anpet
brazilian association for transportation research and education
TRANSPORTES
ISSN: 2237-1346
transportes.anpet.org.br

Inter-evaluator consistency in road safety audits: Evidence from on-site and hybrid evaluations

Consistência entre avaliadores em auditorias de segurança viária: Evidências de avaliações em campo e híbridas

Rogério Lemos Ribeiro[✉,1], Giulliano Renato Molinero[2], Pedro Henrique Vicentini da Silva[3], Barbara Stolte Bezerra[4], and Michelle Andrade[5]

[1]Federal University of Uberlândia, Uberlândia, Minas Gerais, Brazil, rogerio.ribeiro@ufu.br, 0000-0002-6750-8293
[2]Federal University of Uberlândia, Uberlândia, Minas Gerais, Brazil, giulliano.molinero5@gmail.com, 0009-0004-2609-4663
[3]São Paulo State University (UNESP), Bauru, São Paulo, Brasil, pedro.vicentini@unesp.br, 0009-0000-1741-3483
[4]São Paulo State University (UNESP), Bauru, São Paulo, Brasil, barbara.bezerra@unesp.br, 0000-0002-8459-4664
[5]University of Brasília, Brasília, Distrito Federal, Brasil, maccivil@gmail.com, 0000-0002-7660-5853

Received: 12 February, 2026
Final revision: 1 June, 2026
Accepted: 14 June, 2026
Published: 21 August, 2026

Associate Editor:
Flávio José Craveiro Cunto
Federal University of Ceará, Brazil

Keywords:
Road Safety Audit,
Hybrid inspection,
Inter-evaluator variability,
Standardized checklist,
Audit objectivity.

Palavras-chave:
Auditoria de Segurança Viária,
Inspeção híbrida,
Variabilidade entre avaliadores,
Lista de verificação padronizada,
Objetividade da auditoria.

DOI: 10.58922/transportes.v34.e3223

ABSTRACT
Road Safety Audits are a key instrument in proactive risk management for transport infrastructure, though their effectiveness may be influenced by evaluator variability and subjective judgment. This study empirically assessed variability, consistency, and reliability across two inspection modes: conventional on-site inspections and hybrid inspections combining virtual assessment with targeted field verification. Six auditors examined a 10 km section of the BR-116 highway in São Paulo, Brazil, using a 126-item checklist. Statistical analysis, employing the Relative Dispersion Index (RDI) and the Intraclass Correlation Coefficient (ICC, two-way random effects, absolute agreement), indicated moderate reliability for on-site inspections (ICC = 0.584), good reliability for hybrid inspections (ICC = 0.824), and overall moderate reliability (ICC = 0.629). Hybrid inspections demonstrated greater consistency, with variability primarily linked to interpretive complexity of checklist items. These findings provide empirical support for hybrid approaches, fostering more reproducible, scalable, and transparent audit practices.

RESUMO
As Auditorias de Segurança Viária são instrumentos centrais no gerenciamento proativo de riscos em infraestrutura de transportes, embora sua efetividade possa ser afetada pela variabilidade entre avaliadores e pelo julgamento subjetivo. Este estudo avaliou empiricamente a variabilidade, consistência e confiabilidade em dois modos de inspeção: inspeções presenciais convencionais e inspeções híbridas que combinam avaliação virtual com verificação direcionada em campo. Seis auditores examinaram um trecho de 10 km da rodovia BR-116, em São Paulo, Brasil, utilizando uma lista de verificação com 126 itens. A análise estatística, baseada no Índice de Dispersão Relativa (RDI) e no Coeficiente de Correlação Intraclasse (ICC, efeitos aleatórios bidirecionais, concordância absoluta), indicou confiabilidade moderada para inspeções presenciais (ICC = 0,584), boa confiabilidade para inspeções híbridas (ICC = 0,824) e confiabilidade geral moderada (ICC = 0,629). As inspeções híbridas mostraram maior consistência, com variabilidade associada à complexidade interpretativa dos itens. Os resultados oferecem evidências que apoiam abordagens híbridas, promovendo práticas de auditoria mais reprodutíveis, escaláveis e transparentes.

1. Introduction

Road safety is a core component of transportation system performance because infrastructure risk shapes crash outcomes, network reliability, logistics efficiency, public health, and the social legitimacy of mobility systems. Design and operational deficiencies generate costs beyond immediate crashes, affecting productivity, equity, and long-term sustainability. Consequently, road safety governance aligns with several United Nations Sustainable Development Goals addressing health, resilience, and equitable transport access (STA, 2019). Within this framework, Road Safety Audits were introduced in Great Britain and Denmark in the early 1990s as preventive instruments to identify hazards before crashes occur and were later adopted across Europe, Australia, and North America (Elvik, Høye & Sørensen, 2009). Alongside Road Safety Inspections, these audits constitute a central mechanism for proactive infrastructure risk management.

Audits and inspections involve structured, independent evaluations of roadway projects or operating facilities, aiming to reduce exposure to foreseeable hazards. Evidence consistently associates their use with reductions in crash frequency and severity, particularly when applied early in the project life cycle (Elvik, Høye & Sørensen, 2009; Baklanova *et al.*, 2021). As programs matured, transport agencies incorporated digital technologies. High-resolution imagery, geographic information systems, and remote sensing increasingly supplement field visits, giving rise to hybrid inspection protocols that combine remote analysis with targeted on-site verification. While these approaches enhance network coverage and reduce costs, they heighten concerns about whether conclusions reflect stable diagnostic signals or idiosyncratic professional judgment, a critical consideration for regulation, investment, and public accountability.

Despite institutionalization, Road Safety Audits remain heterogeneous. In high-income regions such as the European Union, Road Safety Audits are embedded within regulatory frameworks and are legally mandated as part of infrastructure development. Directive 2008/96/EC (EU, 2008) requires Member States to ensure that road safety audits are systematically carried out across multiple project stages, supported by formal training, certification, and standardized procedures.

However, such regulatory integration is not universally observed. Many jurisdictions still rely on locally developed procedures and instruments, limiting comparability and hindering cumulative learning (Calderón Ramírez *et al.*, 2023). Low and middle-income countries (LMICs) show heterogeneous implementation of road safety audits and star rating systems, with many countries reporting only partial adoption rather than fully institutionalized requirements. While some countries indicate the existence of audits for new road infrastructure, the limited presence of systematic inspections for existing roads and partial design standards for vulnerable users reflect gaps in standardization and enforcement (Bliss & Breen, 2013; WHO, 2018).

In Brazil, audits for new road infrastructure are reported as required, indicating alignment with preventive safety practices; however, design standards for pedestrians and cyclists remain partial, and no systematic inspections or star ratings are reported for existing roads, highlighting inconsistencies in lifecycle safety management (WHO, 2018).

Structured professional development and certification address this variability. Evidence from Southern Europe links audit effectiveness to auditor preparation and certification schemes aligned with European directives (Vardaki, Bekiaris & Kanellaidis, 2018). Programs extend beyond geometric design to include human factors, safety culture, and Safe System principles, reinforcing institutional foundations. Training integrates classroom instruction, supervised field exercises, case-based learning, and periodic recertification (Vardaki, Bekiaris & Kanellaidis, 2018), fostering convergence in diagnostic practices and enabling accumulation of shared analytical experience (Jamroz *et al.*, 2014).

Simultaneously, transport authorities have invested in formal audit instruments. Audits are typically applied during planning and design, while inspections are conducted periodically to detect emerging hazards in operating facilities (Ribeiro & Andrade, 2021). Both rely on structured checklists guiding examination of roadway geometry, roadside features, traffic control devices, and vulnerable user facilities (Nodari & Lindau, 2001). These tools have been developed internationally (Hildebrand & Wilson, 1999) and adapted nationally, including applications in Brazil (Schopf, 2006; Ribeiro, 2019).

Nevertheless, subjectivity persists, as recommendations often rely on professional judgment rather than quantified criteria, allowing for divergent assessments across teams (Schopf, 2006). Cross-national comparisons document substantial variation in procedural requirements, qualification standards, and analytical depth (Ahmed, Che Puan & Ismail, 2013; Calderón Ramírez *et al.*, 2023; Ashish, 2023), while surveys report uneven access to guidance and divergent interpretations of audit principles (AlHamad *et al.*, 2023). Studies using street-level imagery and codified scoring show that operationalized indicators combined with systematic training achieve moderate to high inter-rater agreement, yet interpretive judgments remain prone to divergence (Queralt *et al.*, 2021; Saito *et al.*, 2022; Hanibuchi *et al.*, 2024).

Recent empirical studies have begun to quantify inter-evaluator reliability in structured audit instruments applied to street environments. For instance, Cleland *et al.* (2021) reported intraclass correlation coefficients (ICC) ranging from 0.550 to 0.968 across evaluated attributes, with overall scores achieving excellent agreement (ICC ≈ 0.92). These findings indicate that structured instruments combined with remote assessment technologies can substantially improve consistency between evaluators.

The literature converges on the necessity of training, certification, and harmonized frameworks to ensure audit quality (Jamroz *et al.*, 2014; Vardaki, Bekiaris & Kanellaidis, 2018). Yet, no globally standardized methodology exists, and fragmented guideline development limits cross-national learning (Calderón Ramírez *et al.*, 2023). When supported by coherent regulatory systems and standardized checklists, these mechanisms increase the likelihood that auditors reach comparable diagnoses under similar conditions, strengthening the credibility of safety governance.

Despite these advances, there is still no empirical evidence isolating the combined effects of inter-evaluator variability under different inspection conditions, particularly in hybrid contexts. This gap is consequential as agencies adopt virtual and remote methods without fully understanding how consistency may vary across inspection settings and its implications for regulatory decision-making.

Against this backdrop, the present study aims to empirically assess inter-evaluator variability, consistency, and reliability in on-site and hybrid road safety inspections along a single roadway section. The analysis compares results obtained under different inspection modes (on-site, hybrid, and combined) to evaluate the stability of professional judgments across inspection contexts.

While prior work with the same dataset focused on operational feasibility (Ribeiro & Andrade, 2021), this study emphasizes the reliability of professional judgments under different inspection modes. The analysis is based on audits conducted along a 10 km segment of a federal highway, representing a controlled rural highway environment rather than complex urban settings such as signalized intersections or dense downtown networks. Data are reexamined to quantify inter-evaluator variability and assess consistency across inspection modes.

This study contributes to road safety management by providing empirical evidence on hybrid audit reliability in highway environments, clarifying patterns of professional judgment under different inspection modes, and offering guidance for agencies integrating remote assessment into transparent, reproducible, and scalable Road Safety Audit systems.

2. Method

This study analyzed previously conducted Road Safety Audits and Inspections along a 10 km mountainous segment of the BR-116 dual carriageway (Rodovia Régis Bittencourt), São Paulo, Brazil, between kilometer 508+458 and kilometer 518+458 southbound. The section's geometric and operational complexity generates diverse features for assessment, making it suitable for evaluating inter-evaluator variability and audit objectivity. The location of the study segment, including its start and end points, is presented in Figure 1.

A ground-level view of the study segment, illustrating its geometric characteristics, traffic conditions, and surrounding roadside environment, is presented in Figure 2.

2.1. Evaluators and inspection procedures

Six independent evaluations were analyzed, performed by experienced road safety evaluators. For the purposes of this study, experienced evaluators are defined as professionals with formal training in relevant fields and practical experience in traffic safety engineering. To compose the evaluation groups, six specialists in road safety were selected. The multidisciplinary team comprised three civil engineers, one technologist in earthworks and pavement, one cartographic engineer, and one technologist in logistics. The specialists had approximately 12 years of experience in traffic safety engineering (mean = 11.8 years, median = 12.0 years, SD = 3.97 years).

Three assessments were fully on-site, while three employed a hybrid approach combining remote review of 360-degree video with targeted field verification. This dual-mode design allowed investigation of whether the observation environment affects consistency without altering audit methodology.

Figure 1 • Location of the study segment along the southbound carriageway of BR-116, São Paulo, Brazil, indicating the start (km 508+458) and end (km 518+458) points (adapted from Google, 2026).

Figure 2 • Ground-level view of the study segment, illustrating its geometric characteristics, traffic conditions, and surrounding roadside environment.

All six inspections were conducted within the same week to ensure stability of geometric and operational conditions along the study segment. This temporal proximity minimized potential confounding effects related to infrastructure modifications, traffic management changes, or maintenance activities. Although minor weather variations may have occurred, these were not considered sufficient to materially influence checklist-based evaluations.

Prior to the assessments, all evaluators underwent structured knowledge alignment. This session included presentation of the study segment and its main characteristics, conceptual alignment on road safety principles, including the forgiving roads approach, standardized interpretation of checklist items, and familiarization with the data collection procedures and tools. Each evaluator recorded responses individually using a dedicated electronic data collection application. Inspection planning aspects were also addressed, including definition of field procedures, stopping points, and operational conditions for both on-site and hybrid evaluations. Assessments were conducted independently, following the same predefined procedures. This session aimed to reduce discrepancies arising from differences in conceptual understanding (Ribeiro, 2019). The complete checklist structure, including item hierarchy and descriptions, was used throughout the evaluation process to ensure standardized application across evaluators.

Each inspection aimed to identify actual or potential safety hazards affecting all categories of road users, irrespective of regulatory compliance (Huvarinen *et al.*, 2017). For both inspection modes, evaluators recorded a single consolidated rating for each checklist item after completing all planned site visits or virtual reviews. On-site audits followed a single-stage protocol based on a 234-item checklist. Evaluators traversed the study segment with stops at three predetermined locations, repeating the inspection seven times during daytime conditions and once at night while maintaining the posted speed limit of 60 km/h. Item ratings were finalized after all eight passes had been completed.

Hybrid inspections were conducted in two sequential stages. Stage 1 consisted of a virtual review using 360-degree video and a 191-item checklist, while Stage 2 involved targeted field verification of the remaining 43 items. Items related to roadway alignment, cross-section, auxiliary lanes, intersections, traffic signs and pavement markings, roadside features, bridges and culverts, and general infrastructure were primarily assessed during the virtual review, as these elements can be observed using 360-degree video. In contrast, items requiring closer inspection and detailed assessment, particularly those associated with pavement condition, certain drainage elements, and localized features, were evaluated through targeted field inspections. Evaluators assessed the entire segment during the virtual phase, examining it under various conditions: three travel speeds (30, 50, and 60 km/h), both traffic lanes, and two camera heights (1.53 m and 2.95 m) (Ribeiro & Andrade, 2021). The varying visual perspectives, specifically those resulting from the different camera heights, are illustrated below in Figure 3 with corresponding photographs. This visual representation aims to enhance understanding of how viewpoint changes might affect evaluators' perception of the roadway environment. Final ratings for the hybrid inspections were assigned only after completion of both stages. All evaluations were carried out independently, without communication among evaluators.

Checklist items were rated on a five-point scale, where 1 indicated less than 20 percent presence, 2 represented 20 to 40 percent, 3 corresponded to 40 to 60 percent, 4 indicated 60 to 80 percent, and 5 represented 80 to 100 percent. This resulted in a 234 by 6 matrix of items and evaluators. Items that could not be evaluated in the field or had missing responses were excluded to avoid bias. Field data collection was conducted in January 2019.

The methodological framework adopted in this study is summarized in Figure 4. The process comprises the definition of the study segment, selection and prior alignment of evaluators, application of a standardized checklist, execution of on-site and hybrid inspections, and subsequent data preparation and statistical analysis. After constructing the rating matrix, missing and non-applicable items were excluded, resulting in a final analytical dataset of 126 items from an initial 234-item checklist.

Data were quantitative, paired, and derived from small samples (six evaluations per item). Although checklist ratings were obtained using a five-point Likert scale, which is formally ordinal, parametric analysis was adopted due to the documented robustness of parametric methods to moderate violations of interval-level assumptions in applied research contexts (Norman, 2010).

(a) 1.53 m **(b)** 2.95 m

Figure 3 • Photographs illustrating the two camera heights used during the evaluation. (a) 1.53 m; (b) 2.95 m.

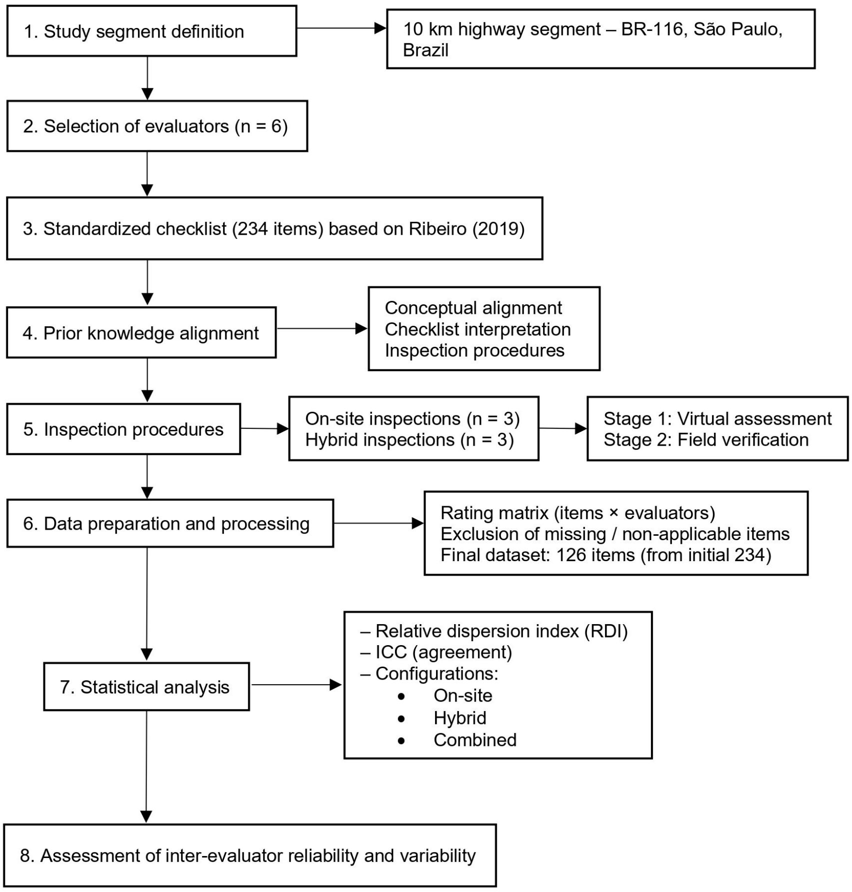

Figure 4 • Methodological framework for evaluating inter-evaluator reliability in on-site and hybrid road safety inspections

Inter-evaluator reliability was assessed using Intraclass Correlation Coefficients (ICC), adopting a two-way random-effects model with absolute agreement (ICC [2,1]). ICC values were calculated separately for the on-site audits, hybrid audits, and the combined dataset.

In addition, Relative Dispersion Indices (RDI) were calculated as descriptive measures of the relative dispersion of evaluator ratings for each checklist item. The RDI corresponds to the coefficient of variation, expressing the standard deviation as a proportion of the mean, according to Equation 1:

$$RDI = \frac{SD}{\text{mean}} \times 100, \tag{1}$$

where SD is the standard deviation of the ratings across evaluators and mean is the corresponding average rating. Higher RDI values indicate greater relative dispersion of scores, whereas lower values indicate more concentrated ratings.

The ICC was calculated according to the following formulation (**??**):

$$ICC(2,1) = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k(MS_C - MS_E)}{n}}, \tag{2}$$

where MS_R is the mean square for rows (subjects/items); MS_C is the mean square for columns (evaluators); MS_E is the residual mean square (error); k is the number of evaluators; and n is the number of items.

This combined analytical approach enabled the assessment of both overall inter-evaluator reliability (through ICC) and relative variability in ratings (through RDI), providing complementary perspectives on agreement patterns.

Analyses were conducted in three configurations: (*i*) on-site inspections, (*ii*) hybrid inspections, and (*iii*) all six inspections combined. Both ICC estimates and RDI values were calculated under each configuration. This structure enabled assessment of how inspection mode influences inter-evaluator variability and objectivity, directly addressing the study's objectives regarding checklist standardization and prior knowledge alignment. Additionally, checklist items were grouped into functional categories, and category-level agreement was assessed using mean relative dispersion indices (RDI) and Intraclass Correlation Coefficients (ICC), following the same analytical framework applied to the full dataset.

3. Results and conclusions

Preliminary examination of the dataset revealed both convergence patterns and localized discrepancies among evaluators. The complete response matrix was used to examine the distribution of evaluator ratings across the different inspection modes.

Overall inter-evaluator reliability was assessed using the Intraclass Correlation Coefficient (ICC). Results indicated moderate reliability for on-site inspections ($ICC = 0.584$; 95% CI: 0.488 to 0.671) and good reliability for hybrid inspections ($ICC = 0.824$; 95% CI: 0.772 to 0.867), according to established ICC interpretation guidelines (Koo & Li, 2016). All ICC estimates were statistically significant ($p < 0.001$), confirming that observed agreement levels were not attributable to chance.

Relative dispersion indices were moderate across all configurations, generally ranging between approximately 15% and 35%, indicating bounded variability within the Likert scale. These values should be interpreted as descriptive indicators of variability rather than measures of agreement, as they do not capture systematic differences among evaluators.

This combined interpretation of ICC and RDI allows for a comprehensive assessment of agreement patterns. While ICC captures the consistency of ratings across evaluators by accounting for both within-item and between-evaluator variance, RDI reflects the relative dispersion of scores around the mean. Together, these metrics provide complementary insights into inter-evaluator variability and audit objectivity.

3.1. Inter-evaluator variability

Analysis was conducted in three configurations: on-site inspections, hybrid inspections, and pooled analysis of all six evaluators.

Across all configurations, variability was not uniformly distributed among checklist items but followed a consistent structural pattern. Items requiring higher levels of interpretive judgment, particularly those related to operational conditions and complex geometric features, tended to exhibit greater dispersion and lower agreement. In contrast, items defined by explicit and observable criteria, such as traffic control devices and roadside features, consistently showed higher levels of agreement.

These results indicate that variability is primarily associated with the interpretive complexity of checklist items rather than the inspection modality itself.

3.1.1. On-site inspections

For on-site evaluations, dispersion levels, as measured by RDI values (approximately 27%), indicate moderate variability among evaluators. As a relative dispersion metric, RDI reflects the spread of ratings but does not capture systematic differences between evaluators.

The ICC (0.584) indicates moderate reliability for on-site audits. The combination of moderate ICC and non-negligible dispersion suggests that, although ratings are relatively clustered, consistent differences among evaluators persist. These differences are likely influenced by environmental and observational variability inherent to field conditions (Koo & Li, 2016).

Overall, the results indicate that on-site audits yield moderate reliability, with evaluator agreement influenced by contextual and operational conditions.

3.1.2. Hybrid inspections

For hybrid evaluations, RDI values (15.50%) indicate lower dispersion compared to on-site conditions, reflecting a higher degree of clustering in evaluator ratings.

The ICC (0.824) indicates good reliability, suggesting stronger agreement among evaluators. In this configuration, dispersion and reliability measures are consistent, indicating that reduced variability is accompanied by improved agreement.

This result is associated with the controlled observation environment provided by virtual assessments, where evaluators are exposed to standardized visual inputs. By reducing environmental and situational variability, hybrid inspections promote more consistent evaluations.

3.1.3. Pooled analysis (all six evaluators)

The purpose of this combined analysis is to evaluate the stability of evaluator agreement under heterogeneous inspection conditions. In practical inspection settings, different stages of assessment may involve both field observations and 360-degree video-based reviews or may be conducted by teams operating under varying conditions.

RDI values (28.41%) indicate moderate dispersion, reflecting the integration of different inspection contexts. The ICC (0.629) indicates moderate reliability, slightly lower than the hybrid configuration. This reduction is associated with the coexistence of heterogeneous observation environments rather than increased inconsistency within each group.

The results indicate that, even under heterogeneous conditions, evaluator agreement remains within acceptable levels. The combined analysis reflects realistic audit conditions in which different inspection modalities coexist, supporting the applicability of the proposed framework in practice.

Overall, the joint interpretation of RDI and ICC confirms that variability is primarily driven by item characteristics and observation conditions, rather than by methodological inconsistencies. To further investigate the structural sources of variability, checklist items were grouped into functional categories (Table 1).

Table 1 presents quantitative differences in inter-evaluator agreement across the four assessment categories. Geometric design items ($n = 19$) achieved the highest ICC value (0.656), followed closely by operational conditions ($n = 12$, $ICC = 0.649$), with both classifications falling within the moderate agreement range. Infrastructure items ($n = 51$) yielded an ICC of 0.582, while traffic control devices ($n = 44$) produced the lowest agreement among the four categories ($ICC = 0.527$), though all values remained within the moderate range. It is worth noting that the two categories comprising the largest item pools, namely traffic control devices and infrastructure, also exhibited the lowest agreement levels, suggesting that broader and more heterogeneous item sets may introduce greater evaluator divergence.

Mean Relative Dispersion Index (RDI) values ranged from 24.17% for traffic control devices to 32.92% for operational conditions, indicating measurable but bounded variability across the Likert scale. Despite operational conditions comprising a comparatively small item set ($n = 12$), its elevated mean RDI of 32.92% points to higher interpretive subjectivity within this category, a pattern not fully captured by the ICC alone. These indices should be interpreted strictly as descriptive measures of spread, given that RDI does not account for systematic directional differences among evaluators.

Taken together, the results presented in Table 1 suggest that variability is driven primarily by the nature of individual items rather than by inspection modality. Categories requiring greater interpretive judgment, such as operational conditions and infrastructure, tend toward higher dispersion and lower

Table 1 • Summary statistics of inter-evaluator variability and agreement (all evaluators combined)

Category	Number of items	Mean RDI (%)	$ICC(2,1)$	Interpretation
Geometric design	19	27.32	0.656	Moderate
Operational conditions	12	32.92	0.649	Moderate
Traffic control devices	44	24.17	0.527	Moderate
Infrastructure	51	31.41	0.582	Moderate

agreement, whereas items with more clearly defined and observable criteria allow for greater evaluator convergence. This finding is consistent with prior research (Queralt *et al.*, 2021; Saito *et al.*, 2022), which demonstrates that well-operationalized indicators paired with structured evaluator training are associated with higher inter-rater agreement, while items relying on contextual or inferential judgment remain inherently more sensitive to individual evaluator perspectives.

3.2. Comparison with previous studies

The results of this study are broadly consistent with prior research emphasizing the role of structured procedures and training in improving consistency in Road Safety Audits. Studies by Jamroz *et al.* (2014) and Vardaki, Bekiaris & Kanellaidis (2018) show that harmonized analytical frameworks and structured professional development are associated with more consistent audit outcomes. A similar pattern is observed in the present analysis, where relatively high levels of agreement were identified across both inspection modes under conditions involving standardized checklists and prior conceptual alignment among evaluators.

Recent literature further highlights the importance of standardized checklists as central instruments for guiding audit processes and reducing ambiguity in evaluation criteria. Calderón Ramírez *et al.* (2023) emphasize that such tools support consistency by structuring the interpretation of roadway conditions across different evaluators and contexts. The agreement patterns observed in this study are consistent with this interpretation, particularly in relation to items defined by explicit and observable criteria.

The higher reliability observed in hybrid audits aligns with findings from Queralt *et al.* (2021) and Saito *et al.* (2022), who report that image-based or virtual evaluations can achieve moderate to high inter-rater agreement when supported by clearly defined scoring systems and appropriate training. Similarly, Hanibuchi *et al.* (2024) indicate that virtual audit approaches can maintain reliability while improving efficiency and reducing operational constraints. The present findings are consistent with these studies, suggesting that controlled observation environments may be associated with increased evaluator agreement.

At the same time, the results indicate that variability persists in items requiring interpretive judgment, regardless of inspection modality. This observation is also consistent with previous research, which shows that even experienced evaluators may diverge in assessments when evaluation criteria are less explicitly defined (Queralt *et al.*, 2021; Saito *et al.*, 2022). This reinforces the importance of clearly operationalized indicators in supporting consistent evaluations.

This study extends the literature by examining on-site and hybrid inspections within the same road segment and under aligned procedural conditions. Unlike Ribeiro & Andrade (2021), which focus on the operational feasibility of hybrid audits, the present analysis allows a more controlled comparison of agreement patterns across inspection contexts. While the study design does not permit causal inference regarding the effects of checklist standardization or conceptual alignment, the observed consistency across configurations is compatible with the interpretation that these elements support more uniform evaluations.

Overall, the findings contribute to a more nuanced understanding of inter-evaluator variability by indicating that agreement levels are associated with both the structure of the audit process and the characteristics of the items being evaluated. In this sense, variability appears to be shaped less by the inspection modality itself and more by the degree of standardization and clarity embedded in the evaluation framework.

3.3. Implications for road safety audit practice

The empirical evidence indicates that inter-evaluator variability in Road Safety Audits can be substantially reduced by integrating three key components: experienced evaluators, pre-alignment of technical concepts, and standardized checklists. Hybrid procedures offer additional benefits by providing controlled observation conditions, which reduce environmental noise and support more uniform evaluations. These insights have practical implications for transportation agencies seeking to enhance the credibility and transparency of audit processes. Checklists should be viewed as dynamic instruments, subject to periodic revision based on field experience and emerging best practices. Structured training and alignment workshops should be integral to audit programs, reinforcing consistency without undermining professional judgment.

3.4. Limitations

Despite its contributions, this study has limitations that must be acknowledged. First, the analysis was confined to a single 10 km segment of BR-116, limiting generalizability to urban environments, road segments with different levels of complexity, or other highway environments. Second, the evaluator sample consisted of six experienced professionals; broader samples could capture greater variability across skill levels, training backgrounds, and professional perspectives. Third, hybrid inspections relied on pre-recorded 360-degree video under fixed conditions; dynamic traffic, weather, or temporary hazards were not captured. Fourth, several checklist items could not be evaluated due to operational constraints or missing data, slightly reducing the completeness of the analysis.

Nonetheless, the study directly addresses the research objectives: evaluating variability in on-site and hybrid inspections and assessing overall judgment consistency under controlled methodological conditions. Future research should expand empirical coverage to multiple road types, larger evaluator cohorts, and real-time hybrid protocols, allowing validation of these findings under more diverse operational conditions and examining their implications for crash risk reduction.

4. Conclusions

This study examined inter-evaluator variability and objectivity in Road Safety Audits and Inspections, considering both conventional on-site and hybrid inspection environments. The findings demonstrate that higher levels of consistency are achieved when audits are structured around standardized instruments and supported by prior conceptual alignment among evaluators.

Comparative analysis shows that both on-site and hybrid modalities can produce reliable assessments when embedded within a shared methodological framework. Intraclass Correlation Coefficient results indicate moderate reliability for on-site audits and good reliability for hybrid inspections, confirming the robustness of structured and harmonized audit protocols. The higher consistency observed in hybrid audits suggests that controlled observation conditions reduce contextual variability, positioning virtual tools as an effective complement, particularly under operational or logistical constraints.

The results challenge the assumption that subjectivity is inherent to Road Safety Audits. Instead, inconsistencies are primarily associated with methodological limitations, including the absence of explicit evaluation criteria, unstandardized instruments, and insufficient alignment among auditors. Addressing these factors enables professional judgment to operate within reproducible and defensible boundaries.

From a practical perspective, the study highlights the importance of continuous refinement of checklists and the systematic implementation of training and alignment procedures. These elements are essential for improving the reliability and transparency of safety evaluations. Although the findings are derived from a specific roadway segment, the methodological approach offers a replicable framework for broader application in transportation safety management.

Theoretically, the study advances the understanding of audit objectivity by demonstrating that it is a function of methodological and institutional design rather than solely of individual expertise. This perspective shifts the focus from evaluator subjectivity to the structure of the audit process, emphasizing the central role of clearly defined technical standards.

Applied implications indicate that checklists should be treated as dynamic instruments, subject to continuous updating based on field experience and evolving roadway conditions. Their systematic use supports more consistent evaluations without constraining professional judgment.

At the institutional level, pre-audit alignment of criteria and procedures emerges as a critical mechanism for reducing inter-evaluator variability. Structured training and ongoing professional development further reinforce consistency while preserving independent assessment.

This study has limitations. The analysis is restricted to a single roadway segment and a group of experienced evaluators, which may limit generalizability. Future research should expand the empirical scope to different road environments, include more diverse auditor profiles, test alternative checklist designs, and examine longitudinally how improvements in evaluator consistency translate into measurable safety outcomes.

Acknowledgements

The authors gratefully acknowledge the National Land Transport Agency (ANTT), Autopista Régis Bittencourt, and Arteris S.A. for the operational support provided for this study.

CRediT authorship contribution statement

Rogério Lemos Ribeiro: investigation, conceptualization, data curation, methodology, project administration, writing – original draft, writing – review & editing (lead); *Giulliano Renato Molinero*: conceptualization, data curation, methodology, writing – original draft; *Pedro Henrique Vicentini da Silva*: methodology, writing – original draft; *Barbara Stolte Bezerra*: conceptualization, writing – original draft, writing – review & editing, supervision. *Michelle Andrade*: conceptualization, supervision, writing – original draft.

Use of artificial intelligence-assisted technology

The authors acknowledge the use of an artificial intelligence–based language tool to assist with translation, linguistic refinement, and verification of citation-reference consistency. The scientific content, analysis, interpretations, and conclusions are entirely the authors' own. The tool was used solely to improve clarity and readability and did not generate original research ideas, data, or results.

Competing interests statement

No competing interests have been declared with regard to this paper or its contents.

Data availability statement

The data that support the findings of this study are available from the corresponding author (Rogério Lemos Ribeiro) upon request.

References

Ahmed, I., O. Che Puan, & C. R. Ismail (2013). A comparative review of road safety audit guidelines of selected countries. *Jurnal Teknologi 65*(3), 67–74. DOI:10.11113/jt.v65.2148.

AlHamad, S., M. Almallah, M. N. Naser, W. K. Alhajyaseen, & M. P. de Ross (2023). Examining the role of road safety audits worldwide: exploring road safety experts' opinions. *International Journal of Injury Control and Safety Promotion 30*(1), 106–115. DOI:10.1080/17457300.2022.2114090.

Ashish, S. Y. M. (2023). Road Safety Audit (RSA) guidelines of selected nations: a comparative review. *Journal of Road Safety 34*(4), 42–50. DOI:10.33492/JRS-D-23-209291.

Baklanova, K., E. Voevodin, E. Cheban, A. Askhabov, & A. Kashura (2021). Road safety audit as a tool for improving safety on the intercity road network. *Transportation Research Procedia 54*, 682–691. DOI:10.1016/j.trpro.2021.02.121.

Bliss, T. & J. Breen (2013). Road safety management capacity reviews and safe system projects guidelines (updated edition). Technical report, Global Road Safety Facility. URL: https://www.worldbank.org/en/topic/transport/publication/road-safety-management-capacity-review-guidelines [visited 06/14/2026].

Calderón Ramírez, J. A., I. Nuñez López, L. G. García Gómez, & M. A. Montoya Alcaraz (2023). Main guidelines in road safety audits: a literature review. *Frontiers in Built Environment 9*, 1–13. DOI:10.3389/fbuil.2023.1282251.

Cleland, C. L., S. Ferguson, F. Kee, P. H. J. Kelly, A. J. Williams, G. Nightingale, A. Cope, C. E. M. Foster, K. E. Milton, M. P. Kelly, R. G. Jepson, & R. F. Hunter (2021). Adaptation and testing of a microscale audit tool to assess liveability using Google Street View: MAPS-liveability. *Journal of Transport & Health 22*, 101226. DOI:10.1016/j.jth.2021.101226.

Elvik, R., T. V. Høye, & M. Sørensen (2009). *The Handbook of Road Safety Measures* (2 ed.). Bingley: Emerald Group Publishing.

EU (2008). *Directive 2008/96/EC of the European Parliament and of the Council of 19 November 2008 on road infrastructure safety management*. European Union. URL: http://data.europa.eu/eli/dir/2008/96/2019-12-16 [visited 06/14/2026].

Google (2026). *Google Earth Pro (Version 7.3)*. Google. URL: https://earth.google.com [visited 06/14/2026].

Hanibuchi, T., S. Nagata, D. Banis, H. Shobe, & T. Nakaya (2024). Urban streetscape changes in Portland, Oregon: a longitudinal virtual audit. *The Professional Geographer 76*(2), 180–193. DOI:10.1080/00330124.2023.2287166.

Hildebrand, E. & F. Wilson (1999). Road safety audit guidelines. Technical report, University of New Brunswick Transportation Group, Canadá.

Huvarinen, Y., E. Svatkova, E. Oleshchenko, & S. Pushchina (2017). Road safety audit. *Transportation Research Procedia 20*, 236–241. DOI:10.1016/j.trpro.2017.01.061.

Jamroz, K., M. Budzyński, W. Kustra, L. Michalski, & S. Gaca (2014). Tools for road infrastructure safety management – Polish experiences. *Transportation Research Procedia 3*, 730–739. DOI:10.1016/j.trpro.2014.10.052.

Koo, T. K. & M. Y. Li (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine 15*(2), 155–163. DOI:10.1016/j.jcm.2016.02.012.

Nodari, C. T. & L. A. Lindau (2001). Auditoria de segurança viária. *Transportes 9*(2), 48–66. DOI:10.14295/transportes.v9i2.170.

Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education 15*, 625–632. DOI:10.1007/s10459-010-9222-y.

Queralt, A., J. Molina-García, M. Terrón-Pérez, E. Cerin, A. Barnett, A. Timperio, J. Veitch, R. Reis, A. A. P. Silva, A. Ghekiere, D. Van Dyck, T. L. Conway, K. L. Cain, C. M. Geremia, & J. F. Sallis (2021). Reliability of streetscape audits comparing on-street and online observations: MAPS-Global in 5 countries. *International Journal of Health Geographics 20*(1), 6. DOI:10.1186/s12942-021-00261-5.

Ribeiro, R. L. (2019). *Procedimento para auditoria de segurança viária incluindo avaliações em ambientes virtuais (Publicação T.TD-006/2019)*. Brasília, DF: Universidade de Brasília.

Ribeiro, R. L. & M. Andrade (2021). Uso de ambientes virtuais para vistorias técnicas em auditoria e inspeção de segurança viária: um estudo de caso em trecho rural de rodovia duplicada. *Transportes 29*(1), 86–102. DOI:10.14295/transportes.v29i1.2216.

Saito, Y., Y. Oguma, S. Inoue, R. Breugelmans, H. Kikuchi, K. Oka, S. Okada, N. Takeda, K. L. Cain, & J. F. Sallis (2022). Inter-rater reliability of streetscape audits using online observations: MAPS Global in Japan. *Preventive Medicine Reports 30*, 102043. DOI:10.1016/j.pmedr.2022.102043.

Schopf, A. R. (2006). Proposição de uma lista de verificação para revisão de segurança viária de rodovias. MSc. thesis, Escola de Engenharia do Rio Grande do Sul, Porto Alegre. URL: http://hdl.handle.net/10183/7540 [visited 06/14/2026].

STA (2019). *Saving Lives Beyond 2020: The Next Steps*. Stockholm: Swedish Transport Administration.

Vardaki, S., E. Bekiaris, & G. Kanellaidis (2018). Addressing education and training needs of road safety auditors. *Cogent Engineering 5*(1), 1525814. DOI:10.1080/23311916.2018.1525814.

WHO (2018). *Global status report on road safety 2018*. Geneva: World Health Organization. URL: https://www.who.int/publications/i/item/9789241565684 [visited 06/14/2026].