation *type* (i.e., first by row and then by column or *vice versa*) and *criteria* are specified.

[5] The OC curve of a sampling plan graphically represents the probability of accepting a lot as a function of the proportion of defective products in that lot. The OC curve is a key tool in quality control, which is used to assess the discriminatory power of the sampling plan between acceptable and unacceptable lot-quality levels (Montgomery, 2019).

4.3 Real-world application

This subsection empirically compares the six aggregation approaches provided in Sect. 4.2, in seven practical situations (*a*, *b*, *c*, *d*, *e*, *f*, and *g*). Figure 4 presents the SSP parameters obtained by applying the ISO 2859-1:1999 (1999) standard (with *normal* inspection level) and the relevant assessment matrix, situation by situation. As specified in Sect. 3.2, this standard allows for the identification of SSP parameters based on a given AQL and inspection level, but it does not account for the presence of multiple inspectors. Since, in the seven practical situations, the AQL values are very similar and the inspection level remains the same (cf., Figure 4), it can be demonstrated that – excluding the aggregation of multiple-inspector assessments – the resulting SSPs do not significantly differ in terms of discriminatory power. However, it will be shown that when multiple-inspector assessments are aggregated, the results can vary significantly depending on the level of agreement (or disagreement) among inspectors and the specific aggregation approach adopted.

The seven situations under investigation are real cases observed and selected within PTN over two months; the different values of the parameters n, c and s are related to contingent reasons (e.g., different values of N and AQL, different availability of inspectors, etc.). It is noteworthy that the number ($s \cdot n$) of replicate inspections conducted by multiple inspectors on the sample units may sometimes be even greater than the total number (N) of inspections that would occur if a single inspector inspected (alone) the entire lot. Consider, for example, situation *e*, where $s \cdot n = 5 \cdot 3 = 15 > N = 12$. This apparent paradox, however, is justified by the considerations already outlined in Sect. 3.2 (mostly visual inspections, inherent subjectivity, not complete interchangeability between inspectors, etc.).

		Situation *a*	Situation *b*	Situation *c*	Situation *d*	Situation *e*	Situation *f*	Situation *g*
Parameters	N	15	18	15	20	12	18	14
	AQL	0.015	0.015	0.02	0.015	0.01	0.02	0.025
	n	3	4	3	5	3	4	3
	c	1	1	1	1	1	1	1
	s	5	3	2	3	5	3	2
Assess. matrix		u_1 u_2 u_3 s_1 ✗ ✓ ✓ s_2 ✓ ✓ ✓ s_3 ✗ ✓ ✓ s_4 ✓ ✓ ✓ s_5 ✓ ✓ ✓	u_1 u_2 u_3 u_4 s_1 ✓ ✓ ✗ ✗ s_2 ✗ ✓ ✓ ✗ s_3 ✗ ✓ ✓ ✗	u_1 u_2 u_3 s_1 ✓ ✗ ✗ s_2 ✓ ✗ ✓	u_1 u_2 u_3 u_4 u_5 s_1 ✓ ✓ ✗ ✗ ✗ s_2 ✓ ✗ ✓ ✗ ✗ s_3 ✗ ✓ ✓ ✗ ✗	u_1 u_2 u_3 s_1 ✓ ✓ ✗ s_2 ✓ ✗ ✓ s_3 ✗ ✓ ✓ s_4 ✗ ✗ ✗ s_5 ✗ ✗ ✗	u_1 u_2 u_3 u_4 s_1 ✓ ✓ ✓ ✓ s_2 ✗ ✗ ✓ ✗ s_3 ✓ ✓ ✗ ✗	u_1 u_2 u_3 s_1 ✗ ✓ ✓ s_2 ✓ ✗ ✗

Figure 4. Overview of the seven practical situations (*a*, *b*, *c*, …, *g*) in which the six aggregation approaches introduced in Sect. 4.2 are applied. These situations are sorted decreasingly according to the κ_G indicator, which is provided in Table 1 and described in Sect. 4.4.

Sect. A.3 (in the appendix) illustrates the step-by-step application of the six aggregation approaches and Table 1 summarises the results obtained. Interestingly, though not surprisingly, the results of the six alternative approaches often display discrepancies. Only in situation *a*, all approaches converge

to the same final LDD of *pass*. In the other situations, results can be discrepant or doubtful, as also evidenced by the "?" symbols, denoting undecidability. Additionally, Table 1 shows that controversial and doubtful results are more likely to occur in situations with a small number of experts and poor agreement in terms of (local) conformity assessments. Particularly problematic are aggregation approaches that apply majority-based criteria, especially when the number of experts (s) – in cases where majority aggregation is applied by columns, as in approaches #1, #3, #4, and #6 – and/or the number of inspected units (n) – in cases where majority aggregation is applied by rows, as in approaches #3, #5, and #6 – are even values. This is evident in situations *c* and *g*, where s is even, and in situations *b* and *f*, where n is even (see Table 1).

Situat.	N	AQL	n	c	s	Appr. #1	Appr. #2	Appr. #3	Appr. #4	Appr. #5	Appr. #6	κ_G
a	15	0.015	3	1	5	*pass*	*pass*	*pass*	*pass*	*pass*	*pass*	0.74
b	18	0.015	4	1	3	*fail*	*fail*	?	*fail*	*fail*	?	0.33
c	15	0.02	3	1	2	?	*fail*	?	?	*fail*	?	0.33
d	20	0.015	5	1	3	*fail*	*fail*	*fail*	*fail*	*fail*	*pass*	0.23
e	12	0.01	3	1	5	*pass*	*fail*	*pass*	*fail*	*fail*	*fail*	-0.15
f	18	0.02	4	1	3	*fail*	*fail*	?	*pass*	*fail*	*pass*	-0.30
g	14	0.025	3	1	2	?	*fail*	?	?	*fail*	?	-1.00

Table 1. Results of the application of the six approaches introduced in Sect. 4.2 to the seven situations schematised in Figure 4. The symbol "?" denotes a result of *undecidability*. The last column shows the corresponding κ_G values, as illustrated in Sect. 4.4.

While the number of experts is an easily countable quantity, the agreement among inspectors is more difficult to quantify. The next sub-section proposes the use of an indicator to carry out this latter quantification, in order to highlight potentially controversial situations in advance.

4.4 Evaluating the agreement between inspectors

The scientific literature includes a plurality of indicators to evaluate the degree of agreement of different inspectors (or *agents*), who independently *rate* certain units (Banerjee et al., 1999; Falotico and Quatto, 2015; Dettori and Norvell, 2020; Franceschini and Maisano, 2021; Franceschini et al., 2022). Very popular are the indicators of the *kappa* family, with the common structure of a ratio between (i) a proxy for the degree of agreement observed in the ratings of interest (numerator) and (ii) a proxy for the degree of agreement that would occur in the case of random ratings (denominator):

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{2}$$

where

P_o is a sub-indicator of the degree of agreement observed on a conventional scale from 0 (maximum disagreement) to 1 (maximum agreement);

P_e is a sub-indicator of the degree of agreement expected if inspectors were to express ratings randomly.

The numerator of the fraction in Eq. 2 can also be interpreted as the "*agreement in excess with respect to random agreement*", while the denominator as the "*maximum achievable excess agreement*" (Banerjee et al., 1999; Tang et al., 2015). From this very general structure introduced by Cohen (1960), a plurality of possible variants of the *kappa* indicator derive, depending on the different definitions of the sub-indicators (P_o and P_e) and the peculiarities of the application context (e.g., number of inspectors, number of categories, nominal/ordinal scale, etc.). Each variant of kappa has its *pros* and *cons* that are not discussed here. For the application context of interest, which is characterised by more than two agents (inspectors), rating certain units (garments) on a nominal binary scale ("✓", "✗") – the version of *kappa* proposed by Gwet (κ_G) is more paradox-resistant than other versions (Fleiss, 1971; Fleiss, 1981; Gwet, 2014; Falotico and Quatto, 2015; Feinstein and Cicchetti, 1990a, 1990b). For this reason, κ_G will be used to assess the inter-inspector agreement in the case study. A brief description of κ_G follows, along with an application example.

Starting from the structure in Eq. 2, the sub-indicator P_o is determined as the average proportion of agreement observed from the perspective of each of the units considered. Returning to the example in Figure 3, with s=5 inspectors evaluating n=3 units, the proportion of agreement relative to the generic i-th unit can be defined as (Gwet, 2014):

$$P_{u_i} = \frac{1}{s\cdot(s-1)} \cdot \sum_{j=1}^{k}\left[s_{u_i,j} \cdot (s_{u_i,j} - 1)\right], \tag{3}$$

where

u_i is the i-th unit;

k is the number of categories used for assessment; in this case k=2 since the categories are "✓" and "✗";

$s_{u_i,j}$ denotes the number of inspectors who assigned the i-th unit to a certain j-th category.

Considering the assessment matrix in Figure 3, the following three P_{u_i} proportions are thus determined:

$$\begin{aligned} P_{u_1} &= \frac{1}{s\cdot(s-1)} \cdot \sum_{j=1}^{k}\left[s_{u_1,a} \cdot \left(s_{u_1,j} - 1\right)\right] = \frac{1}{5\cdot(5-1)} \cdot (3\cdot 2 + 2\cdot 1) = \frac{8}{20} = 0.4 \\ P_{u_2} &= \frac{1}{s\cdot(s-1)} \cdot \sum_{j=1}^{k}\left[s_{u_2,a} \cdot \left(s_{u_2,j} - 1\right)\right] = \frac{1}{5\cdot(5-1)} \cdot \left(5\cdot 4 + 0\cdot(-1)\right) = \frac{20}{20} = 1. \\ P_{u_3} &= \frac{1}{s\cdot(s-1)} \cdot \sum_{j=1}^{k}\left[s_{u_3,a} \cdot \left(s_{u_3,j} - 1\right)\right] = \frac{1}{5\cdot(5-1)} \cdot \left(5\cdot 4 + 0\cdot(-1)\right) = \frac{20}{20} = 1 \end{aligned} \tag{4}$$

Next, P_o is determined through the arithmetic mean of the proportions in Eq. 4:

$$P_o = \frac{1}{n} \cdot \sum_{i=1}^{n} P_{u_i} = \frac{(0.4+1+1)}{3} = 0.8. \quad (5)$$

Shifting the focus to P_e, in the case of binary categories (such as j=1=✓ and j=2=✗ in our case), it is defined as (Gwet, 2014):

$$P_e = \sum_{j=1}^{k} [p_{e,j} \cdot (1 - p_{e,j})], \quad (6)$$

where the term $p_{e,j}$ represents the proportion of agreement that would occur, under the assumption that the ratings related to a specific unit (expressed by different inspectors) were randomly permuted (Banerjee et al. 1999):

$$p_{e,j} = \frac{1}{n \cdot s} \cdot \sum_{i=1}^{n} s_{u_i,j}. \quad (7)$$

The following condition holds: $\sum_{j=1}^{k} p_{e,j} = 1$. Returning to the assessment matrix in Figure 3, it is obtained:

$$\begin{aligned} p_{e,✓} &= \frac{1}{n \cdot s} \cdot \sum_{i=1}^{n} s_{u_i,✓} = \frac{3+5+5}{3 \cdot 5} = 0.87 \\ p_{e,✗} &= \frac{1}{n \cdot s} \cdot \sum_{i=1}^{n} s_{u_i,✓} = \frac{2+0+0}{3 \cdot 5} = 0.13 \end{aligned}. \quad (8)$$

Applying Eq. 6, it is obtained:

$$P_e = 0.87 \cdot (1 - 0.87) + 0.13 \cdot (1 - 0.13) = 0.23. \quad (9)$$

Finally, in line with the general model in Eq. 2, κ_G can be determined as:

$$\kappa_G = \frac{P_o - P_e}{1 - P_e} = \frac{0.8 - 0.23}{1 - 0.23} = 0.74. \quad (10)$$

It can be demonstrated that the domain of κ_G lies in the range [-1,1] (Gwet, 2014; Falotico and Quatto, 2015). A value of 1 indicates perfect agreement between inspectors, while a value of 0 indicates agreement attributable only to chance. A negative value of κ_G indicates worse agreement than chance, which can also be interpreted as systematic disagreement between inspectors. Although there are no unambiguous thresholds for the interpretation of κ_G values, several authors recommended those reported in Table 2. Looking at this scale of interpretation, we can conclude that $\kappa_G < 0.40$ denote potentially doubtful situations, with relatively low inter-inspector agreement.

Range	Inter-inspector agreement
$\kappa_G < 0.20$	Poor
$0.20 \leq \kappa_G < 0.40$	Fair
$0.40 \leq \kappa_G < 0.60$	Moderate
$0.60 \leq \kappa_G < 0.80$	Good
$\kappa_G \geq 0.80$	Very good

Table 2. Commonly used scale of interpretation for κ_G statistic (Landis and Koch, 1977; Dettori and Norvell, 2020).

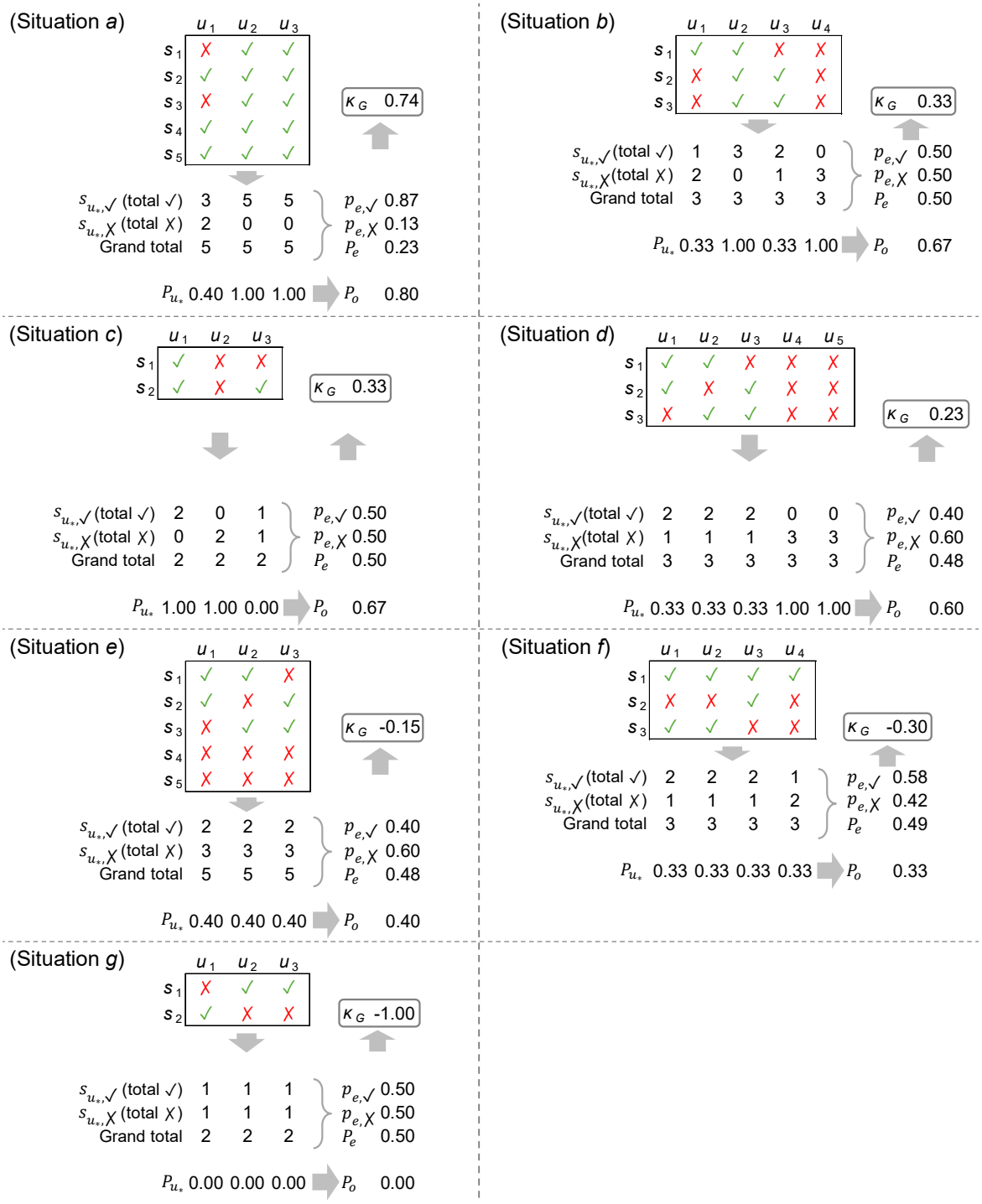

Figure 5. Calculation of κ_G for each of the seven practical situations (*a*, *b*, *c*, *d*, *e*, *f*, *g*) in Figure 4. The sub-indicators ($s_{u_i,j}$, $p_{e,j}$, P_{u_i}, P_e, P_o, with $j \in \{\checkmark, \text{✗}\}$) are described in Sect. 4.4.

Returning to the seven situations exemplified in Table 1, the corresponding κ_G values can be calculated and are reported in Figure 5, which also provides the relevant sub-indicator values ($s_{u_i,j}$, $p_{e,j}$, P_{u_i}, P_e, P_o). The only situation with a relatively high value of κ_G is the first one (0.74), in which

all six approaches lead to the same (*pass*) result (cf. Table 1). In the remaining situations, especially those with doubtful or conflicting results, κ_G responds with relatively low (sometimes even negative) values. These examples therefore show that κ_G can be useful in practice to point out potentially problematic situations in advance, characterized by poor inter-inspector agreement (Kelly, 1989; Franceschini and Maisano, 2018; Franceschini et al., 2022). In these cases, it could be appropriate to trigger corrective actions, such as:

- carry out (joint) re-inspection of the garments that resulted in the most discordant assessments;
- institute additional training for those "misaligned" inspectors (Sect. A.2, in the appendix, exemplifies a training carried out in PTN);
- reinforce conformity assessments by introducing additional (blinded independent) inspections conducted by new inspectors, preferably with advanced technical skills and experience.

The final LDD for the entire lot may be reconsidered after any (further) re-inspections. The selection of the most appropriate corrective actions, typically determined by quality experts, may vary on a case-by-case basis depending on various factors such as the operational context, company policy, customer requirements, etc. Future research will aim to further explore this aspect.

5. Conclusions

The article focused on (post-production) conformity inspections performed by multiple inspectors, which are typically adopted for high-value-added, small-scale, and highly customized productions. This is the case with productions in sectors where safety is of the utmost importance (e.g. nuclear energy, aerospace, defence, etc.) and/or where nonconformities could lead to serious damage to the brand image and dissatisfaction in the customer (e.g., luxury goods, *haute couture*, etc.). A further distinguishing feature is that inspections are almost exclusively manual and require a certain amount of experience from inspectors. Given the high value of the product units subject to inspection, any defects found are usually repaired through appropriate intervention; therefore, rectification sampling schemes are commonly adopted (Montgomery, 2019). Moreover, in these contexts, product models are often highly customised and innovative, with small lots featuring new models. Consequently, inspections may also provide valuable feedback to refine and stabilise production techniques and protocols, especially for newly introduced models.

Although the literature includes a plethora of sampling plan schemes (e.g., *single*, *sequential*, *double*, *skip-slot*, by *variables*, by *attributes*, etc.), they rarely refer to the case of multiple-inspector inspections. This article specifically focused on this case, with a particular emphasis on the aggregation of conformity assessments. By means of some real-world examples, it was showed that

when several inspectors inspect the same product units concurrently, aggregating the relevant conformity assessments becomes necessary. While countless aggregation methods exist and no universally accepted standard has been established, two characteristic features can usually be distinguished (cf. **RQ#1**):

- *Aggregation sequence*. For example, the conformity assessments by individual inspectors may be aggregated *before* applying the LDDs rule according to the sampling plan in use. Alternatively, multiple LDDs may be obtained locally (i.e., at the level of individual inspectors) and the results aggregated *afterwards*.
- *Aggregation criteria.* For example, aggregation based on the *majority*, *unanimous conformity*, or other criteria.

Adapting classical sampling systems to the multiple-inspector case is a non-trivial challenge, since there are no tools to establish the absolute superiority of one aggregation approach over the other (Franceschini et al., 2023). In other words, all possible aggregation approaches are questionable and may sometimes provide discrepant or even conflicting results (Arrow et al., 2010; Franceschini and Maisano, 2019; Franceschini et al., 2023). For this reason, the authors recommend avoiding excessive reliance on any single approach and instead, whenever possible, adopting multiple approaches concurrently. The convergence of results across different techniques can serve as an indication of the relative robustness of the solution; conversely, significant discrepancies should be taken as a warning sign (cf. concept of *wisdom of crowds*) (Franceschini et al., 2022).

The focus of the paper was on the *haute couture* sector, where multiple-inspector conformity assessments are commonly carried out. Based on a real-life case study, six different aggregation approaches were compared, exemplifying several practical situations in which they may yield doubtful or discrepant results. It was also shown that these critical situations most likely occur when the inter-inspector agreement is weak and/or the number of inspectors is small. In addition, majority-based aggregation approaches are hardly applicable with fewer than three inspectors or when the number of inspectors and/or the sample size are even quantities (cf. **RQ#2**).

While it is straightforward to assess the number of inspectors, it is less intuitive to quantify their degree of agreement. To this purpose, this research proposed the use of a statistical indicator, i.e., Gwet's κ_G, as it appears appropriate to the problem of interest. In practice, κ_G can be used as a diagnostic tool before applying any aggregation approach, to pre-emptively identify potentially problematic situations (cf. **RQ#3**). When it is relatively low (e.g., below 0.4, cf. Table 2), it is advisable to avoid hazarding a final LDD that lacks robustness, but rather "take a step back" to investigate the reasons for the misalignment between inspectors, and trigger any corrective actions

(e.g., additional training, more inspectors, etc.). Preventive corrective measures, such as improving the training of inspectors or revising inspection protocols, can then be taken, depending on the judgement of quality experts on a case-by-case basis.

The research pursued has some limitations. Firstly, only a very simple sampling scheme was considered, i.e., an SSP for attributes, the construction of which was guided by ISO 2859-1:1999 (1999). Furthermore, Gwet's κ_G was used as a preliminary alert system to flag situations that require further attention. This does not prevent more complex sampling schemes from being investigated in the future or κ_G from being replaced/complemented by a similar statistical indicator of agreement.

Regarding the future, the investigation will be extended to other areas where multiple-inspector inspections are commonly carried out, beyond *haute couture* (e.g. aerospace or defence). In addition, alternative indicators of agreement beyond κ_G will be explored, with a focus on identifying the most effective actions to address problematic situations as they arise. Finally, a quantitative comparison between multiple-inspector sampling plans is planned, based on the construction of OC curves that take into account the number of inspectors, their propensity for assessment errors, and the characteristics of the aggregation approach adopted.

Acknowledgments: This research was carried out under the MICS (Made in Italy, Circular and Sustainable) Extended Partnership and partially funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza – Missione 4, Componente 2, Investimento 1.3, D.D. 1551.11-10-2022, PE00000004). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Appendix

A.1 Categorization of garment defects

A fundamental requirement of *haute-couture* garments is the "maniacal" care aimed at eliminating possible defects/imperfections that must not reach the end customer. For this reason, the garment manufacturing process includes quality-control and conformity-assessment operations at several stages, such as fabric making, dyeing, printing, spreading, cutting, trimming, stitching, pattern making, finishing, washing, and packing (Ngan et al., 2011). The intention is to identify and correct any defects as soon as possible and in the appropriate places, avoiding dragging them on to subsequent processes. This section focuses on the inspections of finished garments, which are particularly delicate since: (i) they represent the last opportunity to remedy any defects before shipment to the final customer, and (ii) they are meticulously carried out by a plurality of inspectors, who can provide important feedback to improve the entire manufacturing process (Sthephens 2001; Schilling and Neubauer, 2009).

It is common practice in the *haute-couture* field to contemplate garment defects/nonconformities of various kinds, classifying them into three severity categories: *critical*, *major* and *minor* (in decreasing order); the following paragraphs provide a preliminary description in this regard (Hasi and Das, 2011).

Critical defects. A defect that poses a danger to consumer safety and/or violates mandatory standards is defined as critical. Generally, if a critical defect is found in a garment, the customer tends to reject the entire order/lot. A critical defect could harm the customer, compromise the image of retailers throughout the supply chain, damage brand reputation and result in unnecessary expenses in the event of a product recall. Some examples of critical defects in garments are:

- Presence of a needle or other sharp foreign objects included in the finished product packaging;
- Blood stains on the garment (as a result of accidental injury by an operator).
- Broken button;
- Presence of mould on the garment;
- Loose trim and fasteners improperly fastened;
- Drawstring near head or neck in garments for babies and children;
- Excessively long or loose threads or trimmings;
- Lack of warning labels regarding choking and/or more general labelling errors (e.g. wrong size, wrong model, etc.).

Major defects. These defects result in a general dysfunction of the garment or reduced usability that prevents it from being placed on the market. Although they do not pose a threat to user safety, they generally reduce the value of the garment, negatively affect its marketability and saleability, shorten the life cycle, and increase returns for replacement or refund. Some examples of major defects in garments are:

- Open seam;
- Hole in fabric;
- Broken stitch;
- Incorrect colour/design on the garment;
- Damaged fabric;
- Bubbling due to fusing;
- Defective hinge;
- Button not securely fastened (while remaining intact);
- Oil stain.

Minor defects. Minor defects are unlikely to reduce the usability of the product. These are defects in workmanship that go beyond the specifications or requirements agreed with the customer. These defects can usually be repaired with relatively little effort as soon as they are identified (Keist, 2015). Examples of minor defects in garments are:

- Misprinting of the label on a shipping carton;
- Untrimmed thread, missing stitch or uneven stitching on a garment;
- Minor variation in shading between homologous garment pieces;
- Minor error(s) in the label;
- Dirty material that can be easily cleaned.

Figure 6 contains photographs of some actual defects, classified in the *critical* (*a* and *b*) and *major* (*c*, *d* and *e*) categories.

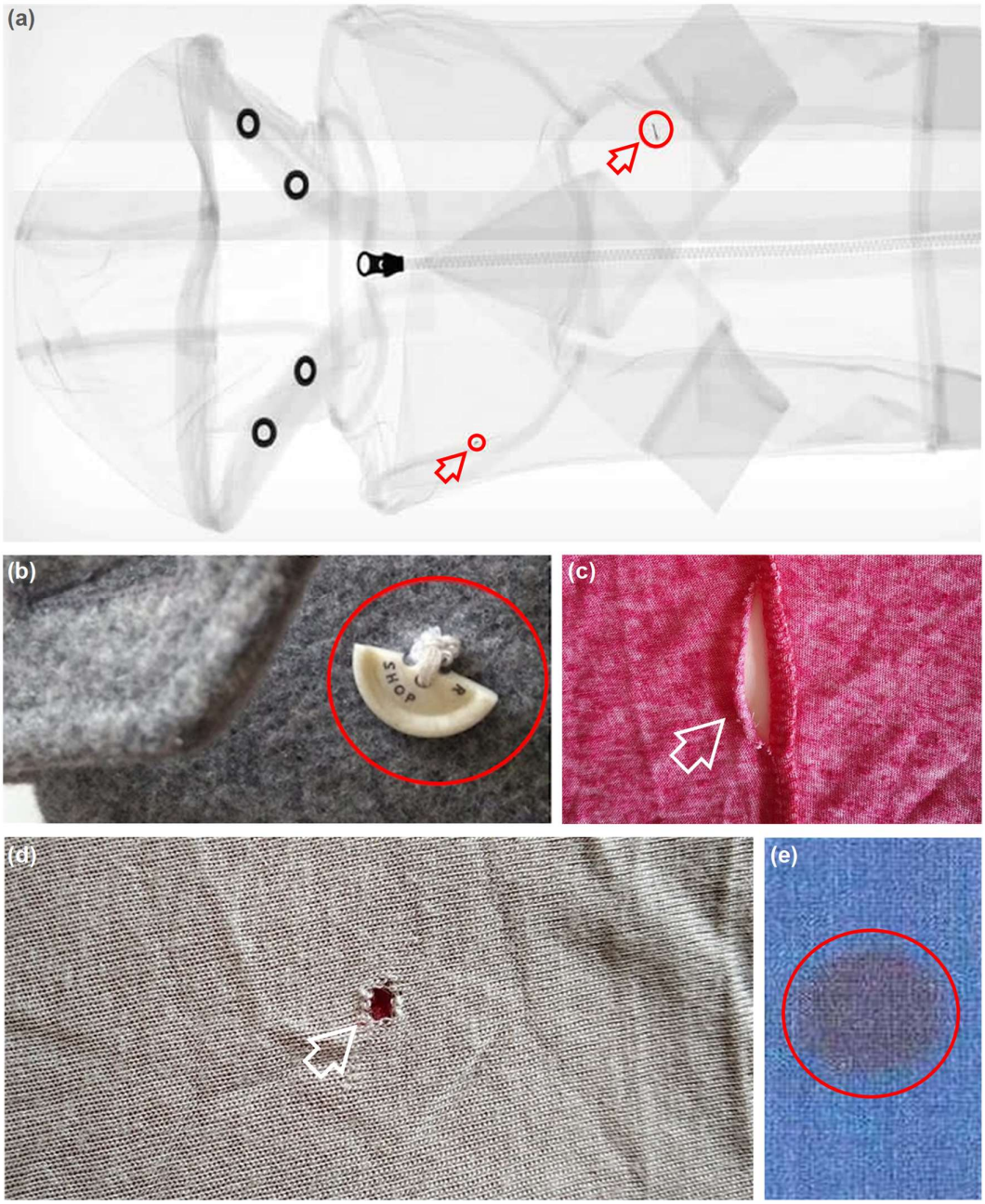

Figure 6. Examples of garment defects: (a) presence of a needle or other sharp object included in the finished product packaging (*critical*), (b) broken button (*critical*), (c) open seam (*major*), (d) hole in fabric (*major*), and (e) oil stain (*major*).

A.2 Training-test example

Two delicate aspects of conformity inspections on *haute-couture* garments can be (i) the non-absolute homogeneity in the inspectors' judgement, and (ii) the fact that inspections are exclusively manual, with inevitable subjectivity, largely linked to the inspectors' experience. For these reasons, it is common practice for several inspectors to independently inspect the same garments, effectively

conducting replicated inspections. Additionally, periodic training is a valuable tool for promoting uniformity of judgment among inspectors. Training activities can take various forms, from preparing simple inspection checklists to offering online courses, coaching sessions, or on-field briefings.

As an example, let us illustrate a training test in the PTN company (cf. Sect. 3) to assess the degree of uniformity in the inspectors' judgements and, consequently, plan possible actions to strengthen their preparation. The test is structured in several steps. First, some specific garments, representative of the company's typical production, are identified. In this specific case: a *bomber jacket*, a *trench coat*, a *windbreaker*, and a *peacoat*. These garments deliberately include several defects (resulting from actual or artfully created manufacturing), as specified in the first two columns of the table in Figure 7. The four garments are individually inspected by each of the (seven) inspectors taking part in the test. The test lasts 10 minutes for each garment (thus 40 minutes in total), within which the inspector has to identify and classify the defects. In doing so, the inspectors are asked to follow a guided procedure, aimed at increasing the effectiveness of inspections while avoiding redundant or unnecessarily operations. The main steps of the procedure adopted by PTN are described below.

Garment	Defect description	Severity	Insp. 1	Insp. 2	Insp. 3	Insp. 4	Insp. 5	Insp. 6	Insp. 7	% of detection per defect
1. Bomber jacket	1.1 Missing quality-control sticker	Critical	✓	✓	✓	✓	✗	✓	✓	85.7%
	1.2 Damaged drawstring pull	Critical	✓	✓	✓	✓	✓	✓	✗	85.7%
	1.3 Missing barcode on polybag packaging	Critical	✓	✓	✓	✓	✓	✓	✗	85.7%
	1.4 Pulled threads inside garment	Major	✓	✓	✗	✓	✓	✓	✓	85.7%
	1.5 Hole in the front	Major	✓	✓	✓	✓	✗	✓	✓	85.7%
2. Trench coat	2.1 Cut composition label	Critical	✗	✓	✗	✓	✗	✗	✓	42.9%
	2.2 Open inner-pocket seam	Critical	✗	✓	✗	✓	✗	✗	✗	28.6%
	2.3 Scratched/damaged button	Critical	✓	✓	✗	✓	✓	✓	✗	71.4%
	2.4 Stained logo label	Critical	✓	✓	✓	✓	✓	✓	✗	85.7%
	2.5 Sleeve seam undone	Major	✗	✓	✓	✓	✗	✓	✓	71.4%
3. Windbreaker	3.1 Stain on inner fabric	Major	✓	✓	✓	✓	✓	✓	✓	100.0%
	3.2 Unsealed thermal tape	Major	✓	✓	✓	✓	✓	✓	✓	100.0%
	3.3 Incomplete swing ticket	Critical	✗	✓	✗	✓	✓	✗	✗	42.9%
	3.4 Partially unstitched logo label	Critical	✓	✓	✓	✓	✓	✓	✗	85.7%
	3.5 Scratched drawstring pull	Critical	✓	✓	✓	✓	✓	✓	✓	100.0%
4. Peacoat	4.1 Loose button placket	Major	✓	✓	✓	✓	✓	✓	✓	100.0%
	4.2 Pulled threads	Major	✓	✓	✓	✓	✗	✓	✗	71.4%
	4.3 Stained composition label	Critical	✗	✓	✗	✓	✓	✓	✗	57.1%
	4.4 Detached collar seam	Major	✗	✗	✓	✓	✗	✓	✓	57.1%
	4.5 Pencil mark under right button	Major	✓	✓	✓	✓	✓	✗	✗	71.4%
	% of detection per inspector		70%	95%	70%	100%	65%	80%	50%	

Figure 7. Results of the test on seven PTN inspectors, with reference to the detection of a set of defects (intentionally created) on four different garments. "✓" and "✗" represent detected and undetected defects respectively. The last column shows the percentage of detection of each specific defect (in a specific row) by all inspectors, while the bottom row shows the percentage of defect detection for each specific inspector (in a specific column).

- **Prerequisites.** Checking the availability of supporting documentation for inspections: drafts, drawings, reference specifications of garment characteristics, lists of defect types and their severity level, etc. (e.g., see Figure 8).
- **Packaging check.** Check that the garment packaging is compliant and that the information on the tag matches that on the bar code affixed to the outer nylon wrapping.
- **Label check.** Check that the information on the outer tag matches that on the inner labels, that all the required labels are present and that they are applied in the correct place and order. Also, check print quality, size, integrity, etc.
- **Check general appearance.** Check that the general appearance of the garment, including the symmetry/proportions of its parts, conforms to the customer's requirements.
- **Accessories check.** Check the quality and functionality of the accessories on the garment. For example, for buttons, check that they fit correctly in the buttonholes, check the correct functioning of zips, snaps, etc.
- **Workmanship check.** Check the quality of the workmanship of the garment, the quality of the fabric and the final ironing.
- **Measurements.** Checking that the garment measurements correspond to the customer's approved specifications and are within the established specifications.
- **Defect identification.** Drawing up a detailed report of all defects detected, classifying them according to a reference list.

Figure 8. Example of a drawing with associated dimensional specifications for certain types of garments, to support inspectors during inspection. For confidentiality reasons, numerical data are hidden.

The columns of Figure 7 contain the types of defects in the test, their respective degree of severity, and their detection ("✓") or non-detection ("✗") by the seven inspectors involved in the test. It is thus possible to determine the percentage of defects detected by each inspector (in the row at the bottom), which is a proxy for the degree of inspection effectiveness. PTN management has conventionally determined that a percentage of less than 70% is inadequate, necessitating programmes to strengthen the skills of the inspectors concerned. In the example, inspectors 5 and 7 failed the test and will be subject to the strengthening programme. It can also be noted that some defect types tend to be more difficult to detect than others, as depicted by the overall percentage of detection of each defect by inspectors, reported in the last column. For example, the most difficult defects to detect are 2.1 and 2.2 on the trench coat. This other piece of information may be useful for management to evaluate possible actions to strengthen the inspectors' skills, if it is realised that certain types of defects are detected with difficulty at a generalised level.

A.3 Details on application examples

The following section shows in more detail the application of the six aggregation approaches in Sect. 4.2 to seven practical situations in Sect. 4.3.

Situation *a*

In this situation, characterised by N=15, AQL=1.5%, s=5 inspectors, and an SSP with parameters n=3 and c=1 (see Figure 3), all six approaches agree in providing a final ***pass*** result, as described below.

Approach #1. In this case, the assessment matrix is aggregated by row by applying the SSP's LDD rule to the (s) rows, resulting in as many local LDDs (*pass* or *fail*), which are reported in the external column. Subsequently, the external-column elements are aggregated into a final LDD, using the *majority* criterion.

Returning to the conformity assessments in Figure 3(1), since the number (d) of defective units detected by each inspector never exceeds the acceptance number specified by the SSP (i.e., $d \leq c = 1$), the application of local LDDs results in *pass* for all inspectors (see the first external column). According to the majority criterion, the final LDD will be ***pass*** too.

Approach #2. This approach is similar to the previous one, except that the aggregation by column uses the local LDDs (in the external column). In this case, the global *pass* is conditional on the local LDDs being *unanimously pass*, which is certainly more stringent than the criterion of majority. Returning to the previous example (with n=3 and c=1), the unanimous local LDDs of *pass* would imply a final LDD of ***pass*** (see Figure 3(2)).

Approach #3. In this case, an initial aggregation by row of the assessment matrix is performed, using the majority criterion. The resulting *local* LDDs (in the external column) are then aggregated by column, using the majority criterion again and resulting in a final LDD. It is worth noting that this approach only uses part of the SSP parameters, i.e., *n* but not *c* (Hati and Das, 2011; Mazza and Alvarez, 2017). Returning to the previous example, the totality of local LDDs of *pass* (in the second external column) would imply a final LDD of ***pass*** (see Figure 3(3)).

Approach #4. In this case, an initial aggregation by column of the assessment matrix is performed, using the majority criterion. The resulting *global* assessments (in the external row) are then aggregated by applying the SSP's LDD rule at a global level, resulting in a final LDD. Returning to the previous example, the following global conformity assessments are obtained (in the first external column):

$$u_1 \Rightarrow \checkmark, u_2 \Rightarrow \checkmark, u_3 \Rightarrow \checkmark. \tag{A.1}$$

from which $d = 0 \leq c = 1$, resulting in a final SSP's LDD of ***pass*** (see Figure 3(4)).

Approach #5. In this case, an initial aggregation by column of the assessment matrix is performed, using the criterion of unanimous conformity. Next, the elements of the external row are aggregated (by row), via the SSP's LDD rule, resulting in a final LDD. This policy is more precautionary than the previous one, as a single (local) "X" is sufficient to determine a corresponding one globally. Returning to the previous example (with *n*=3 and *c*=1), the following global conformity assessments are obtained (in the second external column):

$$u_1 \Rightarrow \mathsf{X}, u_2 \Rightarrow \checkmark, u_3 \Rightarrow \checkmark, \tag{A.2}$$

from which $d = 1 \leq c = 1$, resulting in a final SSP's LDD of ***pass*** (see Figure 3(5)).

Approach #6. Initially, an aggregation by column of the assessment matrix is performed, using the majority criterion. The resulting *global* conformity assessments (in the external row) are then aggregated by row, resulting in a global LDD. Returning to the previous example, the following global conformity assessments are obtained (in the first external row):

$$u_1 \Rightarrow \checkmark, u_2 \Rightarrow \checkmark, u_3 \Rightarrow \checkmark. \tag{A.2}$$

Next, the SSP's LDD rule is applied at a global level ($d = 0 \leq c = 1$), determining a final LDD of ***pass*** (see Figure 3(6)).

Situation *b*

This situation is characterised by N=18, AQL=1.5%, s=3 inspectors, and an SSP with parameters n=4 and c=1. Figure 9 summarises the results obtained through the six aggregation approaches, which are briefly described below.

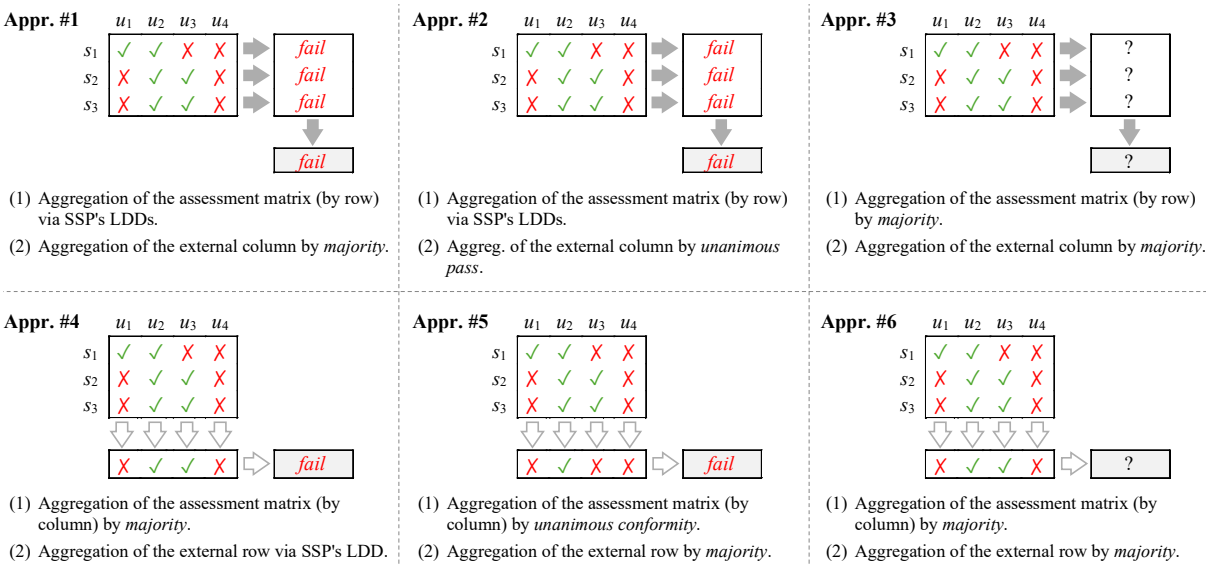

Figure 9. *A*ggregation of the assessment matrix related to the situation *b* (cf. Figure 4), using six different aggregation approaches. For each approach, the aggregation type (i.e., first by row and then by column or *vice versa*) and criteria are specified.

Approaches #1 and #2. Applying the SSP's LDD rule to aggregate the individual-inspector assessments by row, it is obtained (see the external column in Figure 9(1) and (2)):

$$(s_1, s_2, s_3)\ d = 2 > c = 1 \ \Rightarrow fail, \tag{A.4}$$

which results in a final LDD of ***fail*** when aggregating by column, either by *majority* or *unanimous pass*.

Approach #3. Applying the *majority* criterion to aggregate the individual-inspector assessments by row results in *undecidability* for any inspector (see the second external column). Consequently, the final LDD will be undecidability too (symbol "**?**").

Approach #4. The application of the *majority* criterion to the single columns of the assessment matrix results in (see the external row in Figure 9(4)):

$$u_1 \Rightarrow X, u_2 \Rightarrow \checkmark, u_3 \Rightarrow \checkmark, u_4 \Rightarrow X. \tag{A.5}$$

Then, the aggregation by row using the (final) SSP's LDD rule results in $d = 2 > c = 1 \Rightarrow \boldsymbol{fail}$.

Approach #5. The application of the *unanimous-pass* criterion to the assessment matrix results in (see the external row in Figure 9(5)):

$$u_1 \Rightarrow X, u_2 \Rightarrow \checkmark, u_3 \Rightarrow X, u_4 \Rightarrow X. \tag{A.6}$$

Then, the aggregation by row using the (final) SSP's LDD rule results in $d = 3 > c = 1 \Rightarrow \boldsymbol{fail}$.

Approach #6. Applying the *majority* criterion to the single columns of the assessment matrix yields the same results as in Eq. A.5 (see the external row in Figure 9(6)). In the next aggregation by row, no *majority* of "✓" or "✗" is obtained, resulting in *undecidability* (symbol "**?**").

Situation *c*

Considering a lot with *N*=15 and having imposed AQL=2% with *normal* inspection level, the ISO 2859-1 standard determines an SSP with *n*=3 and *c*=1. In this case, there are *s*=2 inspectors. Figure 10 summarises the results obtained through the six aggregation approaches, which are briefly described below.

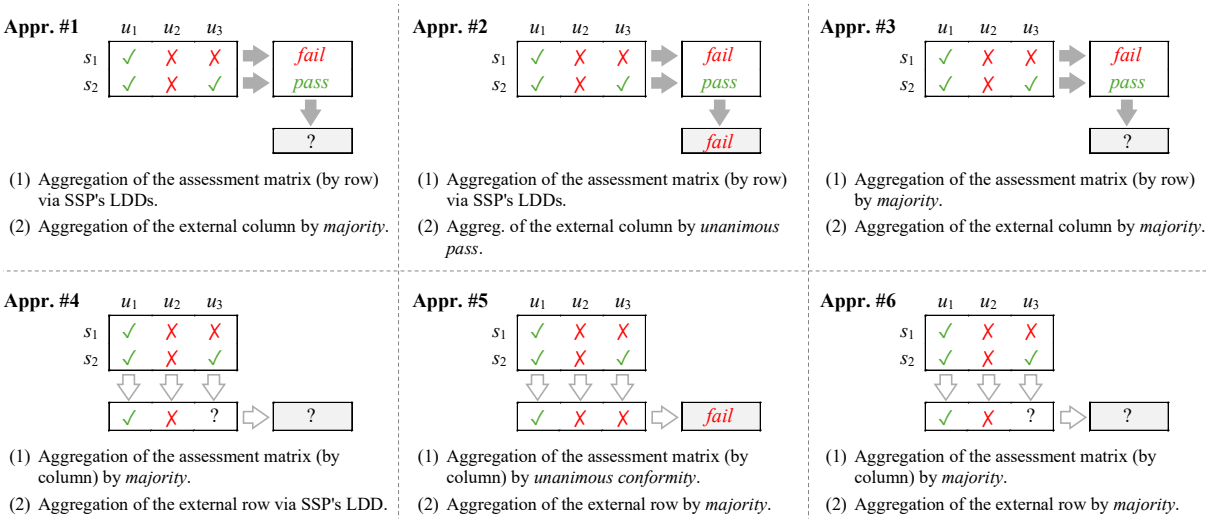

Figure 10. Aggregation of the assessment matrix related to the situation *c* (cf. Figure 4), using six different aggregation approaches. For each approach, the aggregation *type* (i.e., first by row and then by column or *vice versa*) and *criteria* are specified.

Approach #1. Applying the SSP's LDD rule to aggregate the assessment matrix by row, it is obtained (in the external column in Figure 10(1)):

$$\begin{aligned} &(s_1)\ d = 2 > c = 1 \Rightarrow fail, \\ &(s_2)\ d = 1 \leq c = 1 \Rightarrow pass, \end{aligned} \qquad \text{(A.7)}$$

Since no majority is reached (neither as *pass* nor *fail*), the final LDD results in *undecidability* (symbol "**?**").

Approach #2. Failing to meet the condition of unanimous *pass* locally (see the external column), a final SSP's LDD of ***fail*** is determined.

Approach #3. Applying the *majority* criterion locally for individual inspectors yields the same result in Eq. A.7 (see the external column in Figure 10(3)). Since no majority is reached (neither as *pass* nor *fail*), the final LDD results in *undecidability* (symbol "**?**").

Approaches #4 and #6. The aggregation of the assessment matrix by column, using the *majority* criterion, leads to the following result (in the external row in Figure 10(4) and (6)):

$$u_1 \Rightarrow \checkmark, u_2 \Rightarrow X, u_3 \Rightarrow ? \tag{A.8}$$

The subsequent aggregation (by row), either through the SSP's LDD rule or through the *majority* criterion, leads to *undecidability* ("**?**"), as it is not determined whether $d \leq c = 1$ (i.e. *pass* condition for both criteria) or $d > c$ (*fail* condition).

Approach #5. The aggregation of the assessment matrix by column, using the *unanimous-conformity* criterion, leads to the following result (second external row in Figure 10(5)):

$$u_1 \Rightarrow \checkmark, u_2 \Rightarrow X, u_3 \Rightarrow X. \tag{A.9}$$

The subsequent aggregation (by row) through the *majority* criterion therefore leads to ***fail***.

Situation *d*

As *N*=20 and AQL=1.5%, using the ISO 2859-1 standard with *normal* inspection level, the SSP's parameters *n*=5 and *c*=1 are determined. In this case *s*=3. Figure 11 illustrates the application of the six aggregation approaches to the corresponding assessment matrix.

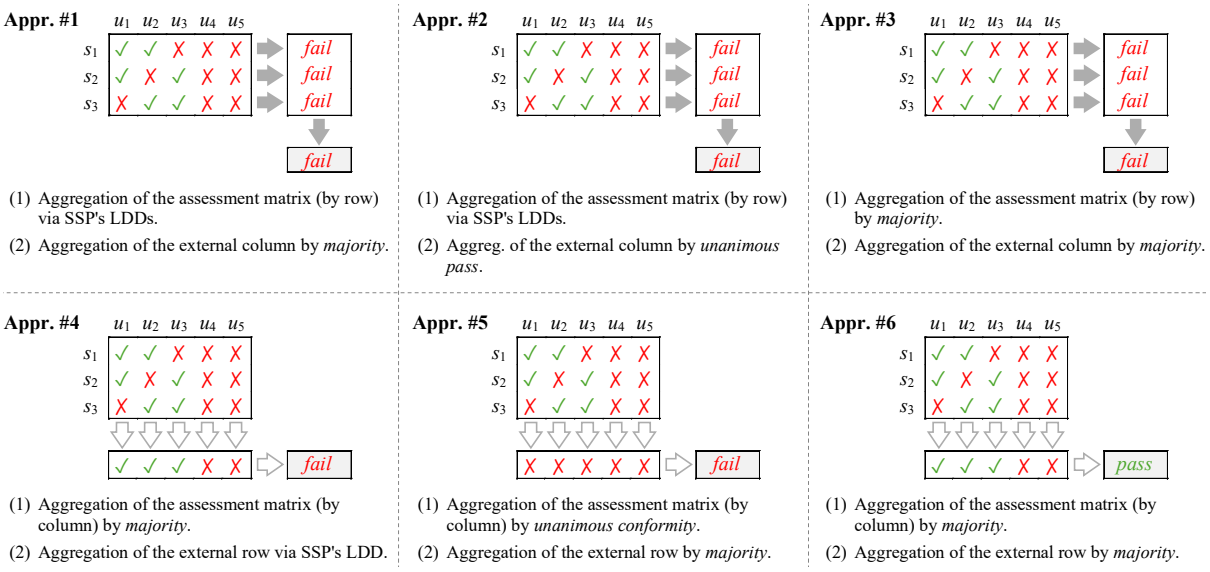

Figure 11. *A*ggregation of the assessment matrix related to the situation *d* (cf. Figure 4), using six different aggregation approaches. For each approach, the aggregation type (i.e., first by row and then by column or *vice versa*) and criteria are specified.

Approaches #1 and #2. Applying the SSP's LDD rule to aggregate the assessment matrix by row, it is obtained (see the external column in Figure 11(1) and (2)):

$$(s_1, s_2, s_3)\ d = 3 > c = 1\ \Rightarrow fail. \tag{A.10}$$

Then, the final LDD is a ***fail***, both when aggregating (by column) by *majority* and by *unanimous pass*.

Approach #3. The aggregation by row of the assessment matrix with the *majority* criterion produces the same result in Eq. A.10 (see the external column in Figure 11(3)). Thus, the final LDD, obtained by a further aggregation by column, is a ***fail***.

Approach #4. The application of the *majority* criterion to the single columns of the assessment matrix results in (see the external row in Figure 11(4)):

$$u_1 \Rightarrow \checkmark, u_2 \Rightarrow \checkmark, u_3 \Rightarrow \checkmark, u_4 \Rightarrow \text{X}, u_5 \Rightarrow \text{X}. \quad \text{(A.11)}$$

Then, the aggregation by row using the (final) SSP's LDD rule results in $d = 2 > c = 1 \Rightarrow \boldsymbol{fail}$.

Approach #5. The application of the *unanimous-conformity* criterion to the single columns of the assessment matrix results in a global non-conformity for all five units (see the external row in Figure 11(5)). Then, the aggregation by row using the (final) SSP's LDD rule results in $d = 5 > c = 1 \Rightarrow \boldsymbol{fail}$.

Approach #6. The application of the *majority* criterion to the single columns of the assessment matrix produces the same results in Eq. A.11 (see the external row in Figure 11(6)). The subsequent aggregation by row with the *majority* criterion then produces a final ***pass***.

Situation *e*

This is similar to situation *a* in terms of SSP parameters (i.e., $n = 3$ and $c = 1$) and number of inspectors ($s = 5$). However, the inspection results are different, as shown in the assessment matrix in Figure 12. Applying the six different approaches leads to the results summarised below.

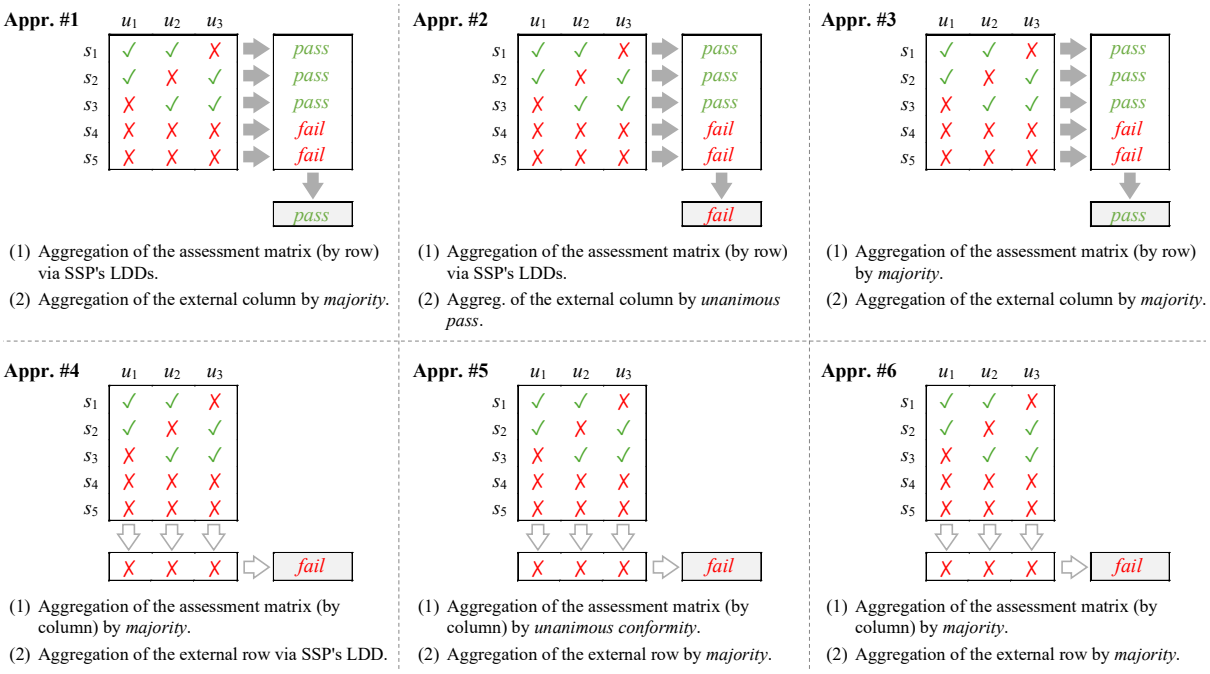

Figure 12. Aggregation of the assessment matrix related to the situation *e* (cf. Figure 4), using six different aggregation approaches. For each approach, the aggregation *type* (i.e., first by row and then by column or *vice versa*) and *criteria* are specified.

Approach #1. Applying the SSP's LDD rule at the level of single-inspector assessments results in (see the external column in Figure 12(1)):

$$\begin{aligned}&(s_1)\ d = 1 \le c = 1 \Rightarrow pass,\\ &(s_2)\ d = 1 \le c = 1 \Rightarrow pass,\\ &(s_3)\ d = 1 \le c = 1 \Rightarrow pass,\\ &(s_4)\ d = 3 > c = 1 \Rightarrow fail,\\ &(s_5)\ d = 3 > c = 1 \Rightarrow fail,\end{aligned} \tag{A.12}$$

from which, based on the *majority* principle (i.e., 3 out of 5), a final LDD of ***pass*** is determined.

Approach #2. Since the *unanimous pass* condition is not fulfilled in the external column (see Figure 12(2)), the final SSP's LDD is of ***fail***.

Approach #3. Applying the *majority* criterion at the level of single-inspector assessments results in the local results reported in the external column in Figure 12(3), which, coincidentally, is equal to that one in Figure 12(1). Then, the final LDD results in a ***pass***, which corresponds to the *majority* of local conformity assessments.

Approach #4. Each of the three units inspected is classified as nonconforming by the *majority* of the inspectors. Therefore, the global conformity assessments are (see the external row in Figure 12(4)):

$$u_1 \Rightarrow X, u_2 \Rightarrow X, u_3 \Rightarrow X, \tag{A.13}$$

from which $d = 3$. Applying the SSP's LDD rule, the final result is of ***fail***.

Approach #5. Since this approach is certainly more stringent than approach #3, the final result will also be ***fail***.

Approach #6. After aggregation by column and by majority of the assessment matrix, the result is the same as in Eq. A.13. Thus, the final LDD results in a ***fail***, which corresponds to the majority of "✗" assessments.

In Fig. 12 the comparison between approach #3 and approach #6 highlights the presence of a condition attributable to the Ostrogorski paradox [Kelly, 1989].

Situation *f*

Let us consider a lot with N=18, AQL=2% and an SSP with n=4 and c=1. In this case, the company provides s=3 inspectors to perform conformity assessments. Figure 13 summarises the results obtained through the six aggregation approaches, which are briefly described below.

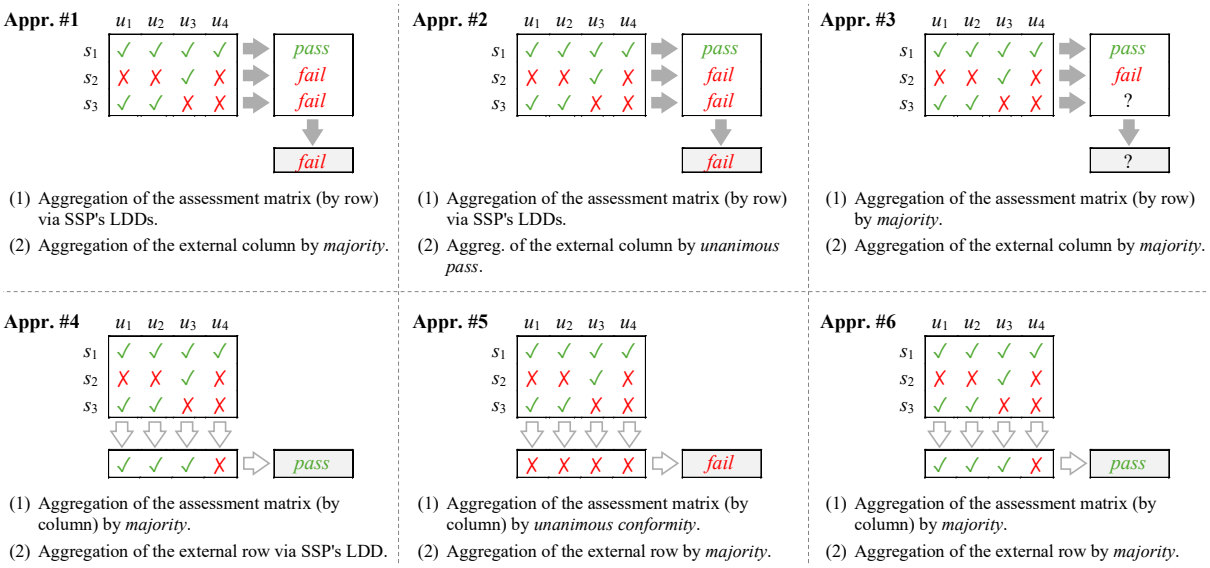

Figure 13. Aggregation of the assessment matrix related to the situation *f* (cf. Figure 4), using six different aggregation approaches. For each approach, the aggregation type (i.e., first by row and then by column or *vice versa*) and criteria are specified.

Approach #1. Applying the SSP's LDD rule to aggregate the assessment matrix by row, it is obtained (see the external column in Figure 13(1)):

(A.14)

$$(s_1)\ d = 0 \leq c = 1 \ \Rightarrow pass,$$
$$(s_2)\ d = 3 > c = 1 \ \Rightarrow fail,$$
$$(s_3)\ d = 2 > c = 1 \ \Rightarrow fail.$$

Aggregating (by column) by *majority*, the final LDD is therefore a ***fail***.

Approach #2. Failing to meet the condition of unanimous *pass* locally (see the same results in Eq. A.14), a final SSP's LDD of ***fail*** is determined.

Approach #3. Applying the *majority* criterion at local level for individual inspectors gives the following result (cf. the external column in Figure 13(3)):

(A.15)

$$(s_1) \Rightarrow pass,$$
$$(s_2) \Rightarrow fail,$$
$$(s_3) \Rightarrow ?$$

The subsequent aggregation (by column), through the *majority* criterion, leads to *undecidability* ("**?**").

Approach #4. The aggregation of the assessment matrix by column, using the *majority* criterion, leads to the following result (see the external row in Figure 13(4)):

$$u_1 \Rightarrow \checkmark, u_2 \Rightarrow \checkmark, u_3 \Rightarrow \checkmark, u_4 \Rightarrow \text{X}. \quad \text{(A.16)}$$

The subsequent aggregation (by row), through the SSP's LDD rule, results in $d = 1 \leq c = 1 \Rightarrow$ $\boldsymbol{pass}$.

Approach #5. The aggregation of the assessment matrix by column, using the criterion of *unanimous conformity*, leads to the following result (see the external row in Figure 13(5)):

$$u_i \Rightarrow \text{X}, \forall i. \quad \text{(A.17)}$$

The subsequent aggregation (by row), through the SSP's LDD rule, results in $d = 4 > c = 1 \Rightarrow$ $\boldsymbol{fail}$.

Approach #6. Applying the *majority* criterion to the single columns of the assessment matrix yields the same results as in Eq. A.16 (see the external row in Figure 13(6)). The next aggregation by row, using the *majority* criterion, results in a final ***pass***.

Situation *g*

This situation is characterized by the parameters $n = 3$, $c = 1$ and $s = 2$. Figure 14 summarises the results obtained through the six aggregation approaches, which are briefly described below.

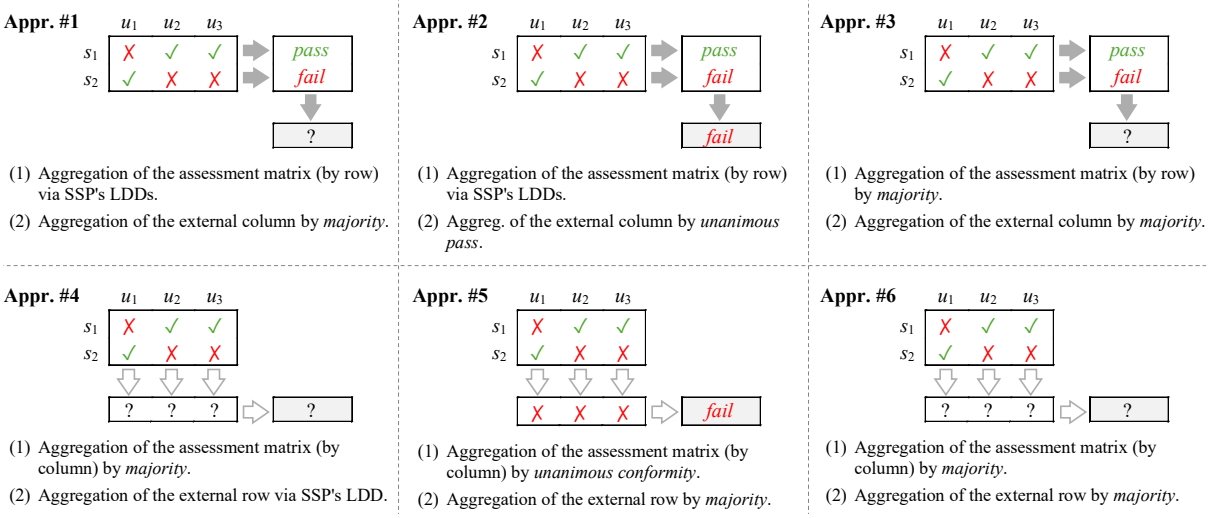

Figure 14. Aggregation of the assessment matrix related to the situation *g* (cf. Figure 4), using six different aggregation approaches. For each approach, the aggregation *type* (i.e., first by row and then by column or *vice versa*) and *criteria* are specified.

Approach #1. Applying the SSP's LDD rule locally for single inspectors, it is obtained (see the external column in Figure 14(1)):

$$(s_1)\ d = 1 \leq c = 1 \ \Rightarrow pass,$$
$$(s_2)\ d = 2 > c = 1 \ \Rightarrow fail, \tag{A.18}$$

Since no majority is reached (neither as *pass* nor *fail*), the final LDD is *undecidable* (symbol "**?**").

Approach #2. Failing to meet the condition of unanimous *pass* locally (see the external column in Figure 14(2)), a final SSP's LDD of ***fail*** is determined.

Approach #3. Applying the *majority* criterion locally for individual inspectors yields the same result in Eq. A.18 (see the external column in Figure 14(3)). Since no majority is reached (neither as *pass* nor *fail*), the final LDD results in *undecidability* (symbol "**?**").

Approaches #4 and #6. For each of the three units, no *majority* of conformity/nonconformity assessments are obtained, generating *undecidability* situations (see the external row in Figure 14(4) and (6)):

$$u_1 \Rightarrow ?, u_2 \Rightarrow ?, u_3 \Rightarrow ?, \tag{A.19}$$

where the symbol "?" indicates *indeterminacy* of the global assessment. Therefore, no further aggregation (by row) can be performed and the resulting final LDD is "**?**".

Approach #5. The global conformity assessments are of nonconformity for all three units (see the external row in Figure 14(5), so the final LDD will be ***fail***.

6. References

Arrow, K.J., Sen, A., Suzumura, K. (2010). Handbook of social choice and welfare (Vol. 2). North Holland, Elsevier.

Aurum A., Petersson H., Wohlin C. (2002), *State-of-the-art: software inspections after 25 years*. Software Testing, Verification and Reliability, 12, 133-154. doi: 10.1002/stvr.243.

Banerjee M., Capozzoli M, McSweeney L., Sinha D. (1999). *Beyond kappa: A review of interrater agreement measures*, The Canadian Journal of Statistics, 27(1),3-23.

BS 6001-0:2006 (2006) *Sampling procedures for inspection by attributes - Part 0 and Part-1*. London, UK.

Chandra, J. and Schall, S. (1998), *The use of repeated measurements to reduce the effect of measurement errors*. IIE Trans., 20, 83–87.

Chun Y.H. (2009), *Improving product quality by multiple inspections: Prior and posterior planning of serial inspection procedures*, IIE Transactions, 41(9), 831-842, doi: 10.1080/07408170802389324.

Chun Y.H. (2016), *Improved method of estimating the product quality after multiple inspections*, International Journal of Production Research, 54(19):5686-5696. doi: 10.1080/00207543.2015.1128128

Cohen J. (1960) *A coefficient of agreement for nominal scales*, Educational and Psychological Measurement,20(1), 37-46, doi: 10.1177/001316446002000104.

Deming, W.E. (2018). Out of the Crisis, reissue. MIT press, Cambridge, ISBN 9780262535946.

Dettori J.R., Norvell D.C. (2020) *Kappa and Beyond: Is There Agreement?* Global Spine Journal. 2020;10(4):499-501. doi:10.1177/2192568220911648

Duffuaa S.O., Khan M. (2005) *Impact of inspection errors on the performance measures of a general repeat inspection plan*. International Journal of Production Research. 43(23), 4945–4967. doi:10.1080/00207540412331325413.

Duffuaa S.O., El-Ga'aly A. (2015). Impact of inspection errors on the formulation of a multi-objective optimization process targeting model under inspection sampling plan. Computers & Industrial Engineering, 80, 254-260.

Falotico, R., Quatto P. (2015) *Fleiss' kappa statistic without paradoxes*. Quality & Quantity. 49, 463–470. doi:10.1007/s11135-014-0003-1.

Feinstein, A.R., Cicchetti D.V. (1990a), *High agreement but low kappa: I. The problems of two paradoxes*. Journal of Clinical Epidemiology. 43(6), 543–549. doi: 10.1016/0895-4356(90)90158-L

Feinstein, A.R., Cicchetti D.V. (1990b), *High agreement but low kappa: II. Resolving the paradoxes*. Journal of Clinical Epidemiology. 43(6), 551–558. doi: 10.1016/0895-4356(90)90159-M.

Felsenthal D.S., Nurmi H. (2018) Voting Procedures for Electing a Single Candidate, Springer, Cham, Switzerland.

Ferrari, A., Carlin, A., Rafele, C., Zenezini, G. (2024). A method for developing and validating simulation models for automated storage and retrieval system digital twins. The International Journal of Advanced Manufacturing Technology, 131(11), 5369-5382.

Fleiss, J.L., (1971), *Measuring nominal scale agreement among many raters*. Psychological Bulletin. 76(5), 378–382. doi:10.1037/h0031619.

Fleiss, J.L., (1981), *Statistical methods for rates and proportions*. J. Wiley & Sons, New York.

Franceschini, F., Maisano, D., Mastrogiacomo, L. (2016). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of informetrics*, *10*(4), 933-953.

Franceschini, F., Maisano, D. (2018). *Classification of objects into quality categories in the presence of hierarchical decision-making agents*. Accreditation and Quality Assurance, 23(1): 5-17.

Franceschini F., Galetto M., Genta G., Maisano D., (2018), *Selection of Quality-Inspection Procedures for short-run Productions*. International Journal of Advanced Manufacturing Technology, 99(9-12): 2537-2547. doi: 10.1007/s00170-018-2648-8.

Franceschini, F., Maisano, D. (2019). *Design decisions: concordance of designers and effects of the Arrow's theorem on the collective preference ranking*. Research in Engineering Design, 30(3), 425-434.

Franceschini, F., Maisano, D. (2021). *Aggregating multiple ordinal rankings in engineering design: the best model according to the Kendall's coefficient of concordance*. Research in Engineering Design, 32(1), 91-103.

Franceschini, F., Maisano, D., Mastrogiacomo, L. (2022) *Rankings and Decisions in Engineering: Conceptual and Practical Insights*. International Series in Operations Research & Management Science Series, Vol. 319, Springer International Publishing, Cham (Switzerland), ISSN: 0884-8289.

Franceschini, F., Maisano, D.A., Mastrogiacomo, L. (2023) *The ranking-aggregation problem in manufacturing: potential, pitfalls, and good practices*, Materials Research Proceedings, Vol. 35, pp 276-285, 2023

Genta G., Galetto M., Franceschini F. (2018), *Product complexity and design of inspection strategies for assembly manufacturing processes*. International Journal of Production Research, 56(11)-4056-4066. doi:10.1080/00207543.2018.1430907.

Genta G., Galetto M., Franceschini F. (2020), *Inspection procedures in manufacturing processes: recent studies and research perspectives*. International Journal of Production Research. 58(15): 4767-4788. doi: 10.1080/00207543.2020.1766713.

Gwet, K.L., (2008), *Computing inter-rater reliability and its variance in the presence of high agreement*. British Journal of Mathematical and Statistical Psycology. 61, 29–48. doi:10.1348/ 000711006X126600.

Gwet, K.L., (2015), *Testing the difference of correlated agreement coefficients for statistical significance*. Educational and Psychological Measurement. 76(4), 609–637. doi:10.1177/ 0013164415596420.

Gwet K.L. (2014) *Handbook of Inter-Rater Reliability*. Fourth Edition, Advanced Analitics LLC, Gaithersburg. MD, USA, ISBN: 978-0-9708062-8-4.

Hati S., Das B.R. (2011) *Seam Pucker in Apparels: A Critical Review of Evaluation Methods*. Asian Journal of Textile, 1: 60-73.

ISO 2859-1:1999 (1999) *Sampling procedures for inspection by attributes - Part 1*. Genève, Switzerland.

Keist C.N. (2015). *Quality Control and Quality Assurance in the Apparel Industry*. Garment Manufacturing Technology, 405–26. doi:10.1016/B978-1-78242-232-7.00016-3.

Kelly, J.S., (1989), *The Ostrogorski Paradox*. Social Choice and Welfare. 6(1), 71–76. doi:jstor.org/stable/41060227.

Knight J.K., Myers E.A. (1991), *Phased inspections and their implementation*. - ACM SIGSOFT Software Engineering Notes, 16(3), 29-35.

Landis J.R., Koch G. (1977), *The measurement of observer agreement for categorical data*. Biometrics, 33, 159-174.

Mandroli S.S., Shrivastava A.K., Ding Y. (2006), *A survey of inspection strategy and sensor distribution studies in discrete-part manufacturing processes*, IIE Transactions, 38:4, 309-328, doi: 10.1080/07408170500327352.

Mazza C., Alvarez J.L. (2017). *Haute couture and prêt-à-porter: the popular press and the diffusion of management practices*. In The Aesthetic Turn in Management (pp. 157-178). Routledge.

Montgomery D.C. (2019), *Introduction to Statistical Quality Control*, 8th Edition, Wiley, New York.

Ngan H.Y.T., Pang G.K.H., Yung N.H.C. (2011), *Automated fabric defect detection—A review*, Image and Vision Computing, 29(7), 442-458.

Odakura M., Kometani Y., Koike M., Tooma M. (2009), *Advanced inspection technologies for nuclear power plants*, Hitachi Review, 58(2), 82-87.

Schilling E.G., Neubauer D.V. (2009), *Acceptance Sampling in Quality Control*, 2nd Edition, CRC Press, BocaRaton, FL, USA. ISBN-13: 978-1584889526,

Stephens K.S. (2001), *The Handbook of Applied Acceptance Sampling*: *Plans, Procedures & Principles*. ASQ Quality Press, Milwaukee, Wisconsin, USA. ISBN-13: 978-0873894753.

Tang, W., Hu, J., Zhang, H., Wu, P., He, H. (2015). Kappa coefficient: a popular measure of rater agreement. Shanghai archives of psychiatry, 27(1), 62.

Verna E., Galetto M., Genta G., Franceschini F. (2021), *Inspection planning by defect prediction models and inspection strategy maps*. Production Engineering, 15(6):897-915. doi: 10.1007/s11740-021-01067-x.

Yuen C.W.M., Wong W.K., Qian S.Q., Chan L.K., Fung E.H.K., (2009), *A hybrid model using genetic algorithm and neural network for classifying garment defects*, Expert Systems with Applications, 36(2), 2037-2047.

Declaration

Ethical Approval

The authors respect the Ethical Guidelines of the Journal.

Consent to Participate

Not applicable.

Consent to Publish

Not applicable.

Authors Contributions

The authors have provided an equal contribution to the drafting of the paper.

Competing Interests

The authors do not have conflict of interest.

Availability of data and materials

Not applicable.