frontiers
in Applied Mathematics
and Statistics

TYPE Original Research
PUBLISHED 11 March 2026
DOI 10.3389/fams.2025.1687194

OPEN ACCESS

EDITED BY
Eva Cantoni,
University of Geneva, Switzerland

REVIEWED BY
Pankaj Tiwari,
University of Kalyani, India
Karen Elliott,
University of Birmingham, United Kingdom

*CORRESPONDENCE
Muhammad Umair Danish
mdanish3@uwo.ca

RECEIVED 17 August 2025
REVISED 06 January 2026
ACCEPTED 04 February 2026
PUBLISHED 11 March 2026

CITATION
Umair Danish M, Rehman U and Grolinger K (2026) Monotone delta: an order-theoretic tournament graph approach for internal consistency assessment.
Front. Appl. Math. Stat. 12:1687194.
doi: 10.3389/fams.2025.1687194

Monotone delta: an order-theoretic tournament graph approach for internal consistency assessment

Muhammad Umair Danish[1]*, Umair Rehman[2] and Katarina Grolinger[1]

[1]Department of Electrical and Computer Engineering, Western University, London, ON, Canada, [2]Department of Computer Science, Western University, London, ON, Canada

This paper introduces Monotone Delta (δ), an order-theoretic measure for assessing the internal consistency of survey-based instruments. Classical coefficients such as Cronbach's Alpha and McDonald's Omega can yield misleading estimates under practical violations, including redundancy, multidimensional constructs, and correlated errors. Monotone Delta avoids parametric and factor-model assumptions by quantifying internal consistency through contradiction minimization with a weighted tournament formulation, aligning responses to an optimal unidimensional latent order. In controlled synthetic studies across four scenarios (tau-equivalence, redundancy, multidimensionality, and non-normal/correlated errors), Monotone Delta stays closest to the theoretical reliability, with absolute error $\leq$ 0.02 in the challenging scenarios where Alpha and Omega deviate by as much as 0.22 and 0.14, respectively. On a 350-participant human study for AI-generated image assessment, Monotone Delta agrees with Alpha/Omega under near-ideal conditions (overall $\delta = 0.91$ vs. $\alpha = 0.92$, $\omega = 0.94$) while remaining stable under redundancy and non-normal perturbations (overall $\delta = 0.84$ and $\delta = 0.81$, respectively, where Alpha drops to 0.95 and 0.35). These results position Monotone Delta as a practical alternative for reliability assessment in socio-technical systems, human factors, healthcare, and interactive system design.

KEYWORDS
human-AI interaction, internal consistency, order theory, reliability, socio-technical systems, surveys

1 Introduction

Reliability assessment is essential for evaluating survey-based instruments used in human-centered domains such as human-robot collaboration [1], healthcare [2], AI-generated content evaluation [3], and education [4]. These instruments are often deployed in situations where human responses inform the adaptive behavior of intelligent agents, decision-making systems, and simulation models. Reliability refers to the precision of scores across replications, while internal consistency is a key aspect of reliability. Internal consistency gauges how homogeneously the items tap a single latent construct, distinguishing it from broader reliability, which focuses on test-retest stability or parallel forms. This distinction is critical, as internal consistency ensures that collected responses

reflect the intended behavioral or cognitive factors, supporting the development of more accurate and stable social system models [5, 6].

Internal consistency is also fundamental in the computational modeling of user behavior and social perception, particularly in human-machine interaction and AI-assisted decision support [3, 7]. For example, survey-based evaluations that guide the development of user-centered systems, trust calibration models, or cognition-aware automation rely on the stability of item alignment within instruments. Poor reliability can propagate uncertainty in downstream modeling tasks, such as user simulation or behavior prediction. Robust reliability measures help mitigate this risk by ensuring coherent representation of latent dimensions, reducing the effects of redundancy or structural noise, and enabling the simulation of reliable and interpretable behavioral patterns in social systems [8, 9].

Existing reliability measures such as Cronbach's Alpha and McDonald's Omega are widely used to assess the internal consistency of survey-based instruments [10]. However, these methods are built on assumptions that usually fail in practical applications. For instance, Cronbach's Alpha assumes tau-equivalence, which requires all items to represent the latent construct equally, an assumption rarely met in real-world datasets [10]. It is also sensitive to redundancy, artificially inflating reliability when similar items are included [29, 30]. McDonald's Omega addresses some limitations by allowing for unequal item contributions, but still, it relies heavily on factor models sensitive to small sample sizes and uneven data distributions [5]. Both measures assume uncorrelated errors (local independence), though they do not require strict normality, which is frequently violated in multidimensional or heterogeneous datasets [11]. Addressing these gaps requires a fundamentally different technique that avoids reliance on parametric assumptions, adapts to diverse data distributions, and ensures stability across the varying nature of data. The need for an assumption-free, scalable, and robust reliability measure presents an opportunity to advance internal consistency assessment and improve reliability evaluation across disciplines.

This work is motivated by the growing mismatch between classical internal consistency measures and the structural properties of modern survey data used in human-centered and AI-mediated systems. Contemporary survey instruments increasingly capture context-dependent and heterogeneous human judgments, where item contributions are uneven, response patterns are non-Gaussian, and latent constructs may be only partially ordered rather than strictly linear. In such settings, variance-based and factor-analytic reliability measures can obscure meaningful response structure, conflate redundancy with coherence, and produce unstable or misleading reliability estimates. These limitations motivate the need for a reliability framework that directly reflects the ordinal and relational nature of survey responses, rather than relying on distributional or parametric assumptions.

To address the challenges of traditional reliability measures, this paper proposes Monotone Delta δ, an order-theoretic method designed to assess internal consistency by leveraging ordinal relationships among item responses. The core principle of Monotone Delta is to minimize ordinal contradictions in data by arranging responses along an optimal unidimensional latent order. This involves constructing a weighted tournament graph that captures pairwise dominance relationships between items and respondents. Monotone Delta identifies and resolves contradictions to quantify the alignment of responses with a coherent latent structure using the tournament graph technique. The proposed Monotone Delta operates without relying on assumptions such as tau-equivalence, normality, or factor models. Our method is fundamentally and theoretically different from Cronbach's Alpha and McDonald's Omega because it is based on order theory [12, 13], which makes it resilient against redundant items, multidimensional constructs, and distributional irregularities. This paper evaluates Monotone Delta through a controlled synthetic data and human-centered study on AI-generated image assessments. The proposed method remains adaptable to other domains utilizing survey-based instruments.

The novelty of this work lies in framing internal consistency assessment as an order-theoretic problem rather than a covariance-based. Unlike existing approaches that quantify reliability through variance decomposition or latent factor recovery, Monotone Delta evaluates internal consistency by measuring the degree to which survey responses admit a coherent latent ordering with minimal ordinal contradictions. By modeling pairwise response dominance relationships using a weighted tournament graph, the proposed method captures consistency as a structural property of response orderings rather than numerical agreement. This perspective introduces a fundamentally different notion of reliability that is assumption-free, robust to redundancy and multidimensionality, and aligned with the ordinal semantics inherent in many human-centered survey instruments. The main contributions of this paper are as follows:

1. Design of Monotone Delta, an order-theoretic measure quantifying internal consistency by minimizing ordinal contradictions, operating without parametric assumptions, and ensuring robustness against multidimensionality, redundancy, and data irregularities.
2. Design of a systematic evaluation to assess reliability measures under challenging conditions, including tau-equivalence, redundancy, multidimensionality, and non-normal distributions.
3. Theoretical evaluation of Monotone Delta, Cronbach's Alpha, and McDonald Omega, proving Monotone Delta's resilience to multidimensionality, redundancy inflation, and non-normality.
4. Experimental comparison of Monotone Delta with traditional measures, verifying its stable performance across diverse scenarios.

Before proceeding to the formal development of Monotone Delta, Table 1 provides a concise comparative overview of the proposed method against widely used reliability measures. The table highlights the core assumptions, structural sensitivities, and robustness properties of Cronbach's Alpha, McDonald's Omega, and Monotone Delta. This comparison serves as a summary, clarifying how classical variance- and factor-based measures respond to redundancy, multidimensionality, and non-normality,

and how the proposed order-theoretic formulation addresses these limitations.

The remainder of the paper is organized as follows: Section 2 presents the formal constructs and discusses limitations of existing measures, Section 3 details the proposed Monotone Delta, Section 4 presents the evaluation, and Section 5 concludes the paper.

2 Formal constructs and limitations of existing reliability measures

This section describes data and variable representation and provides theoretical evidence of challenges associated with Cronbach's Alpha and McDonald's Omega. This section also describes existing alternative methods and introduces Order Theory as a foundation for our work.

2.1 Data and variable representation

Let $R = \{r_1, r_2, \ldots, r_N\}$ represent the set of N respondents, where r_j denotes a single respondent, and let $I = \{i_1, i_2, \ldots, i_K\}$ be the set of K items or questions in the survey-based instrument. Each respondent r_j provides a vector of responses:

$$\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jK}) \in \mathbb{R}^K, \tag{1}$$

where $x_{j\ell}$ denotes the response of respondent r_j to item i_ℓ. We define the response vector across all respondents for a fixed item, denoted as $\mathbf{X}_\ell$, representing the set of all responses to an item i_ℓ from the N respondents, as:

$$\mathbf{X}_\ell = (x_{1\ell}, x_{2\ell}, \ldots, x_{N\ell}) \tag{2}$$

The total response score for each item i_ℓ, aggregating responses from all respondents, is given by:

$$X_\ell = \sum_{j=1}^{N} x_{j\ell}. \tag{3}$$

Here, $x_{j\ell}$ represents individual responses, $\mathbf{X}_\ell$ denotes the vector of responses for the item i_ℓ across all respondents, and X_ℓ is the aggregate score for the item i_ℓ.

2.2 Cronbach's Alpha

Cronbach's Alpha, denoted by α, is the most widely used method for measuring the internal consistency of a survey-based instrument [14]. It is defined as:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{\ell=1}^{k} \sigma^2_{X_\ell}}{\sigma^2_T}\right), \tag{4}$$

where k is items, $\sigma^2_{X_\ell}$ is variance of item i_ℓ, and σ^2_T is variance of total score $T = \sum_{\ell=1}^{k} X_\ell$. Cronbach's Alpha assumes tau-equivalence: items have equal true-score loadings but possibly different error variances [14]. This implies equal covariances between items but varying variances due to errors. Tau-equivalence rarely holds, leading to biased estimates. Alpha is also sensitive to item count: as k increases, $\alpha \to 1$, even with redundant items, as adding items reduces the error variance proportion. In multidimensional data, the covariance matrix becomes block-diagonal, violating unidimensionality and yielding misleading estimates. For example, a survey with separate factors for "usability" and "satisfaction" would show inflated alpha despite measuring distinct constructs. This assumption implies:

$$\mathrm{Cov}(X_\ell, X_m) = \sigma^2_T, \qquad \mathrm{Var}(X_\ell) = \sigma^2_T + \sigma^2_{\epsilon_\ell}, \qquad \forall \ell \neq m. \tag{5}$$

where $\mathrm{Cov}(X_\ell, X_m)$ is the covariance between items i_ℓ and i_m, and $\mathrm{Var}(X_\ell)$ is the variance of item i_ℓ. However, tau-equivalence rarely holds in practice, as items may differ in their measurement properties, leading to biased estimates. The second issue with Cronbach's Alpha is that it is sensitive to the number of items. As the number of items k increases, the value of α approaches one, even if the additional items are redundant or do not enhance the quality of the instrument. This behavior can be described as follows:

$$\alpha \to 1 \quad \text{as} \quad k \to \infty. \tag{6}$$

When the survey-based instrument captures multiple latent constructs [29, 30], the covariance matrix Σ of the item responses becomes block-diagonal:

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, \tag{7}$$

where Σ_1 and Σ_2 represent covariances within subsets of items measuring distinct constructs. This violates the assumption of unidimensionality, resulting in misleading reliability estimates. These limitations show that while Cronbach's Alpha is widely used, its assumptions and sensitivity to specific conditions restrict its effectiveness as a universal reliability measure.

2.3 McDonald Omega

McDonald's Omega, denoted as ω, is the second most used technique after Cronbach's Alpha [5, 15, 16]: it quantifies internal consistency by partitioning the total score variance into variance explained by a common latent factor and unique item variances. The total score T_j for respondent r_j is:

$$T_j = \sum_{\ell=1}^{K} x_{j\ell}, \tag{8}$$

where $x_{j\ell}$ represents the response of respondent r_j to item i_ℓ. The variance of the total score T_j is expressed as:

$$\sigma^2_T = \sum_{\ell=1}^{K} \lambda^2_\ell \sigma^2_F + \sum_{\ell=1}^{K} \sigma^2_{\epsilon_\ell}, \tag{9}$$

TABLE 1 Comparison of reliability measures.

Criterion	Cronbach's α	McDonald's ω	Monotone δ (Proposed)
Assumptions	Tau-equivalence required	Factor model assumptions	None
Handling multidimensionality	Produces misleading results	Moderately sensitive	Robust against violations
Sensitivity to Item redundancy	Inflates reliability scores	Overestimates reliability	Resilient to redundancy
Model dependence	No explicit model required	Relies on factor models	Independent of parametric models
Robustness to non-normality	Limited robustness	Susceptible to deviations	Fully robust
Computational complexity	Low	Moderate	Moderate

where λ_ℓ denotes the factor loading of item i_ℓ, σ_F^2 represents the variance of the common latent factor F_j, and $\sigma_{\epsilon_\ell}^2$ denotes the unique variance of item i_ℓ. McDonald's Omega is formally defined as:

$$\omega = \frac{\left(\sum_{\ell=1}^{K} \lambda_\ell\right)^2 \sigma_F^2}{\left(\sum_{\ell=1}^{K} \lambda_\ell\right)^2 \sigma_F^2 \; + \; \sum_{\ell=1}^{K} \sigma_{\epsilon_\ell}^2}. \tag{10}$$

This expression measures the proportion of total variance in the responses attributable to the common latent factor F_j. The common latent factor F_j represents the shared variance across all measurement instrument items and reflects the measured construct. The unique variances ($\sigma_{\epsilon_\ell}^2$) correspond to item-specific variability not explained by the common factor, and these are assumed to be uncorrelated across items:

$$\text{Cov}(\epsilon_{j\ell}, \epsilon_{jm}) = 0 \quad \text{for} \quad \ell \neq m. \tag{11}$$

The limitations of McDonald's Omega arise from specific factor model assumptions inherent in its computation, and the Assumption of Uncorrelated errors is often violated in practice. The overlapping content among items refers to items that assess highly similar aspects of a construct and can introduce error correlations, which leads to biased estimates of ω:

$$\text{Cov}(\epsilon_{j\ell}, \epsilon_{jm}) \neq 0 \quad \text{for} \quad \ell \neq m. \tag{12}$$

Weak factor loadings ($\lambda_\ell \approx 0$) reduce the contribution of items to the numerator:

$$\sum_{\ell=1}^{K} \lambda_\ell^2 \sigma_F^2, \tag{13}$$

This disproportionately inflates the denominator due to increased unique variance, resulting in an underestimated reliability. Redundancy among items inflates the total score variance σ_T^2 for items measuring identical constructs. The variances compound, resulting in poor reliability. Such inflation artificially raises ω, undermining its interpretive value. Moreover, in multidimensional datasets, items may correspond to distinct latent factors, leading to a block-diagonal covariance structure akin to the ω-specific form:

$$\Sigma = \begin{bmatrix} \Sigma_{\phi_1}\Theta_1 & 0 \\ 0 & \Sigma_{\phi_2}\Theta_2 \end{bmatrix}, \tag{14}$$

where Σ_{ϕ_1} and Σ_{ϕ_2} represent the factor covariance matrices for two latent dimensions, and Θ_1 and Θ_2 denote their respective residual variances. This structure violates the unidimensionality assumption, potentially rendering ω an inadequate measure of reliability. From the discussed challenges, summarized in Table 1, it is evident that both measures have limitations that limit their applicability across a wide range of applications. As the sophistication of questionnaire development continues to evolve, there is an urgent need for new measures to address the challenges both techniques face.

2.4 Alternative methods

In addition to Cronbach's Alpha and McDonald's Omega, other techniques, such as the Greatest Lower Bound (GLB) and Split-Half Reliability, have been proposed as alternative measures of internal consistency. The GLB provides a lower bound on reliability by optimizing the covariance matrix under the congeneric model, often exceeding alpha [17]. It does not assume unidimensionality but can be unstable and computationally intensive. Split-Half Reliability partitions items and correlates subset scores; it's independent of alpha/omega but sensitive to partitioning [18]. Both are useful benchmarks, but share limitations in handling redundancy and noise. Recent work has revisited the role of classical reliability coefficients, particularly Cronbach's Alpha, in light of long-standing critiques regarding their misuse and interpretability. Doval et al. [19] provide a contemporary reassessment of coefficient alpha and omega, emphasizing that neither measure should be treated as universally valid. Their analysis highlights that both coefficients can yield reasonable estimates only under restrictive conditions, such as approximate unidimensionality, near-normal item distributions, and weak error correlations. When these conditions are violated, which is common in practical survey data, reliability estimates become unstable or misleading. Importantly, this line of work reinforces the broader view that reliability assessment should be driven by the structural properties of the data rather than by default reliance on variance-based coefficients.

Other recent studies have focused on specific failure modes of Cronbach's Alpha that directly threaten its validity. Alkhadim and Alsharif [**?**] demonstrate that semantic overlap between survey items can substantially inflate alpha, even when the resulting reliability estimate is statistically indistinguishable from values obtained under random response patterns. Complementary evidence is provided by Hoekstra et al. [20], who empirically show that misunderstandings and misinterpretations of alpha

remain widespread among applied researchers. Together, these findings suggest that classical reliability coefficients often conflate conceptual redundancy, response artifacts, and genuine internal consistency. These limitations motivate the development of alternative reliability frameworks that do not rely on covariance structure or distributional assumptions and instead assess consistency through relational or structural properties of responses.

Split-Half Reliability [18] is a notable measure that partitions items into two subsets and evaluates the correlation between their scores. Despite its simplicity, the method is sensitive to how items are divided, leading to variability in reliability estimates. This measure also does not consider ordinal relationships, which is a considerable limitation in datasets with ties or noise. While alternative measures such as GLB and Split-Half Reliability present more options for assessing internal consistency, they share common limitations due to their theoretical reliance on Cronbach's Alpha and McDonald's Omega. These techniques extend or modify either Cronbach's Alpha and McDonald's Omega; for example, GLB refines Cronbach's Alpha through covariance matrix optimization, and Split-Half Reliability evaluates subset correlations and simplifies Omega by focusing on inter-item relationships. However, their shared assumptions, including unidimensionality and pairwise independence of items, limit their applicability to handle redundancy, noise, and multidimensionality. Given the widespread usage and theoretical prominence of Cronbach's Alpha and McDonald's Omega, they remain the most impactful benchmarks for comparison. We address these gaps by introducing a novel order-theoretic method that explicitly quantifies contradictions and incorporates robust handling of ties and noise, delivering a more reliable and scalable solution for modern datasets.

2.5 Order theory

We employ order theory as a foundation to overcome the limitations of traditional reliability measures. It provides a mathematical framework for analyzing hierarchical and sequential relationships, such as greater than, less than, and precedes [12, 13]. By formalizing these intuitive relationships through the lens of partial orders, this framework provides a robust mechanism for evaluating ordering and coherence within datasets. A partial order constitutes a binary relation $\preceq$ on a set P that adheres to three fundamental properties:

$$a \preceq a \quad \text{(reflexivity)}, \tag{15}$$

$$a \preceq b \text{ and } b \preceq a \implies a = b \quad \text{(antisymmetry)}, \tag{16}$$

$$a \preceq b \text{ and } b \preceq c \implies a \preceq c \quad \text{(transitivity)}. \tag{17}$$

In the field of measurement instruments, the response set R and items I establish a partially ordered set (poset) when their responses reflect an inherent order based on a latent trait. For example, higher scores typically signify a greater alignment with the measured construct. Order-preserving (monotone) functions are pivotal in evaluating the internal consistency of measurement instruments. A function $f : P \to Q$ is considered monotone if it satisfies the condition:

$$a \preceq b \implies f(a) \preceq f(b). \tag{18}$$

Monotonicity ensures the preservation of the latent ordering of responses under transformations, thereby facilitating meaningful interpretations of aggregated scores. Contradictions arise when observed responses deviate from the assumed latent ordering. For a poset P with relation $\preceq$, these contradictions become evident through pairs $(a, b) \in P \times P$ such that:

$$a \preceq b \quad \text{and} \quad b \prec a. \tag{19}$$

These violations disrupt the dataset's unidimensionality, complicating the interpretation of reliability measures. Addressing these contradictions is critical for deriving reliable internal consistency estimates. To solve challenges faced by both Cronbach's Alpha and McDoland Omega, we aim to employ the principles of order theory to quantify internal consistency by minimizing ordinal contradictions. The order theory can assess the alignment of item responses with a latent order, defined by a poset P in which items i_ℓ and responses $x_{j\ell}$ fulfill the requirement:

$$x_{j\ell} \preceq x_{jm} \implies i_\ell \preceq i_m. \tag{20}$$

This ordinal relationship enables practical evaluation of internal consistency across complex and heterogeneous questionnaires.

3 Monotone Delta

This section describes the proposed Monotone Delta, including the Theoretic Formulation of Monotone Delta, Monotone Delta Definition and Normalization, its properties, and theoretical examination.

3.1 Theoretic formulation of Monotone Delta

We use the notation for R, I, and $\mathbf{x}_j$ as defined in Section 2.1 and introduce additional symbols. Let π denote a permutation that orders respondents based on their responses. The function $W(j, k)$ counts the number of items where respondent r_j outperforms r_k, and it is used to construct weighted tournaments. A weighted tournament refers to a type of directed graph used in order theory to represent pairwise relationships among elements, such as respondents or items [21, 22]. The symbol $C(\pi)$ represents the total contradiction count for a given ordering π, quantifying deviations from the latent order.

The purpose is to evaluate how well the responses align with a single monotone latent dimension by minimizing contradictions. Consider a poset $(R, \preceq)$, where $\preceq$ represents a hypothesized latent order that reflects the unidimensional trait being measured. The

goal is to align the respondents $R = \{r_1, r_2, \ldots, r_N\}$ with this latent order. Let $\pi : \{1, \ldots, N\} \to \{1, \ldots, N\}$ denote a permutation that provides a linear extension of the poset, meaning the respondents are arranged such that:

$$r_{\pi(1)} \preceq r_{\pi(2)} \preceq \cdots \preceq r_{\pi(N)}. \tag{21}$$

The respondents' responses should respect this ordering if the data are perfectly unidimensional and free of noise. For any pair of respondents j and k where $\pi(j) < \pi(k)$, the responses for all items should satisfy:

$$\pi(j) < \pi(k) \implies x_{\pi(j)\ell} \leq x_{\pi(k)\ell}, \quad \forall \ell \in \{1, \ldots, K\}. \tag{22}$$

Here $x_{\pi(j)\ell}$ represents the response of the j-th respondent (according to the permutation π) to the ℓ-th item. The inequality $x_{\pi(j)\ell} \leq x_{\pi(k)\ell}$ implies that respondent $r_{\pi(j)}$ shows a response no stronger than respondent $r_{\pi(k)}$ for all items, consistent with the hypothesized latent order. Contradictions occur due to multidimensionality, noise, or redundant patterns. A contradiction is defined as a violation of Equation 22, i.e., there exists $j < k$ and an item ℓ such that:

$$x_{\pi(j)\ell} > x_{\pi(k)\ell}. \tag{23}$$

Understanding the core intuition behind Monotone Delta begins with viewing the dataset as a collection of ordinal comparisons between respondents. Each respondent's vector of responses can be seen as a profile whose relative ordering conveys latent information about the measured trait. The permutation π attempts to arrange respondents along a single latent continuum consistent with their observed responses. However, real-world data often contains noise, multidimensionality, or response patterns that conflict with any perfectly linear ordering. To capture these conflicts, Monotone Delta formalizes the notion of contradictions as instances where a respondent expected to rank lower actually scores higher on some items. The method transforms the reliability assessment into an optimization problem that seeks an ordering minimizing such inconsistencies, thereby aligning empirical data with the underlying theoretical latent trait as closely as possible.

The degree of contradiction measures how far the data deviates from a perfect unidimensional ordering. To quantify contradictions in respondent scores, we use the concept of a "weighted tournament," a directed graph where vertices correspond to respondents, and directed edges indicate dominant relationships based on their responses. The edge weights quantify in how many items one respondent outperforms another, and this computes the analysis of pairwise contradictions and the optimization of respondent orderings. The function $W(j, k)$ is defined as:

$$W(j, k) = \#\{\ell : x_{j\ell} > x_{k\ell}\}, \tag{24}$$

where $W(j, k)$ represents the number of items (ℓ) for which respondent r_j scores higher than respondent r_k. The symbol $\#\{\ldots\}$ denotes the cardinality of the set (i.e., the count of elements in the set). For example, if respondent r_j scores higher than r_k on 3 out of 5 items, then $W(j, k) = 3$. This structure induces a *weighted tournament* on N vertices, with directed edges weighted by $W(j, k)$.

To analyze contradictions, we consider a linear extension of the poset, a specific ordering π of respondents that respects the poset's partial order as much as possible. A linear extension arranges respondents $r_1, \ldots, r_N$ in a total order, such that if $r_j \preceq r_k$ in the poset, then r_j appears before r_k in π. However, due to noise or multidimensionality, responses may not perfectly align with the poset's partial order, leading to contradictions. For a given ordering π, the total contradiction count is:

$$C(\pi) = \sum_{1 \leq j < k \leq N} \#\{\ell : x_{\pi(j)\ell} > x_{\pi(k)\ell}\}. \tag{25}$$

This equation counts the number of item-level violations of the ordering π. A contradiction occurs when $x_{\pi(j)\ell} > x_{\pi(k)\ell}$ despite $\pi(j) < \pi(k)$, indicating that respondent $r_{\pi(j)}$ unexpectedly outperforms $r_{\pi(k)}$ on some items. To find the optimal ordering, we iteratively refine π by evaluating pairwise swaps of respondents and accepting swaps that reduce the contradiction count $C(\pi)$. The process continues until $C(\pi)$ converges to its minimum value C^*.

$$C^* = \min_{\pi} C(\pi). \tag{26}$$

The optimal ordering π^*, obtained through this refinement, aligns responses as closely as possible to the hypothesized latent order. This method is equivalent to solving a minimum feedback arc set problem [23] on the weighted tournament defined by $W(j, k)$.

3.2 Monotone Delta definition and normalization

The maximum possible contradiction count, $C_{\max}$, occurs if for every pair (r_j, r_k) with $j < k$, the ordering π is reversed relative to their observed dominance. Each pair can contribute up to K contradictions, and there are $N(N-1)/2$ pairs, thus:

$$C_{\max} = K \cdot \frac{N(N-1)}{2}. \tag{27}$$

We define the Monotone Delta as:

$$\delta = 1 - \frac{C^*}{C_{\max}}. \tag{28}$$

The value $\delta = 1$ indicates perfect unidimensional coherence, while lower values of δ reflect weaker coherence due to increased contradictions. As the dataset complexity increases (e.g., multiple latent dimensions, correlated errors, redundant items), C^* increases, reducing δ and signaling weaker unidimensional coherence. The ties in $W(j, k)$ and noise ($\epsilon_{j\ell}$) are handled by ensuring they do not artificially inflate C^* by maintaining the reliability. Algorithm 1 describes all the computational steps of the proposed Monotone Delta (δ). The concept of a weighted tournament graph provides a powerful abstraction for analyzing pairwise relationships within the response data. Each respondent corresponds to a vertex, and directed edges encode dominance based on the number of items one respondent scores higher on relative to another. The weight of each edge, $W(j, k)$,

effectively summarizes the strength of dominance between pairs, transforming the multidimensional response matrix into a network structure amenable to combinatorial optimization. Minimizing contradictions then corresponds to finding a vertex ordering that minimizes the total backward edge weight, akin to solving the minimum feedback arc set problem.

3.3 Properties and theoretical results

Theorem 1 (Scale Invariance)**.** Consider any strictly increasing transformation $g_\ell : \mathbb{R} \to \mathbb{R}$ applied item wise, i.e., $x_{j\ell} \mapsto g_\ell(x_{j\ell})$. Then, the relative ordering among responses is preserved, implying that:

$$C(\pi), \quad C^*, \quad \text{and} \quad \delta \quad \text{are unaffected by} \quad g_\ell. \tag{29}$$

Proof: Since g_ℓ is strictly increasing, we have

$$x_{j\ell} > x_{k\ell} \iff g_\ell(x_{j\ell}) > g_\ell(x_{k\ell}). \tag{30}$$

No new contradictions can be introduced or removed by such transformation. The structural properties of the weighted tournament (and thus the minimal contradiction count) remain unchanged. Therefore, δ is unaffected by scale changes. Further discussion on scale invariance in ordinal methods can be found in Bowen and Masa [24].

This proof verifies that Monotone Delta remains robust to scale changes, unlike Cronbach's Alpha, which is sensitive to such transformations [29].

Theorem 2 (Sensitivity to multidimensionality)**.** Let there be $d > 1$ latent dimensions, each affecting a distinct subset of items $I = I_1 \cup I_2 \cup \cdots \cup I_d$. If these dimensions are sufficiently distinct, then for large N there exists $\beta(N, K, d) > 0$ such that

$$\mathbb{E}[C^*] \geq \beta(N, K, d), \tag{31}$$

and therefore,

$$\delta \leq 1 - \frac{\beta(N, K, d)}{C_{\max}}. \tag{32}$$

Proof: If items are truly governed by multiple dimensions, a single total ordering cannot perfectly satisfy all item-response relations. The resulting "dimension conflicts" impose a positive lower bound on the minimal contradiction count. Formally, one can decompose the weighted tournament into sub-tournaments driven by each dimension and show via the minimum feedback arc set approach that these independent structures force additional contradictions. This returns $\beta(N, K, d)$ as a lower bound on $\mathbb{E}[C^*]$.

This indicates that as multidimensional conflicts intensify, Monotone Delta decreases, detecting deviations from unidimensionality that Cronbach's Alpha or McDonald Omega fail to reveal.

Theorem 3 (Redundancy resistance)**.** If r redundant items identical (up to small perturbations ϵ) to an existing item subset are added, the minimal contradiction count remains stable:

$$C^*(N, K + r) \approx C^*(N, K). \tag{33}$$

Input: Response matrix $X \in \mathbb{R}^{N \times K}$, where $x_{j\ell}$ is the response of respondent r_j to item i_ℓ.
Output: Monotone Delta δ, a measure of internal consistency in $[0, 1]$.
Step 1: construct weighted tournament
Initialize a directed, weighted graph $G = (V, E)$ with vertices $V = \{r_1, \ldots, r_N\}$.
for $j = 1$ to N **do**
  for $k = 1$ to N with $k \neq j$ **do**
    Compute $W(j, k) = \#\{\ell \mid x_{j\ell} > x_{k\ell}\}$.
    Add a directed edge from r_j to r_k with weight $W(j, k)$ to G.
  end for
end for
Step 2: initial ordering
Compute the mean score for each respondent r_j: $\bar{x}_j = \frac{1}{K} \sum_{\ell=1}^{K} x_{j\ell}$.
Sort respondents according to $\bar{x}_j$ to obtain an initial permutation π.
Step 3: local search optimization
Set $C(\pi) = \sum_{1 \leq j < k \leq N} \#\{\ell : x_{\pi(j)\ell} > x_{\pi(k)\ell}\}$.
repeat
    Select a pair $(r_{\pi(p)}, r_{\pi(q)})$ at random, with $p < q$.
    Create a new permutation π' by swapping $r_{\pi(p)}$ and $r_{\pi(q)}$.
    Compute $C(\pi')$.
    if $C(\pi') < C(\pi)$ **then** accept $\pi' \leftarrow \pi$ and $C(\pi) \leftarrow C(\pi')$.
until no improving swap is found after several attempts.
Step 4: compute minimal contradiction count
After convergence, let π^* be the final permutation found and $C^* = C(\pi^*)$ be the minimal contradiction count obtained.
Step 5: calculate Monotone Delta
Compute the maximum possible contradiction count $C_{\max} = K \cdot \frac{N(N-1)}{2}$.
Compute $\delta = 1 - \frac{C^*}{C_{\max}}$.
return δ

Algorithm 1. Monotone Delta (δ).

Proof: Let the response vector for respondent r_j be $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jK})$. Redundant items are defined as:

$$\begin{aligned} x_{j(K+m)} &= x_{j\ell} + \epsilon_{j(K+m)}, \\ m &= 1, \ldots, r, \\ \ell &\in \{1, \ldots, K\}. \end{aligned} \tag{34}$$

where $\epsilon_{j(K+m)}$ represents small independent perturbations. The updated weight function $W'(j,k)$ is:

$$W'(j,k) = W(j,k) + \sum_{m=1}^{r} \mathbb{I}\big(x_{j(K+m)} > x_{k(K+m)}\big). \tag{35}$$

where $\mathbb{I}(\cdot)$ is the indicator function. For redundant items, assuming small $\epsilon_{j(K+m)}$, we have:

$$\text{If } x_{j\ell} > x_{k\ell}, \quad \text{then } x_{j(K+m)} > x_{k(K+m)}, \quad \forall m. \tag{36}$$

Thus, redundant items preserve the relative ordering between r_j and r_k, contributing no additional contradictions. The total contradiction count for an ordering π after adding redundant items is:

$$C'(N,K+r) = C(N,K) + \sum_{j<k} \sum_{m=1}^{r} \mathbb{I}\big(x_{\pi(j)(K+m)} > x_{\pi(k)(K+m)}\big). \tag{37}$$

Since

$$\mathbb{I}(x_{\pi(j)(K+m)} > x_{\pi(k)(K+m)}) = \mathbb{I}(x_{\pi(j)\ell} > x_{\pi(k)\ell}). \tag{38}$$

the contradictions remain unchanged:

$$C'(N,K+r) = C(N,K). \tag{39}$$

Thus, the minimal contradiction count satisfies:

$$C^*(N,K+r) = C^*(N,K). \tag{40}$$

For small ϵ, the perturbations introduced by redundant items are negligible, ensuring:

$$C^*(N,K+r) \approx C^*(N,K). \tag{41}$$

This proof verifies that Monotone Delta δ is inherently resilient to redundancy, unlike traditional measures such as Alpha or Omega, which inflate reliability scores when redundant items are added.

4 Experimental evaluation

This section describes synthetic Data, a human study, and an evaluation through four scenarios and computational complexity.

4.1 Synthetic data experiment

To demonstrate the proposed method's performance under controlled conditions, we generated *synthetic* data where we already know the True reliability because the data is synthetic, such as the baseline Cronbach's alpha would be near perfect reliability. Then we created four scenarios using manually controlled perturbations such as *Redundancy*, *Multidimensionality*, and *Non-Normal/Correlated Errors* and compared them with *Tau-Equivalence ideal condition*, which will help to understand how traditional reliability methods and the proposed method respond to different situations. Synthetic data allows us to precisely define a *ground truth* reliability, which is challenging in real-world datasets (which can compound multiple factors). We then compare our proposed *Monotone Delta* with Cronbach's Alpha, McDonald's Omega, Greatest Lower Bound, and Split-Half Reliability.

4.1.1 Ground truth and data generation

Since we have synthetically controlled data, it is easy to find the ground truth. I the single-factor setup, let F denote the latent factor with variance σ_F^2, and let each item i have a loading λ_i and an error variance $\sigma_{\epsilon_i}^2$. The theoretical reliability [25] is then given by:

$$\text{TrueReliability} = \frac{\left(\sum_{i=1}^{K} \lambda_i\right)^2 \sigma_F^2}{\left(\sum_{i=1}^{K} \lambda_i\right)^2 \sigma_F^2 + \sum_{i=1}^{K} \sigma_{\epsilon_i}^2}. \tag{42}$$

where $\sum_{i=1}^{K} \lambda_i^2 \sigma_F^2$ represents the total variance accounted for by the latent factor, and $\sum_{i=1}^{K} \sigma_{\epsilon_i}^2$ represents the cumulative item-specific error variance. This allowed us to systematically simulate different structural violations found in real-world data and establish a theoretical reliability value for each case using the definition from Equation 42.

The data were generated in matrix form $X \in \mathbb{R}^{N\times K}$, where $N = 250$ respondents and $K = 20$ items. Each item was generated to align with a latent trait, with controlled Gaussian noise and scaling to ensure a valid ordinal structure. We introduced specific perturbations for each scenario to simulate realistic failure modes observed in reliability estimation. These include item redundancy, multidimensionality, and non-normal or correlated error structures. The ground truth reliability in each case was computed using the known latent variables and their contribution to item scores; this is theoretically correct because data is created under controlled conditions to prove certain conditions specifically.

4.1.2 Results and discussion

Table 2 presents the estimated reliability values from five commonly used techniques, including Cronbach's Alpha, McDonald's Omega, Greatest Lower Bound, Split-Half Reliability, and the proposed Monotone Delta, across four key scenarios. In the tau-equivalence scenario, where the ideal assumptions of classical reliability estimation hold, all methods perform similarly well. The theoretical reliability is 0.95, and each method estimates a value close to this: Cronbach's Alpha and McDonald's Omega yield 0.94, while GLB slightly increases to 0.96. Split-half reliability gives a slightly lower estimate of 0.93. Monotone Delta returns 0.95, resulting in a mean absolute error (MAE)

TABLE 2 Comparison of reliability measures vs. the theoretical reliability across four synthetic scenarios.

Scenario	True Rel.	Cronbach's Alpha	McDonald's Omega	GLB	Split-Half	Monotone Delta
Tau-equivalence	0.95	0.94 (0.01)	0.94 (0.01)	0.96 (0.01)	0.93 (0.02)	0.95 (0.01)
Redundancy	0.88	0.97 (0.09)	0.92 (0.04)	1.00 (0.12)	0.81 (0.02)	0.87 (0.01)
Multidimensional	0.82	0.69 (0.13)	0.75 (0.07)	0.67 (0.15)	0.74 (0.08)	0.80 (0.02)
Non-Normal	0.85	0.63 (0.22)	0.71 (0.14)	0.68 (0.17)	0.62 (0.23)	0.83 (0.02)

Values in parentheses indicate absolute difference from the ground truth.

of 0.01, matching the best of the traditional methods under ideal conditions. This confirms that Monotone Delta preserves fidelity to the ground truth when classical assumptions are met.

In the redundancy scenario, we introduced highly collinear items with existing ones (with 0.98 Pearson correlation), simulating a common issue in poorly designed surveys. The theoretical reliability remains at 0.88, but Cronbach's Alpha increases to 0.97 with an MAE of 0.01, significantly overestimating reliability due to its sensitivity to item count and inter-item correlation. GLB reaches 1.00 with an MAE of 0.12, stressing its extreme inflation under redundancy. McDonald's Omega shows a moderate overestimation at 0.92 with an MAE of 0.04. Split-half reliability provides a better estimate at 0.81 with an MAE of 0.07, but deviates from the ground truth. Monotone Delta returns 0.87, maintaining the lowest MAE of 0.01 and confirming its robustness in mitigating redundancy-driven inflation.

In the multidimensionality scenario, items were generated from two latent traits that each influenced half of the instrument. This structure violates the assumption that all items measure a single latent construct. The theoretical ground truth reliability was 0.82, while Cronbach's Alpha dropped sharply to 0.69 with an MAE of 0.13, and GLB fell further to 0.67 with an MAE of 0.15. McDonald's Omega performs moderately better at 0.75 with MAE of 0.07, which might be acceptable for most surveys but still underestimates the theoretical reliability of 0.82. Split-half reliability also declines to 0.74. Monotone Delta estimates the reliability at 0.80 with an MAE of 0.02, only two points away from the ground truth, again showing good sensitivity to multidimensional inconsistency.

Next, the non-normal and correlated error scenario introduces heavy-tailed item distributions and shared error structures between item pairs manually to the same synthetic dataset. These violate parametric assumptions required by most traditional reliability measures. Here, the theoretical ground truth was 0.85, while Cronbach's Alpha dropped to 0.63, and Split-half reliability performed similarly at 0.62. GLB and McDonald's Omega also deteriorate, giving estimates of 0.68 and 0.71 with MAEs of 0.17 and 0.14, respectively, compared to the true reliability of 0.85. Monotone Delta produces a score of 0.83 with an MAE of 0.02, resulting in the lowest error of 0.02 across all methods.

These findings confirm that Monotone Delta remains aligned with theoretical reliability under ideal and challenging conditions. Due to its order-theoretic foundation, Monotone Delta consistently tracks the proper internal consistency, unlike traditional methods, which are prone to inflation or deflation depending on structural violations in the data. This supports the method's practical utility in evaluating modern survey instruments that are often non-normal, redundant, or multidimensional.

4.2 Human study

We have also extended experiments on a real-world human subject study named Visual Verity [26] with a sample size of 350 participants for AI-generated images. But in case of real-world human studies, calculating a definitive ground truth for internal consistency is inherently challenging due to the absence of observable latent constructs and the complex interactions between respondent behaviors and item semantics. Since human-rated responses are influenced by numerous uncontrolled factors such as individual interpretation, cultural bias, and cognitive variability, there is no objective formula to compute the true reliability. Therefore, evaluations of human data remain inherently empirical, relying on comparative consistency across methods rather than direct measurement against, but still including human studies for experiment-proven applicability of the proposed method in real-world surveys and questionnaires.

In this survey, which is published in our previous works such as Aziz et al. [3, 26]. The AI-generated image dataset was chosen for its direct impact on technology, society, and social systems. The study consists of 22 questions assessing four distinct constructs to evaluate the perceptual quality and experiential responses to AI-generated images from three commercial models and camera-captured images. We got ethics approval from the Non-Medical Research Ethics Board at Western University, Ontario, to ensure ethical compliance in participant recruitment and data collection. We recruited participants via an online platform, namely prolific [27], which is known for its diverse pool.

The dataset evaluates images generated by three commercial AI models—DALL·E 3, DALL·E 2, and Stable Diffusion—and camera-captured images. These models represent different strategies for image generation and provide a diverse range of outputs regarding photorealism, coherence, and quality. The questionnaire given in Table 3 assesses multiple dimensions of image evaluation: demographics, photorealism, image quality, and caption consistency. It uses a mix of Likert-scale, multiple-choice, and open-ended questions designed to gather comprehensive feedback from participants.

TABLE 3 Visual verity questionnaire.

Question ID	Question text	Scale
Demographic Questions (DQ)		
DQ1	What is your gender?	Multiple choice
DQ2	What is your age?	Open-ended
DQ3	What is your educational qualification?	Multiple choice
DQ4	Experience with AI or computer-generated images.	Likert (1–5)
DQ5	Frequency of viewing digital images/graphics.	Likert (1–5)
DQ6	Experience in graphic design or photography.	Yes/No
DQ7	What is your country of residence?	Open-ended
Photorealism Assessment (PR)		
PR1	The image looks like a photograph of a real scene.	Likert (1–5)
PR2	I can easily imagine seeing this image in the real world.	Likert (1–5)
PR3	The visual details in this image make it appear realistic.	Likert (1–5)
PR4	The textures in the image look natural and real.	Likert (1–5)
PR5	The lighting and shadows in the image contribute to its realism.	Likert (1–5)
Image Quality (IQ)		
IQ1	The image is clear and sharp.	Likert (1–5)
IQ2	The colors in the image are vibrant and lifelike.	Likert (1–5)
IQ3	I am satisfied with the overall quality of this image.	Likert (1–5)
IQ4	The image has no visible artifacts or distortions.	Likert (1–5)
IQ5	The resolution of the image meets my expectations.	Likert (1–5)
Caption Consistency (CC)		
CC1	The image perfectly aligns with the given caption.	Likert (1–5)
CC2	The elements in the image correspond to the described scene in the caption.	Likert (1–5)
CC3	If I were to describe this image with a caption, it would closely match the provided one.	Likert (1–5)
CC4	The image misses some details mentioned in the caption.	Likert (1–5)
CC5	I feel the image is a true representation of the given caption.	Likert (1–5)

The questionnaire is a reliable foundation for internal consistency experiments due to its diversity and complexity; for example, data was collected against four constructs, totaling 67 questions and allowing us to assess response alignment and coherence across multiple evaluation dimensions. We have presented results in Figure 1, and Table 4 shows the overall average results, which demonstrate that camera images are highly realistic, achieving the highest scores in photorealism and text-image alignment. Since the purpose of this paper is to assess the internal consistency of the questionnaire, this assessment will focus less on questionnaire results but will emphasize examining internal consistency.

4.3 Scenario 1: Tau-equivalence (near-ideal condition)

As shown in Figure 2, under near-ideal conditions Monotone Delta performs similarly to CronbachŠs Alpha and McDonaldŠs Omega across the evaluated datasets. To establish the validity of our method, Monotone Delta, we first evaluate its performance under ideal conditions and compare it with established baseline measures, Cronbach's Alpha and McDonald's Omega. This will give confidence that the proposed method performs nearly equal to established baseline measures under normal conditions. We also extend comparisons with other measures such as GLB and Split-Half Reliability. Table 5 summarizes the reliability scores across four datasets such as Camera, DALL·E2, DALL·E3, and Stable Diffusion, and their combined overall dataset. All reliability measures show strong internal consistency, with values close to 1 indicating high reliability and values closer to 0 reflecting weak internal consistency. For the Camera dataset, Cronbach's Alpha scored 0.89, indicating strong internal consistency. McDonald's Omega aligns closely with a score of 0.90, further validating the reliability of the dataset. Monotone Delta, our proposed method, scored 0.88, showing close agreement with Cronbach's Alpha and McDonald's Omega. This alignment with established measures establishes the validity of Monotone Delta under ideal conditions, as it performs similarly to these well-established measures, instilling confidence in its use for further evaluation under more complex scenarios.

For the DALL·E2 dataset, Cronbach's Alpha reaches a higher value of 0.94, explaining stronger internal consistency. McDonald's Omega closely follows, with a score of 0.95. Monotone Delta

also performs similarly in this scenario, achieving a score of 0.92. For the DALL·E3 dataset, Cronbach's Alpha scored 0.93, McDonald's Omega achieved 0.94, and Monotone Delta scored 0.91, reflecting consistent agreement between the three measures. For the stable diffusion and overall dataset, the proposed method performs similarly to the established baselines, which ensures we now perturb our dataset to create another scenario and determine whether Monotone Delta and other measures give stable results or not.

4.4 Scenario 2: inflation by redundant items

In this scenario, we manually apply redundancy to the datasets by adding new items that are highly similar to existing ones. The redundant items were generated as linear combinations of original items with a redundancy factor of 0.95, meaning the new items were almost identical to the originals, with a small amount of random noise added. This modification aimed to assess the resilience of reliability measures against inflation caused by redundant items, which artificially increase item correlations and often lead to inflated reliability scores.

Table 6 presents the reliability measures across datasets, such as Cronbach's Alpha, sensitive to the number of items and their correlations, which showed inflated scores across all datasets. For example, in the Stable Diffusion dataset, Cronbach's Alpha increased to 0.98, indicating an artificially high level of internal consistency. This result reflects the measure's susceptibility to redundancy, as adding redundant items leads to overestimating reliability. McDonald's Omega also displayed inflated scores, though to a slightly lesser extent compared to Cronbach's Alpha. In the Stable Diffusion dataset, McDonald's Omega reached 0.93, confirming that it, too, is influenced by redundant items, albeit less dramatically than Cronbach's Alpha.

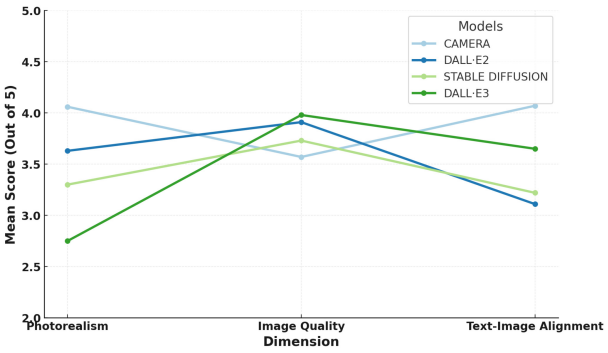

FIGURE 1
Comparison of participant responses across models (Camera, DALL·E2, Stable Diffusion, and DALL·E3) for three evaluation dimensions: Photorealism, Image Quality, and Text-Image Alignment. DALL·E3 and Stable Diffusion show contrasting trends in Image Quality, while Camera scores consistently high across dimensions.

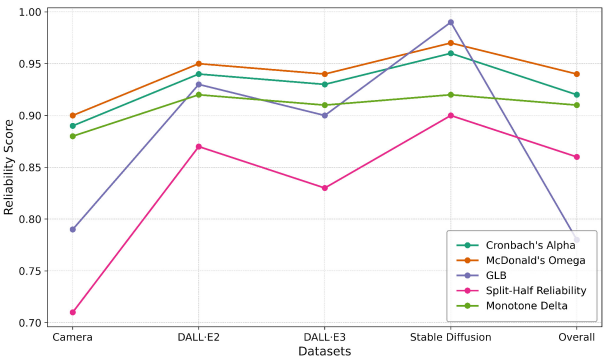

FIGURE 2
Readability scores under ideal conditions show that Monotone delta performs similarly to Cronbach's Alpha and McDonald's Omega.

TABLE 4 Mean participant responses (out of 5).

Dimension	Camera	DALL·E2	GLIDE	Stable diffusion	DALL·E3
Photorealism	4.06	3.63	2.04	3.30	2.75
Photorealism	4.06	3.63	2.04	3.30	2.75
Image quality	3.57	3.91	2.10	3.73	3.98
Photorealism	4.06	3.63	2.04	3.30	2.75
Text-image align.	4.07	3.11	2.03	3.22	3.65

TABLE 5 Reliability measures across datasets under ideal conditions.

Dataset	Cronbach's Alpha	McDonald's Omega	GLB	Split-half	Monotone Delta
Camera	0.89	0.90	0.79	0.71	0.88
DALL·E2	0.94	0.95	0.93	0.87	0.92
DALL·E3	0.93	0.94	0.90	0.83	0.91
Stable diffusion	0.96	0.97	1.01	0.90	0.92
Overall	0.92	0.94	0.78	0.86	0.91

TABLE 6 Reliability measures across datasets under redundancy scenario.

Dataset	Cronbach's Alpha	McDonald's Omega	GLB	Split-half	Monotone Delta
Camera	0.93	0.94	0.80	0.86	0.82
DALL·E2	0.96	0.97	0.93	0.93	0.83
DALL·E3	0.95	0.90	0.88	0.86	0.85
Stable diffusion	0.98	0.93	1.00	0.95	0.80
Overall	0.95	0.91	0.79	0.83	0.84

GLB also showed inflation under redundancy. For instance, in the Stable Diffusion dataset, GLB scored 1.00, surpassing all other measures and showing strong internal consistency, but it is misleading. Split-half reliability also performed poorly under redundancy, showing varying degrees of inflation. Split-half reliability in the DALL·E2 dataset increased to 0.93, reflecting the influence of redundant items on these measures. Monotone Delta, in contrast, showed resilience to redundancy across all datasets. In the Stable Diffusion dataset, it scored 0.80, closely aligning with its performance under ideal conditions and remaining unaffected by adding redundant items. The score in other datasets' similarity decreases, which shows Monotone Delta's resilience.

4.5 Scenario 3: multidimensionality

In this innovative scenario, we introduced multidimensionality into the datasets by splitting the items into two subsets, each influenced by a separate latent trait. This modification intentionally disrupted the unidimensional structure assumed by traditional reliability measures, introducing complexity that challenges their validity. By introducing multidimensionality, the items no longer measure a single cohesive construct, making it difficult for measures that rely on unidimensional assumptions to provide accurate reliability estimates. Table 7 summarizes the reliability scores across datasets, such as, Cronbach's Alpha, which assumes unidimensionality, showed a noticeable decline compared to its performance under ideal conditions. For instance, in the Camera dataset, Cronbach's Alpha dropped to 0.84, indicating a weaker internal consistency. This reduction stresses the measure's sensitivity to multidimensionality, as it conflates the distinct latent traits into a single reliability estimate.

McDonald's Omega, which accounts for varying item contributions but still relies on factor models, showed a slightly better performance than Cronbach's Alpha. In the DALL·E3 dataset, McDonald's Omega scored 0.89, reflecting moderate sensitivity to multidimensionality. GLB, which optimizes covariance matrices, also struggled with the multidimensional structure. For instance, in the Stable Diffusion dataset, GLB scored 0.78, confirming its inability to account for multiple latent traits fully. Split-half reliability performed poorly and reduced their scores across all datasets. Monotone Delta, however, showed resilience in the presence of multidimensionality. In the Camera dataset, Monotone Delta scored 0.75, verifying its ability to detect and quantify the impact of multidimensional constructs. Monotone Delta does not rely on assumptions of unidimensionality or factor structures. Instead, it minimizes ordinal contradictions, more accurately measuring the internal consistency.

4.6 Scenario 4: non-normal and correlated errors

In this scenario, we examined the robustness of reliability measures under conditions of non-normal distributions and correlated errors. Modifications deliberately violated the assumptions of normality and independent errors that many traditional measures rely on, providing a rigorous test of their effectiveness in handling real-world irregularities. Non-normal distributions caused the data to become uneven and stretched, leading to skewness (a shift in balance) and kurtosis (sharp peaks or flatness).

Table 8 presents the reliability scores for each dataset, such as Cronbach's Alpha, which assumes tau-equivalence and uncorrelated errors, declined across all datasets. For example, in the Camera dataset, Cronbach's Alpha dropped to 0.25, showing its inability to assess internal consistency under non-normal conditions accurately. This decline stresses its reliance on stringent assumptions often violated in real-world data. McDonald's Omega, which partially relaxes some of Cronbach's Alpha's assumptions, also showed reduced performance. In the DALL·E3 dataset, McDonald's Omega scored 0.48, reflecting moderate sensitivity to non-normality and correlated errors. However, its dependence on factor models limits its robustness in such scenarios, as these models struggle with non-linear and non-independent relationships. GLB, which optimizes covariance matrices, performed slightly better than Cronbach's Alpha and McDonald's Omega. Split-half reliability was proven to be the least reliable; for example, it scored 0.25 in the Camera dataset, showing its limitations in addressing the dependencies and non-linearity introduced by correlated errors.

Monotone Delta, on the other hand, showed superior robustness. In the Camera dataset, it scored 0.73, giving a stable measure. This performance stresses Monotone Delta's strength in capturing internal consistency without relying on assumptions of normality or independent errors. The results emphasize the limitations of traditional measures when confronted with non-normal distributions and correlated errors.

TABLE 7 Reliability measures across datasets under multidimensionality scenario.

Dataset	Cronbach's Alpha	McDonald's Omega	GLB	Split-half reliability	Monotone Delta
Camera	0.84	0.86	0.68	0.22	0.75
DALL·E2	0.85	0.88	0.72	0.37	0.77
DALL·E3	0.87	0.89	0.75	0.41	0.78
Stable diffusion	0.89	0.91	0.78	0.46	0.79
Overall	0.90	0.92	0.75	0.43	0.78

TABLE 8 Reliability measures across datasets under non-normal and correlated errors scenario.

Dataset	Cronbach's Alpha	McDonald's Omega	GLB	Split-Half	Monotone Delta
Camera	0.25	0.42	0.55	0.25	0.73
DALL·E2	0.28	0.45	0.57	0.28	0.75
DALL·E3	0.30	0.48	0.60	0.30	0.77
Stable diffusion	0.33	0.51	0.63	0.33	0.79
Overall	0.35	0.53	0.65	0.35	0.81

TABLE 9 Computation times for reliability measures (seconds).

Dataset	Cronbach's Alpha	McDonald's Omega	GLB	Split-Half	Monotone Delta
Camera	0.12	0.18	0.14	0.10	14.51
DALL·E2	0.14	0.20	0.15	0.12	15.24
DALL·E3	0.11	0.19	0.13	0.11	13.02
Stable diffusion	0.13	0.21	0.14	0.12	14.82
Overall	0.34	0.52	0.43	0.38	38.11

We also evaluated the computation times across all four scenarios, as summarized in Table 9. Computation times were measured using an AMD Ryzen Threadripper PRO 5955WX processor [28], ensuring test consistency and reliability. Monotone Delta consistently required more time than the other methods due to the iterative optimization process inherent to its computation.

5 Conclusion and future work

This paper proposed a Monotone Delta (δ) measure designed to address the limitations of traditional methods under diverse scenarios. Monotone Delta utilizes order theory to minimize ordinal contradictions and quantify reliability without relying on restrictive assumptions, and improves reliability assessment by addressing challenges such as redundancy, multidimensionality, and non-normality, presenting a reliable alternative to conventional measures. The theoretical and experimental evaluation was conducted across diverse scenarios and proved that Monotone Delta remains reliable and steady across diverse data, stressing its stability and accuracy in challenging conditions.

In contrast to classical reliability measures such as Cronbach's Alpha and McDonald's Omega, which estimate internal consistency through variance decomposition and latent factor assumptions, Monotone Delta evaluates reliability as a structural property of response orderings. Existing studies have shown that Alpha and Omega can produce inflated or deflated estimates under redundancy, multidimensionality, or correlated errors, which are common in modern survey-based instruments. Our experimental results confirm these findings and further demonstrate that Monotone Delta provides more stable and interpretable reliability estimates under the same conditions, without requiring assumptions such as tau-equivalence, unidimensionality, or error independence.

Compared with alternative approaches such as Greatest Lower Bound and Split-Half Reliability, which partially mitigate some weaknesses of Alpha and Omega but remain sensitive to data partitioning or covariance structure, Monotone Delta consistently aligns with theoretical reliability in controlled settings and exhibits robust behavior in real-world human studies. These comparisons position Monotone Delta as a complementary and, in many scenarios, more suitable reliability measure for contemporary survey data characterized by ordinal judgments, heterogeneous constructs, and structural noise.

Future work will focus on optimizing Monotone Delta's computational efficiency for larger datasets. It will also explore its possible integration with probabilistic and Bayesian frameworks to extend its applicability to larger datasets and further enhance its utility across diverse domains.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by Western University Non Medical Research Ethics Board (NMREB). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

MU: Software, Conceptualization, Writing – review & editing, Visualization, Methodology, Writing – original draft, Formal analysis, Data curation, Validation, Investigation. UR: Supervision, Project administration, Investigation, Conceptualization, Methodology, Writing – original draft, Visualization. KG: Investigation, Methodology, Software, Resources, Writing – review & editing, Project administration, Supervision, Validation, Funding acquisition.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the Climate Action and Awareness Fund (EDF-CA-2021i018, Environment and Climate Change Canada) and the Canada Research Chairs Program (CRC-2022-00078).

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. Grammarly's AI-based editor was used exclusively for proofreading and language polishing.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Hoffman G. Evaluating fluency in human-robot collaboration. *IEEE Trans Hum-Mach Syst*. (2019) 49:209–18. doi: 10.1109/THMS.2019.2904558

2. Lu Z, Zhou Y, Hu L, Zhu J, Liu S, Huang Q, et al. A wearable human-machine interactive instrument for controlling a wheelchair robotic arm system. *IEEE Trans Instrum Meas*. (2024) 73:4005315. doi: 10.1109/TIM.2024.3376685

3. Aziz M, Rehman U, Danish MU, Grolinger K. Global-local image perceptual score (GLIPS): evaluating photorealistic quality of AI-generated images. *IEEE Trans Hum-Mach Syst*. (2025) 55:223–33. doi: 10.36227/techrxiv.173933251.15884096/v1

4. Martín S, Lopez-Martin E, Moreno-Pulido A, Meier R, Castro M. The future of educational technologies for engineering education. *IEEE Trans Learn Technol*. (2021) 14:613–23. doi: 10.1109/TLT.2021.3120771

5. Hayes AF, Coutts JJ. Use omega rather than Cronbach's alpha for estimating reliability. But… *Commun Methods Meas*. (2020) 14:1–24. doi: 10.1080/19312458.2020.1718629

6. Ahmed I, Ishtiaq S. Reliability and validity: importance in medical research. *J Pak Med Assoc*. (2021). 71:2401–6. doi: 10.47391/JPMA.06-861

7. Mu J, Di Benedetto A. Networking capability and new product development. *IEEE Trans Eng Manag*. (2011) 59:4–19. doi: 10.1109/TEM.2011.2146256

8. Stadler M, Sailer M, Fischer F. Knowledge as a formative construct: a good alpha is not always better. *New Ideas Psychol*. (2021) 60:100832. doi: 10.1016/j.newideapsych.2020.100832

9. Moenaert RK, De Meyer A, Souder WE, Deschoolmeester D. R&D/marketing communication during the fuzzy front-end. *IEEE Trans Eng Manag*. (1995) 42:243–58. doi: 10.1109/17.403743

10. Barbera J, Naibert N, Komperda R, Pentecost TC. Clarity on Cronbach's alpha use. *J Chem Educ*. (2020) 98:257–58. doi: 10.1021/acs.jchemed.0c00183

11. Stensen K, Lydersen S. Internal consistency: from alpha to omega. *Tidsskr Nor Laegeforen*. (2022) 142:tidsskr.22.0112. doi: 10.4045/tidsskr.22.0112

12. Davey B. *Introduction to Lattices and Order*. Cambridge: Cambridge University Press. (2002). doi: 10.1017/CBO9780511809088

13. Chen X, Yu H, Hao F. Prescribed-time event-triggered bipartite consensus of multiagent systems. *IEEE Trans Cybern*. (2020) 52:2589–98. doi: 10.1109/TCYB.2020.3004572

14. Kennedy I. Sample size determination in test-retest and Cronbach alpha reliability estimates. *Br J Contemp Educ*. (2022) 2:17–29. doi: 10.52589/BJCE-FY266HK9

15. Orçan F. Comparison of Cronbach's alpha and McDonald's omega for ordinal data: are they different? *Int J Assess Tools Educ*. (2023) 10:709–22. doi: 10.21449/ijate.1271693

16. Cho E. Neither Cronbach's Alpha nor McDonald's Omega: a commentary on Sijtsma and Pfadt. *Psychometrika*. (2021) 86:877–86. doi: 10.1007/s11336-021-09801-1

17. Ten Berge JM, Sočan G. The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*. (2004) 69:613–25. doi: 10.1007/BF02289858

18. Chakrabartty SN. Best split-half and maximum reliability. *IOSR J Res Method Educ*. (2013). 3:1–8. doi: 10.9790/7388-0310108

19. Doval E, Viladrich C, Angulo-Brunet A. Coefficient alpha: the resistance of a classic. *Psicothema*. (2023) 35:5. doi: 10.7334/psicothema2022.321

20. Hoekstra R, Vugteveen J, Warrens M, Kruyen P. An empirical analysis of alleged misunderstandings of coefficient alpha. *Int J Soc Res Methodol*. (2019) 22:351–64. doi: 10.1080/13645579.2018.1547523

21. Connelly BL, Tihanyi L, Crook TR, Gangloff KA. Tournament theory: thirty years of contests and competitions. *J Manage*. (2014) 40:16–47. doi: 10.1177/0149206313498902

22. Rajkumar A, Veerathu V, Mir AB. A theory of tournament representations. *arXiv [preprint]*. (2021). doi: 10.48550/arXiv.2110.05188

23. Younger D. Minimum feedback arc sets for a directed graph. *IEEE Trans Circuit Theory*. (1963) 10:238–245. doi: 10.1109/TCT.1963.1082116

24. Bowen NK, Masa RD. Conducting measurement invariance tests with ordinal data: a guide for social work researchers. *J Soc Social Work Res*. (2015) 6:229–49. doi: 10.1086/681607

25. Raykov T. Estimation of composite reliability for congeneric measures. *Appl Psychol Meas*. (1997) 21:173–84. doi: 10.1177/01466216970212006

26. Aziz M, Rehman U, Danish MU, Ali S, Abbasi AZ. Towards a unified evaluation framework: integrating human perception and metrics for AI-generated images. *Multimed Syst*. (2025) 31:180. doi: 10.1007/s00530-025-01769-7

27. Albert DA, Smilek D. Comparing attentional disengagement between Prolific and MTurk samples. *Sci Rep*. (2023) 13:20574. doi: 10.1038/s41598-023-46048-5

28. Danish MU, Grolinger K. Leveraging hypernetworks and learnable Kernels for consumer energy forecasting across diverse consumer types. *IEEE Trans Power Deliv*. (2024) 40:75–87. doi: 10.31224/4361

29. Alkhadim GS. Cronbach's alpha and semantic overlap between items: A proposed correction and tests of significance. *Front. Psychol.* (2022) 13:815490. doi: 10.3389/fpsyg.2022.815490

30. Agbo AA. Cronbach's alpha: Review of limitations and associated recommendations. *J. Psychol. Afr.* (2010) 20:233–239. doi: 10.1080/14330237.2010.10820371