Leveraging Order-Theoretic Tournament Graphs for Assessing Internal Consistency in Survey-Based Instruments Across Diverse Scenarios

Muhammad Umair Danish, *Student Member, IEEE*, Umair Rehman, *Member, IEEE*, and Katarina Grolinger, *Member, IEEE*

***Abstract*—This paper introduces Monotone Delta (δ), an order-theoretic measure designed to enhance the reliability assessment of survey-based instruments in human-machine interactions. Traditional reliability measures, such as Cronbach's Alpha and McDonald's Omega, often yield misleading estimates due to their sensitivity to redundancy, multidimensional constructs, and assumptions of normality and uncorrelated errors. These limitations can compromise decision-making in human-centric evaluations, where survey instruments inform adaptive interfaces, cognitive workload assessments, and human-AI trust models. Monotone Delta addresses these issues by quantifying internal consistency through the minimization of ordinal contradictions and alignment with a unidimensional latent order using weighted tournaments. Unlike traditional approaches, it operates without parametric or model-based assumptions. We conducted theoretical analyses and experimental evaluations on four challenging scenarios: tau-equivalence, redundancy, multidimensionality, and non-normal distributions, and proved that Monotone Delta provides more stable reliability assessments compared to existing methods. The Monotone Delta is a valuable alternative for evaluating questionnaire-based assessments in psychology, human factors, healthcare, and interactive system design, enabling organizations to optimize survey instruments, reduce costly redundancies, and enhance confidence in human-system interactions.**

***Index Terms*—Reliability assessment, Internal consistency, survey-based instruments, Cronbach's Alpha, Non-parametric methods.**

I. 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]. The assessment ensures that elements within survey questionnaires are internally consistent and collectively measure the intended construct, affirming the truthfulness of the collected data [5], [6]. Internal consistency is a type of reliability assessment that measures how well items in a survey-based instrument contribute to evaluating the same latent construct. Thus, internal consistency is essential for assessing any measuring methodology for collecting primary data. It also enhances the validity of research findings and strengthens their credibility, making it essential for rigorous scientific inquiry [7], [8].

Internal consistency assessment is fundamental in human-centered systems to ensure the reliability and accuracy of survey-based instruments used in human-machine interaction, cognitive systems, and AI-assisted decision support [3], [9]. For instance, survey-based evaluations that inform strategies in user experience design, trust in automation, and AI-driven decision systems must reliably capture the constructs they intend to measure. This ensures that each item in a questionnaire meaningfully contributes to its intended construct, enhancing internal consistency and optimizing the instrument's dimensional structure [10]. Moreover, robust reliability measures mitigate the risks of flawed conclusions by addressing issues such as redundant items, correlated errors, or multidimensional constructs, which can obscure accurate reliability scores and lead to biased managerial decisions [11]. For practitioners, this translates into improved confidence in research findings and their applicability to real-world challenges [12].

Existing reliability measures such as Cronbach's Alpha and McDonald's Omega are widely used to assess the internal consistency of survey-based instruments [13]. 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 [13]. It is also sensitive to redundancy, artificially inflating reliability when similar items are included [3], [14]. 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 also depend on assumptions of normality and uncorrelated errors, which are frequently violated in multidimensional or heterogeneous datasets [15]. Addressing these gaps requires a fundamentally different technique that avoids reliance on parametric assumptions, adapts to diverse data distributions, and ensures stability across 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.

Muhammad Umair Danish and Katarina Grolinger are with the Department of Electrical and Computer Engineering, The University of Western Ontario, London, ON, Canada (e-mails: {mdanish3, kgroling}@uwo.ca).

Umair Rehman is with the Department of Computer Science, The University of Western Ontario, London, ON, Canada (e-mail: urehman6@uwo.ca).

Corresponding author: Katarina Grolinger (e-mail: kgroling@uwo.ca).

This work was supported by the Canada Research Chairs Program under Grant CRC-2022-00078 (K. Grolinger), NSERC Discovery Grant RGPIN-2018-06222 (K. Grolinger), SSHRC Insight Development Grant File No. 430-2024-01140 (U. Rehman), and NSERC Discovery Grant RGPIN-2024-05191 (U. Rehman). Computation was enabled in part by the Digital Research Alliance of Canada.

TABLE I
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

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 technique is fundamentally and theocratically different from Cronbach's Alpha and McDonald's Omega because it is based on order theory [16], [17], which makes it resilient against redundant items, multidimensional constructs, and distributional irregularities. This paper evaluates Monotone Delta through a human-centered study on AI-generated image assessments. The proposed method remains adaptable to other domains utilizing survey-based 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.

The remainder of the paper is organized as follows: Section II presents the formal constructs and discusses limitations of existing measures, Section III details the proposed Monotone Delta, Section IV presents the evaluation, and Section V concludes the paper.

II. 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.

A. 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.

B. 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 [18]. It is defined as:

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

where k represents the number of items, $\sigma_{X_\ell}^2$ is the variance of responses for item i_ℓ, and σ_T^2 is the variance of the total composite score T, which is defined as:

$$T = \sum_{\ell=1}^{k} X_\ell. \tag{5}$$

Cronbach's Alpha is an excellent measure, but it has several limitations, including its assumption of tau-equivalence and the requirement that all items have equal true-score variances. This assumption implies:

$$\text{Cov}(X_\ell, X_m) = \text{Var}(X_\ell), \quad \forall \ell, m \in \{1, 2, \ldots, k\}, \tag{6}$$

where $\text{Cov}(X_\ell, X_m)$ is the covariance between items i_ℓ and i_m, and $\text{Var}(X_\ell)$ 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:

$$\alpha \to 1 \quad \text{as} \quad k \to \infty. \tag{7}$$

When the survey-based instrument captures multiple latent constructs [10], the covariance matrix Σ of the item responses becomes block-diagonal:

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, \tag{8}$$

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.

C. McDonald Omega

McDonald's Omega, denoted as ω, is the second most used technique after Chronback Alpha [5], [19], [20]: 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:

$$T_j = \sum_{\ell=1}^{K} x_{j\ell}, \tag{9}$$

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:

$$\sigma_T^2 = \sum_{\ell=1}^{K} \lambda_\ell^2 \sigma_F^2 + \sum_{\ell=1}^{K} \sigma_{\epsilon_\ell}^2, \tag{10}$$

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:

$$\omega = \frac{\sum_{\ell=1}^{K} \lambda_\ell^2 \sigma_F^2}{\sigma_T^2}. \tag{11}$$

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, reflecting 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:

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

The limitations of McDonald's Omega arise from specific factor model assumptions inherent in its computation, and Uncorrelated errors are 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:

$$\text{Cov}(\epsilon_{j\ell}, \epsilon_{jm}) \neq 0 \quad \text{for} \quad \ell \neq m. \tag{13}$$

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

$$\sum_{\ell=1}^{K} \lambda_\ell^2 \sigma_F^2, \tag{14}$$

This disproportionately inflates the denominator due to increased unique variance, leading to 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:

$$\Sigma = \begin{bmatrix} \Sigma_{\phi_1}\Theta_1 & 0 \\ 0 & \Sigma_{\phi_2}\Theta_2 \end{bmatrix}, \tag{15}$$

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, which are also summarized in Table I, it is evident that both measures have limitations, which 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.

D. Alternative Methods

In addition to Cronbach's Alpha and McDonald's Omega, other techniques such as Greatest Lower Bound (GLB) and Split-Half Reliability have been proposed as an alternative measure of internal consistency. The GLB [21], [22] estimates reliability by optimizing the covariance matrix of items. The GLB usually outperforms Cronbach's Alpha, but it requires intensive computational resources and fails to address redundancy and multidimensional data. Moreover, its reliance on matrix optimization limits scalability to large datasets [21], [22]. Split-Half Reliability [23] 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.

E. 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 [16], [17]. 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:

$$a \preceq a \quad \text{(reflexivity)}, \tag{16}$$

$$a \preceq b \text{ and } b \preceq a \implies a = b \quad \text{(antisymmetry)}, \tag{17}$$

$$a \preceq b \text{ and } b \preceq c \implies a \preceq c \quad \text{(transitivity)}. \tag{18}$$

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:

$$a \preceq b \implies f(a) \preceq f(b). \tag{19}$$

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:

$$a \preceq b \quad \text{and} \quad b \prec a. \tag{20}$$

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 Chronback 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:

$$x_{j\ell} \preceq x_{jm} \implies i_\ell \preceq i_m. \tag{21}$$

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

III. 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.

A. Theoretic Formulation of Monotone Delta

We use the notation for R, I, and $\mathbf{x}_j$ as defined in Subsection II-A 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 [24], [25]. 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:

$$r_{\pi(1)} \preceq r_{\pi(2)} \preceq \cdots \preceq r_{\pi(N)}. \tag{22}$$

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:

$$\pi(j) < \pi(k) \implies x_{\pi(j)\ell} \leq x_{\pi(k)\ell}, \quad \forall \ell \in \{1, \ldots, K\}. \tag{23}$$

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 (23), i.e., there exists $j < k$ and an item ℓ such that:

$$x_{\pi(j)\ell} > x_{\pi(k)\ell}. \tag{24}$$

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:

$$W(j,k) = \#\{\ell : x_{j\ell} > x_{k\ell}\}, \tag{25}$$

where $W(j,k)$ represents the number of items (ℓ) for which respondent r_j scores higher than respondent r_k. The symbol $\#\{\dots\}$ 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 $r_1, \dots, r_N$ 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, the responses may not perfectly align with the poset's partial order, resulting in contradictions. For a given ordering π, the total contradiction count is:

$$C(\pi) = \sum_{1 \le j < k \le N} \#\{\ell : x_{\pi(j)\ell} > x_{\pi(k)\ell}\}. \tag{26}$$

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^*.

$$C^* = \min_{\pi} C(\pi). \tag{27}$$

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 [26] on the weighted tournament defined by $W(j,k)$.

B. 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:

$$C_{\max} = K \cdot \frac{N(N-1)}{2}. \tag{28}$$

We define the Monotone Delta as:

$$\delta = 1 - \frac{C^*}{C_{\max}}. \tag{29}$$

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).

C. 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:

$$C(\pi), \quad C^*, \quad \text{and} \quad \delta \quad \text{are unaffected by} \quad g_\ell. \tag{30}$$

Proof. Since g_ℓ is strictly increasing, we have

$$x_{j\ell} > x_{k\ell} \iff g_\ell\big(x_{j\ell}\big) > g_\ell\big(x_{k\ell}\big). \tag{31}$$

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 [27]. □

This proof verifies that Monotone Delta remains robust to scale changes, unlike Cronbach's Alpha, which is sensitive to such transformations [14].

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

$$\mathbb{E}\big[C^*\big] \;\ge\; \beta(N,K,d), \tag{32}$$

and therefore,

$$\delta \;\le\; 1 \;-\; \frac{\beta(N,K,d)}{C_{\max}}. \tag{33}$$

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:

$$C^*(N, K+r) \approx C^*(N, K). \tag{34}$$

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:

$$\begin{aligned} x_{j(K+m)} &= x_{j\ell} + \epsilon_{j(K+m)}, \\ m &= 1, \ldots, r, \\ \ell &\in \{1, \ldots, K\}. \end{aligned} \tag{35}$$

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

$$\begin{aligned} W'(j,k) = W(j,k) \\ + \sum_{m=1}^{r} \mathbb{I}\big(x_{j(K+m)} > x_{k(K+m)}\big). \end{aligned} \tag{36}$$

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

$$\begin{aligned} &\text{If } x_{j\ell} > x_{k\ell}, \\ &\text{then } x_{j(K+m)} > x_{k(K+m)}, \quad \forall m. \end{aligned} \tag{37}$$

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:

$$\begin{aligned} C'(N, K+r) = C(N, K) \\ + \sum_{j<k} \sum_{m=1}^{r} \mathbb{I}\big(x_{\pi(j)(K+m)} > x_{\pi(k)(K+m)}\big). \end{aligned} \tag{38}$$

Since

$$\begin{aligned} &\mathbb{I}(x_{\pi(j)(K+m)} > x_{\pi(k)(K+m)}) \\ &\qquad = \mathbb{I}(x_{\pi(j)\ell} > x_{\pi(k)\ell}). \end{aligned} \tag{39}$$

the contradictions remain unchanged:

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

Thus, the minimal contradiction count satisfies:

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

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

$$C^*(N, K+r) \approx C^*(N, K). \tag{42}$$

□

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.

Algorithm 1 Monotone Delta (δ)

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

IV. Experimental Evaluation

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

A. Human Study

We designed and conducted a human subject study named Visual Verity, with a sample size of 350 participants for AI-generated images. The AI-generated image dataset was chosen for its direct impact on managerial decisions in engineering management, including marketing, product design, and strategic innovation. 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

TABLE II
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)

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 [28], which is known for its diverse variety of 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 II) 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.

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 III shows overall average results that show 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.

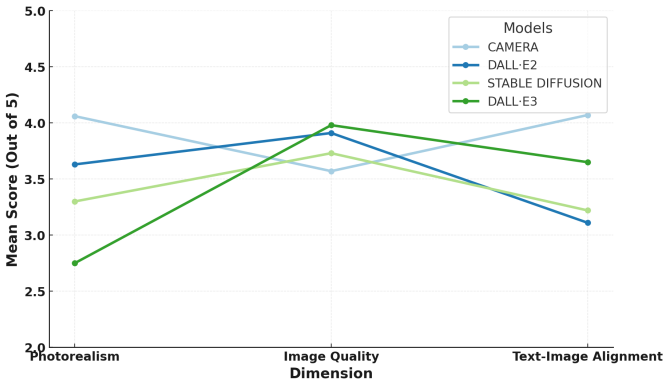

Fig. 1. Comarison of participant responses across models (Camera, DALL·E2, Stable Diffusion, and DALL·E3) for three evaluation dimensions: Photorealism, Image Quality, and Text-Image Alignment. DALL·E3 and Stable Diffusion show contrasting trends in Image Quality, while Camera scores consistently high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

TABLE III
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
Image Quality	3.57	3.91	2.10	3.73	3.98
Text-Image Align.	4.07	3.11	2.03	3.22	3.65

B. Scenario 1: Tau-Equivalence (Near-Ideal Condition)

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

TABLE IV
RELIABILITY MEASURES ACROSS DATASETS UNDER IDEAL CONDITIONS

Dataset	Cronbach's Alpha	McDonald's Omega	GLB	Split-Half Reliability	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

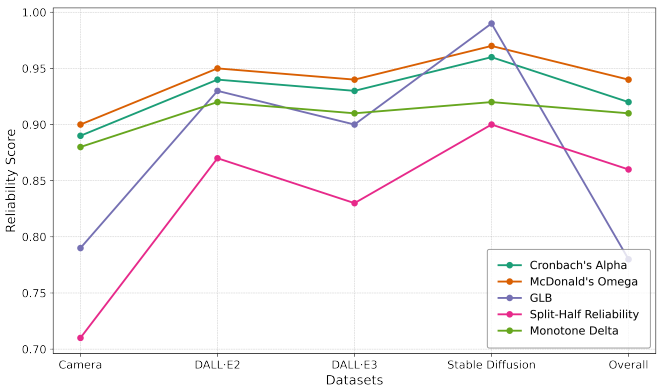

Fig. 2. Readability scores under ideal conditions show that Monotone Delta performs similarly to Chronback's Alpha and McDoland Omega.

comparisons with other measures such as GLB and Split-Half Reliability. Table IV 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.

C. 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 V 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.

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.

D. 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

TABLE V
RELIABILITY MEASURES ACROSS DATASETS UNDER REDUNDANCY SCENARIO

Dataset	Cronbach's Alpha	McDonald's Omega	GLB	Split-Half Reliability	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

TABLE VI
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 VII
RELIABILITY MEASURES ACROSS DATASETS UNDER NON-NORMAL AND CORRELATED ERRORS SCENARIO

Dataset	Cronbach's Alpha	McDonald's Omega	GLB	Split-Half Reliability	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

VI 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 fully account for multiple latent traits. 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.

E. 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 VII 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 accurately assess internal consistency under non-normal conditions. 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 proven least reliable, for example, 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 the stable measure. This performance stresses Monotone Delta's strength in capturing internal consistency without relying on assump-

TABLE VIII
COMPUTATION TIMES FOR RELIABILITY MEASURES (SECONDS)

Dataset	Cronbach's Alpha	McDonald's Omega	GLB	Split-Half Reliability	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

tions of normality or independent errors. The results emphasize the limitations of traditional measures when confronted with non-normal distributions and correlated errors.

We also evaluated the computation times for Cronbach's Alpha, McDonald's Omega, and Monotone Delta across all four scenarios, as summarized in Table VIII. Computation times were measured using an AMD Ryzen Threadripper PRO 5955WX processor [29], 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.

V. 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. The experiments also showed that Monotone Delta is computationally expensive and suitable for most human studies but can be prone to NP-Hardness for larger datasets, which require a careful optimization strategy. 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.

REFERENCES

[1] G. Hoffman, "Evaluating fluency in human–robot collaboration," *IEEE Transactions on Human-Machine Systems*, 2019.

[2] Z. Lu, Y. Zhou, L. Hu, J. Zhu, S. Liu, Q. Huang, and Y. Li, "A wearable human-machine interactive instrument for controlling a wheelchair robotic arm system," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[3] M. Aziz, U. Rehman, M. U. Danish, and K. Grolinger, "Global-local image perceptual score (glips): Evaluating photorealistic quality of ai-generated images," *IEEE Transactions on Human-Machine Systems*, 2025.

[4] S. Martín, E. Lopez-Martin, A. Moreno-Pulido, R. Meier, and M. Castro, "The future of educational technologies for engineering education," *IEEE Transactions on Learning Technologies*, 2021.

[5] A. F. Hayes and J. J. Coutts, "Use omega rather than cronbach's alpha for estimating reliability. but…," *Communication Methods and Measures*, 2020.

[6] I. Ahmed and S. Ishtiaq, "Reliability and validity: Importance in medical research." *Methods*, 2021.

[7] M. Tavakol and R. Dennick, "Making sense of cronbach's alpha," *International journal of medical education*, vol. 2, p. 53, 2011.

[8] S. P. Y. Karakaya and Z. N. Alparslan, "Sample size in reliability studies: A practical guide based on cronbach's alpha," *Psychiatry and Behavioral Sciences*, 2022.

[9] J. Mu and A. Di Benedetto, "Networking capability and new product development," *IEEE Transactions on Engineering Management*, 2011.

[10] C. G. Forero, "Cronbach's alpha," in *Encyclopedia of quality of life and well-being research*. Springer, 2024.

[11] M. Stadler, M. Sailer, and F. Fischer, "Knowledge as a formative construct: A good alpha is not always better," *New Ideas in Psychology*, 2021.

[12] R. K. Moenaert, A. De Meyer, W. E. Souder, and D. Deschoolmeester, "R&d/marketing communication during the fuzzy front-end," *IEEE transactions on Engineering Management*, 1995.

[13] J. Barbera, N. Naibert, R. Komperda, and T. C. Pentecost, "Clarity on cronbach's alpha use," *Journal of Chemical Education*, 2020.

[14] A. A. Agbo, "Cronbach's alpha: Review of limitations and associated recommendations," *Journal of Psychology in Africa*, vol. 20, no. 2, pp. 233–239, 2010.

[15] K. Stensen and S. Lydersen, "Internal consistency: from alpha to omega," *Tidsskrift for den Norske Laegeforening: Tidsskrift for Praktisk Medicin, ny Raekke*, vol. 142, no. 12, 2022.

[16] B. Davey, *Introduction to Lattices and Order*. Cambridge University Press, 2002.

[17] X. Chen, H. Yu, and F. Hao, "Prescribed-time event-triggered bipartite consensus of multiagent systems," *IEEE Transactions on Cybernetics*, 2020.

[18] I. Kennedy, "Sample size determination in test-retest and cronbach alpha reliability estimates," *British Journal of Contemporary Education*, 2022.

[19] F. Orçan, "Comparison of cronbach's alpha and mcdonald's omega for ordinal data: Are they different?" *International Journal of Assessment Tools in Education*, vol. 10, no. 4, pp. 709–722, 2023.

[20] E. Cho, "Neither cronbach's alpha nor mcdonald's omega: A commentary on sijtsma and pfadt," *Psychometrika*, 2021.

[21] J. M. Ten Berge and G. Sočan, "The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality," *Psychometrika*, 2004.

[22] E. Cho, "Reliability and omega hierarchical in multidimensional data: A comparison of various estimators." *Psychological Methods*, 2022.

[23] S. N. Chakrabartty, "Best split-half and maximum reliability," *IOSR Journal of Research and Method in Education*, 2013.

[24] B. L. Connelly, L. Tihanyi, T. R. Crook, and K. A. Gangloff, "Tournament theory: Thirty years of contests and competitions," *Journal of management*, 2014.

[25] A. Rajkumar, V. Veerathu, and A. B. Mir, "A theory of tournament representations," *arXiv preprint arXiv:2110.05188*, 2021.

[26] D. Younger, "Minimum feedback arc sets for a directed graph," *IEEE Transactions on Circuit Theory*, 1963.

[27] N. K. Bowen and R. D. Masa, "Conducting measurement invariance tests with ordinal data: A guide for social work researchers," *Journal of the Society for Social Work and Research*, 2015.

[28] D. A. Albert and D. Smilek, "Comparing attentional disengagement between prolific and mturk samples," *Scientific Reports*, 2023.

[29] M. U. Danish and K. Grolinger, "Leveraging hypernetworks and learnable kernels for consumer energy forecasting across diverse consumer types," *IEEE Transactions on Power Delivery*, 2024.