TPM Vol. 32, No.1, 2025
ISSN: 1972-6325
https://www.tpmap.org/

CRONBACH'S ALPHA COEFFICIENT BETWEEN CLASSICAL TEST THEORY (CTT) AND ITEM RESPONSE THEORY (IRT): A CRITICAL REVIEW

BAKR HUSSEIN
AL- IRAQIA UNIVERSITY

Abstract: This article provides a comprehensive critical analysis of Cronbach's alpha from a theoretical and conceptual perspective. It traces the historical development of the coefficient from its roots in Kuder and Richardson's KR-20 equations to Cronbach's (1951) generalization, which made it applicable to multilevel data. It discusses prerequisites for its proper use, such as unidimensionality and error independence, and highlights common interpretation errors, such as considering it as evidence of validity rather than reliability. It also presents criticisms of it from an Item Response Theory (IRT) perspective, which reveal its limitations due to its sample dependence and assumption of consistent accuracy across all subjects. Conversely, the article reviews more accurate contemporary alternatives, such as McDonald's Omega and IRT-based reliability indices, offering practical recommendations that call for factor analysis before calculating alpha, the use of concomitant coefficients, and the adoption of modern models to ensure a more responsible assessment of the reliability of measurement instruments.
Keywords: : Cronbach's alpha, Scale reliability, Classical Test theory, Item Response theory, Reliability coefficient .

INTRODUCTION:

Reliability is one of the fundamental pillars of assessing the quality of measurement tools in the fields of psychology and education. Standardized tools must be as free of random errors as possible to ensure the validity and generalizability of results (AERA et al., 2014; Furr, 2018). Among the various reliability measures, internal consistency reliability has emerged as one of the most common indicators in research that relies on multi-item scales, due to its effectiveness in estimating the reliability of a tool through a single application (Cortina, 1993). Since its introduction in Cronbach's (1951) seminal work, Cronbach's alpha has dominated the field of internal consistency assessment for over seven decades. It represents an important mathematical generalization of the Kuder-Richardson coefficient (KR-20), allowing it to be applied to multilevel items rather than just binary items. Its ease of calculation, coupled with the widespread availability of statistical packages, has made it a default standard in many research fields, to the point that it has been described as "the most widely used coefficient in the social sciences for which it is perhaps unsuitable." (Sijtsma, 2009, p. 107)

However, the rise of Item Response Theory (IRT) as a more sophisticated theoretical framework has raised fundamental questions about the appropriateness of alpha in the modern measurement era. While classical test-response theory (CTT) and its models, including alpha, provide a holistic view based on the characteristics of the test as a whole, IRT focuses on analyzing the characteristics of each item individually and seeks to achieve criterion invariance, in which individual ability estimates and item characteristics change independently of the sample used (Embretson & Reise, 2000; Hambleton et al., 1991).

Comparative empirical studies reveal a methodological gap between the two approaches. While alpha measures test reliability through the variance of total scores, making it sample-dependent, IRT introduces the concept of a test information function (TIF), which reveals variation in measurement accuracy across different ability levels (Baker & Kim, 2004). This variation in accuracy, which alpha ignores as a single digit, has profound implications in diagnostic and high-stakes choices.

This gap deepens with the development of statistical alternatives within the CTT framework itself. McDonald's Omega emerges as a more theoretically robust alternative, especially when it violates the tau-equivalence assumption. Studies by Zinbarg et al. (2005) and Dunn et al. (2014) indicate that Omega provides more accurate estimates of true reliability, while alpha tends to be underestimated in many practical contexts. Despite these criticisms, alpha retains its place in research practice. As Raykov & Marcoulides (2017) note, alpha's historical and widespread use, along with its ease of interpretation, ensures it remains a useful tool for preliminary assessments and exploratory research. Hayes & Coutts (2020) also emphasizes that the practical differences between alpha and omega may be minimal in practical applications with simple factor structures and homogeneous items.

From this perspective, this critical review aims to analyze Cronbach's alpha at the crossroads of CTT an IRT. The article will address the following:

• Trace the historical and mathematical development of the coefficient.
• Analyze the theoretical foundations and underlying assumptions of both CTT and IRT.
• Critically evaluate the criticisms of alpha from an IRT perspective.
• Review contemporary alternatives such as omega and IRT scales.
• Provide practical recommendations for practitioners and researchers.
This review also aims to provide an integrated conceptual framework that helps researchers understand when Cronbach's alpha is sufficient and when the research context requires moving to more advanced models, thus contributing to improving the quality and reliability of standard practices in the fields of psychology and education.

HISTORICAL ORIGINS AND THEORETICAL DEVELOPMENT

Although Lee J. Cronbach is most widely associated with this coefficient, the conceptual roots of the idea of internal consistency predate his 1951 article "Coefficient Alpha and the Meaning of Internal Consistency in Tests." Researchers had previously sought to develop practical methods for estimating test reliability through a single application, to avoid the problems of repeated application in the test-retest method.
In the same context, Kuder and Richardson (1937) presented a set of formulas, the most prominent of which was the formula known as the KR-20 coefficient. This provided an elegant solution for estimating the reliability of tests consisting of binary (true/false) items. The formula relies on the variance of individual items and the variance of the total score, and it represented a qualitative leap at the time (Nunnally & Bernstein, 1994).
What Cronbach (1951) did was not simply introduce a new coefficient, but rather an important mathematical generalization. He realized that the KR-20 formula was limited to binary items (0:1), while psychological and educational research was increasingly turning to multilevel scales such as Likert scales. In his article, he presented an expanded formula that could be applied to any item with a response scale, while maintaining the basic statistical logic. He explained that his coefficient, which he called alpha (α), could be understood as the average of all possible reliability coefficients that could be obtained from dividing the test into two parts in all possible ways (Cronbach, 1951). This generalization made alpha a more powerful and reliable measure than traditional split-half methods, which rely on a single splitting method that may not be optimal (Cortina, 1993). Cronbach's role was not limited to providing the mathematical formula, but rather expanded the discussion about the theoretical meaning of internal consistency itself. He discussed in more depth what is meant by "item homogeneity" and linked this to the theoretical construct to be measured, which contributed to the development of the relationship between the concepts of reliability and validity (Cronbach & Meehl, 1955). The ease of calculating the formula with subsequent computer advances made it more popular, to the point that it became synonymous with internal consistency reliability in the minds of many researchers. This later led to criticisms about its misuse and application in inappropriate contexts, as Sijtsma (2009) warned. The emergence of Cronbach's alpha represents a natural evolution in metric thought, from the exclusivity of binary items with the KR-20 to the flexibility of scaled measurement, and from the reliability of single-item division to the average reliability of all possible divisions.

MATHEMATICAL FORMULA AND TERMS OF USE

Cronbach's alpha value, or alpha coefficient (a), is calculated using the following formula:

$$a = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_T^2}\right)$$

Where:

- K is the number of items on the scale.
- σ_T^2 is the variance of the total score on the scale.
- $\sum_{i=1}^{k}\sigma_i^2$ is the sum of the variances of the individuals' scores on each item on the scale.

INTERPRETING CRONBACH'S ALPHA (A)

Interpreting Cronbach's alpha is a crucial step, and it is often oversimplified by relying on general rules that can be misleading. While the classic classifications presented by Nunnally (1978)which suggest a value of 0.70 as the minimum acceptable reliability in research are a useful starting point, accurate interpretation requires a deeper understanding of the context and the theory under consideration.

.1. What does alpha actually mean?

A high alpha value indicates that the items are well interrelated and that they measure a common characteristic. Mathematically, it can be interpreted as the proportion of variance in total scores attributable to true inter-individual variance versus variance due to random error between items. However, it is important to emphasize that alpha is an indicator of internal consistency and not necessarily unidimensionality (Cortina, 1993). A test can be multidimensional and still yield a high alpha value if the dimensions are strongly interrelated, which highlights the need for factor analysis before relying on alpha.

2. Interpretation in the context of the type and purpose of the instrument:

The required alpha level varies depending on the nature and purpose of the instrument:

• In the exploratory stages of research: A value between 0.60 and 0.70 may be acceptable for developing a new instrument, where the goal is to improve the instrument in the future (Nunnally & Bernstein, 1994).
• In basic and applied research: A value of 0.70 or higher is a common and widely accepted standard for supporting the consistency of an instrument.
• In diagnostic or high-stakes tests (such as professional competency certifications or clinical diagnosis), a value of at least 0.80 or 0.90 is expected to ensure high accuracy and reduce errors in making critical decisions (Streiner, 2003).

3. Interpretation Limitations and Common Fallacies:

There are several fallacies to be aware of when interpreting alpha:

1. A high alpha does not imply validity: A high alpha value ensures that the instrument measures "something" consistently, but it does not guarantee that this "something" is the stated theoretical construct. Validity is an independent characteristic that must be proven by other evidence (AERA et al., 2014).

2. A very high alpha (>0.95) may not be desirable: This may indicate the presence of redundant items, where items are almost identical in wording and meaning. This prolongs the administration time without adding new information, and may reduce the content validity of the test (Sijtsma, 2009).

3. Alpha is not a measure of temporal stability: As discussed previously, a high alpha value does not guarantee that the test will produce similar results if readministered two weeks later. This type of reliability requires calculating a test-retest coefficient.

4. Alpha is affected by the number of items: The coefficient is sensitive to the length of the test. The greater the number of items (even if their quality is average), the greater the alpha value. Therefore, the value must always be interpreted taking into account the number of items from which it was calculated (Cortina, 1993).

3.4. Accompanying Measures for More Accurate Interpretation

To avoid these misconceptions, accompanying measures should be reported when presenting the alpha value:

- Mean Inter-Item Correlation: This measure is less affected by the number of items, with the literature indicating that the optimal value for this mean is between 0.15 and 0.50 (Piedmont, 2014). Values below 0.15 indicate poor homogeneity, while values above 0.50 indicate possible duplication between items.
- Corrected Item-Total Correlation: Each item must contribute positively and significantly to the measure. Items with a corrected correlation below 0.30 are typically considered candidates for deletion unless there is a strong theoretical justification for retaining them (Field, 2018).

CONDITIONS FOR PROPER USE

Cronbach's alpha coefficient cannot be applied correctly and its value cannot be interpreted meaningfully unless a set of basic conditions are met. Ignoring these conditions is one of the main reasons for misuse and misinterpretation, which many researchers have warned against (Sijtsma, 2009; Tavakol & Dennick, 2011).

1. Unidimensionality Assumption

The most important condition for using Cronbach's alpha is that the test be unidimensional, meaning that all of its items measure a single, homogeneous theoretical dimension. This condition stems from the basic logic of alpha as an average of item correlations. If the test measures multiple dimensions (such as a test measuring anxiety and depression together), the correlations between items belonging to different dimensions will be low, leading to a misleadingly low alpha value that does not reflect the true homogeneity within each dimension individually (Cortina, 1993). In such cases, calculating alpha for the test as a whole becomes meaningless, and separate alpha values must be calculated for each distinct dimension (factor) identified through exploratory or confirmatory factor analysis (Hayes & Coutts, 2020; Furr, 2018).

2. Local Independence:

Random errors (e) in individuals' responses to items are required to be statistically independent of each other. This means that the response to one item does not systematically or directly influence the response to another item, taking into account the underlying construct being measured (Hambleton et al., 1991). This assumption is violated, for example, if two items share a common scenario or stimulus, or if the response to one item directly predicts the response to a subsequent item. Violating local independence artificially inflates the alpha value, because the shared variance between items will result not only from the construct being measured, but also from their interdependence (Raykov & Marcoulides, 2017).

3. Nature of Data and Level of Measurement:

Cronbach's alpha was originally designed for continuous or semi-continuous data. While it is commonly applied to Likert-scale data (which are essentially ordinal data), this application is practically acceptable when the number of scale categories is five or more and the distributions are not severely skewed (Zumbo et al., 2007). For nominal-level items (such as multiple-choice items with unordered answers) or binary items, more specialized formulas such as the KR-20 coefficient are theoretically more appropriate, given that alpha is considered a generalization of the KR-20, as previously mentioned (Kuder & Richardson, 1937; Cronbach, 1951).

4. Essential Tau-Equivalence:

This is a strict theoretical requirement that assumes that all items have the same true correlation with the target construct, i.e., they contribute equally to its measurement. In practice, this requirement is rarely fully met. However, alpha is considered a reasonable estimate of reliability as long as the items are sufficiently

homogeneous (i.e., they have positive correlations with each other and with the total score). If the item correlations with the construct vary greatly, alpha tends to estimate a lower bound on true reliability. Other measures, such as McDonald's omega (ω), may be more accurate (Dunn et al., 2014; Zinbarg et al., 2005). A large variation in item correlations with the total score indicates a potential violation of this assumption.

ALPHA (A) AND RELIABILITY METHODS IN CLASSICAL TEST THEORY (CTT)

The value of Cronbach's alpha lies not only in its practicality as a tool, but also in its position within the interconnected fabric of scales within Classical Test Theory (CTT). This unified mathematical framework provides a deeper understanding of the essential relationships between alpha and other reliability measurement methods, such as split-half reliability and the Kuder-Richardson 20 (KR-20), and reveals its nature as a minimum estimate of true reliability.

- **Relationship to split-half reliability and the Spearman-Brown formula: Mathematical Generalization:**

As mentioned, the idea of split-half reliability is based on dividing a test into two equal halves (X_1 and X_2) and calculating the correlation coefficient between them (P $\{X_1\ X_2\}$). Since this correlation measures the reliability of one half of the test, the Spearman-Brown formula is used to predict the reliability of the entire test:

$$\rho xx = \frac{2\,\rho x_1 x_2}{1 + \rho x_1 x_2}$$

- where ρxx is the reliability of the full test after correction.
- $\rho x_1\ x_2$ is the reliability of the half test.

Cronbach's (1951) main mathematical contribution was to demonstrate that his coefficient, alpha, does not depend on a single method of division. Rather, he showed that if a test of k items is divided into two parts in all possible ways, the alpha value calculated by the formula:

$$a = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_T^2}\right)$$

It equals the mean value of all split-half reliability values calculated using Spearman-Brown for those splits (Cortina, 1993). This means that alpha is a mathematical generalization of split-half reliability, making it a more stable and reliable estimate because it is not affected by the randomness of choosing a specific splitting method that may not be representative of the consistency of the test.

• Relationship with the Kuder-Richardson 20 (KR-20): The Special Case:

The relationship with the KR-20 coefficient represents the most striking example of alpha being an extension and generalization of the previous measures. Kuder and Richardson (1937) introduced the KR-20 formula specifically for binary-item tests (true/false), which are given either (1:0) or (1:1).

$$KR_{20} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} piqi}{\sigma_x^2}\right)$$

where pi is the proportion of correct answers to item i, and qi = 1 - pi is the proportion of incorrect answers.

Mathematical verification reveals that pi qi is actually the variance of the two-item variance σ^2. Substituting $\Sigma\ \sigma^2$ in the alpha formula for Σpi qi, the two formulas are completely identical (Kuder & Richardson, 1937; Cronbach, 1951). Thus, Cronbach's alpha coefficient is a generalization of the KR-20 coefficient to include multilevel items (such as Likert scales), while the KR-20 remains the special case of alpha when items are two-items.

Also within the classical test theory (CTT) model, both alpha and test-retest reliability can be viewed as estimates of the ratio of true variance to total variance, but they address different sources of measurement error.

- Test-retest reliability (ρ{XX-T}): Estimates reliability under the assumption that the primary source of error is variation in individuals or conditions over time. It is the Pearson correlation between the scores of the two tests.
- Cronbach's alpha (α): Estimates reliability under the assumption that the primary source of error is heterogeneity of item content within a single test session.

If all CTT assumptions are met, including construct stability across time and basic item homogeneity (Tau-equivalence), the two values should be close. However, theoretical work indicates that alpha tends to be a conservative estimate or a lower bound for true reliability in most practical situations, especially when assumptions of complete item homogeneity are not met (Sijtsma, 2009). The difference between alpha and test-retest reliability values indicates the dominance of a certain type of measurement error—internal versus temporal—which requires the researcher to interpret it within the context of their research and the nature of the construct being measured (Furr, 2018). This derivation and relationship demonstrate that Cronbach's alpha did not arise in a vacuum, but was a natural and well-established development within a continuous series of statistical measures aimed at estimating reliability. It generalizes and extends what Kuder and Richardson began, and unifies the concept of split-half into a single, robust formula that is independent of the method of division. This mathematical understanding enhances the researcher's awareness that the choice of a reliability measure and its interpretation must be based on the nature of the data (binary or graded) and the

underlying theoretical assumptions, ensuring a more accurate and responsible interpretation of the characteristics of the measuring instrument.

CRONBACH'S ALPHA (A) AND ITEM RESPONSE THEORY (IRT)

While Cronbach's alpha is deeply entrenched within the framework of classical test theory (CTT), the emergence of item response theory (IRT) as a more modern and powerful framework for psychological and educational measurement has shed new light on the strengths and weaknesses of alpha. IRT is not only an alternative to alpha, but it also provides a theoretical framework for understanding when and why alpha may fail to provide an accurate picture of test characteristics. This can be illustrated by the following:

1. Fundamental Differences Between CTT and IRT Perspectives:

The fundamental differences lie in the founding principles of each theory:

- In CTT and alpha, attention is focused on the characteristics of the test as a whole (such as mean, variance, and total score reliability). Alpha is based on total score variances and item variances, and is sample-dependent. That is, the alpha value can vary significantly depending on the group of individuals to whom the test is administered (Hambleton et al., 1991).
- In IRT, attention is focused on the properties of each individual item (item-level parameters), such as difficulty and discrimination, and the estimate of the individual's latent ability (θ). IRT aims to be invariant, meaning that the properties of the item should remain relatively constant regardless of the sample from which it was drawn, and that the individual's ability should remain constant regardless of the set of items they answer (Embretson & Reise, 2000).

2. Does IRT offer a fundamental critique of alpha?

From an IRT perspective, alpha suffers from several fundamental limitations:

- Sample dependence: As mentioned, alpha is not a constant for a test. If the test is administered to a very homogeneous sample (with low variance in the measured trait), the variance of the total score, σ^2, will decrease, leading to a lower alpha value, even if the items are excellent. The opposite is true for heterogeneous samples. In IRT, however, the item discrimination coefficient is estimated separately from the ability distribution in the sample.
- Assumption of Equivalence in Measurement Accuracy for All: CTT and alpha assume that measurement error is equal for all individuals. In reality, tests measure more accurately those individuals whose ability levels are close to the difficulty level of the items. IRT, on the other hand, provides an Item and Test Information Function (ITIF) that accurately indicates at which ability level (θ) the test is more accurate (i.e., lower measurement error). This demonstrates that the test has multiple invariances that depend on ability level, which the single-digit alpha provides ignores (Baker & Kim, 2004).
- Insensitivity to the ordinal nature of the data: Alpha is typically calculated using a Pearson correlation matrix, which assumes interval data, whereas Likert data are ordinal. Many IRT models explicitly address the ordinal nature of the data, providing a more appropriate estimate.

3. Practical Integration: What Does Alpha Offer in the Age of IRT?

Despite these limitations, alpha still has a role in research practice, even in the era of IRT dominance:

- Alpha as a preliminary and quick indicator: Alpha remains a useful and easy-to-calculate tool for a quick, initial estimate of internal consistency during initial instrument development.
- Alpha as a verification of the unidimensionality assumption of IRT: One of the basic assumptions underlying the application of most unidimensional IRT models is that items measure a single, dominant dimension. A high alpha value can be used (along with exploratory factor analysis) as preliminary evidence of this common unidimensionality before moving on to the more complex analyses specific to IRT (Zickar, 2020).
- Reliability as a shared goal: Ultimately, both approaches seek to reduce measurement error. What IRT offers is a more sophisticated means of understanding and enhancing this reliability across different ability levels.

SUMMARY OF THE RELATIONSHIP

It is clear from the above that the coefficient a occupies a pivotal position in the framework of classical CTT. However, an examination of the relationship between Cronbach's alpha and IRT reveals that alpha is a product of a theoretical framework with its well-known limitations. While it remains a practical and useful tool in many contexts, especially for summary reporting and preliminary analysis, modern measurement theory (IRT) reminds us that the single number provided by alpha is an oversimplification of the complex reality of psychological measurement. The modern perspective pushes us beyond the question "What is the reliability of a test?" to the more precise question: "What is the reliability of a test, for which group of individuals, and at what level of the trait measured?" In rigorous research, the use of alpha should complement, and not replace, the more sophisticated analyses provided by item response theory whenever possible. The comparison between classical CTT and IRT in their treatment of Cronbach's alpha can be summarized in the following table:

TABLE 1 Comparison between CTT and IRT in dealing with Cronbach's alpha coefficient

Comparison dimension	CTT	IRT
theoretical status	Alpha is the cornerstone and one of the most famous indicators of reliability, and is interpreted as the average of all possible split-half coefficients (Cronbach, 1951; Cortina, 1993).	Alpha is viewed as a limited, primitive index; it provides a preliminary estimate of reliability but does not accurately reflect item properties or accuracy variability across ability levels (Hambleton et al., 1991; Embretson & Reise, 2000).
Sample reliance	It is directly dependent on the sample, as its value decreases if the total variance is poor even with good paragraph quality.	It is clear that the properties of the items (discrimination and difficulty) should be independent of the sample, so α is not relied upon as the primary criterion for reliability.
Measurement assumptions	Unidimensionality and Tau-Equivalence assumption	Show that these assumptions are often unrealistic, and that α fails when dimensions are multidimensional or when item discrimination is different.
Interpretation of accuracy	It is presented as a single value encompassing the Reliability of the tool.	α is criticized because it ignores that measurement error is not constant, but varies across ability levels, which is illustrated by information functions in IRT (Baker & Kim, 2004).
Practical role	A standard and popular tool for assessing internal consistency, especially in exploratory research.	α is used only as a first step before moving on to more sophisticated indicators such as omega or information functions.

Contemporary Criticisms and Alternatives

Despite the widespread use of Cronbach's alpha for decades, the modern scientific literature, particularly in the past two decades, has witnessed an increase in substantive criticisms directed at it. It is no longer viewed as the undisputed gold standard, but rather as one of several reliability measures that carry significant limitations. This criticism is pushing researchers toward adopting more robust and theoretically and statistically appropriate alternatives.

1. Main Criticisms of Cronbach's alpha

The most prominent criticisms of the coefficient (a) can be summarized as follows:

1.1. High sensitivity to the number of items rather than their quality: One of the most prominent problems with alpha is that it is significantly affected by the number of items. The greater the number of items (even if they are of average or low quality), the higher the alpha value. This may lead researchers to add unnecessary or redundant items just to increase the reliability value, which lengthens the test without adding real new information and reduces the efficiency of the application (Cortina, 1993; Sijtsma, 2009).

1.2. The unrealistic assumption of unidimensionality and perfect homogeneity: As discussed earlier, alpha assumes that the test measures a single dimension. However, many psychological constructs are complex and multidimensional in nature (such as critical thinking or quality of life). Calculating alpha for the test as a whole in this case would be misleading, as it would yield a low value that does not reflect the true homogeneity within each subdimension (Zinbarg et al., 2005).

1.3. Being merely a "lower bound" for reliability under strict assumptions: Critics point out that alpha is merely an estimate of true reliability, and only if the assumption of "substantial equivalence" of the items is met. In practice, these assumptions are rarely met, meaning that alpha is often a low estimate of true reliability, especially for tests that are completely heterogeneous (Sijtsma, 2009).

1.4. Reliance on the Pearson correlation matrix for ordinal data: Alpha is typically calculated using the Pearson covariance or correlation matrix, which assumes that the data are on an interval scale. While common psychometric scales (such as Likert scales) are primarily ordinal, this discrepancy can lead to bias in estimation, although studies indicate that this bias is practically limited if the number of categories is large (5 or more) (Zumbo et al., 2007).

2. More Powerful Statistical Alternatives:

In response to these criticisms, statistical alternatives have emerged that address many of the limitations of alpha:

2.1. McDonald's Omega (ω): This is currently the most recommended and recommended alternative in the modern literature. In an important seminal study, Zinbarg, Revelle, Yovel, & Li (2005) examined the relationships between alpha, omega, and another alternative, beta. The researchers concluded that omega, particularly hierarchical omega (ωh), provides a more accurate estimate of reliability, especially for tests with

complex factorial structures. Their results showed that Cronbach's alpha tends to underestimate true reliability when the factor loadings of items vary greatly, which violates the tau-equivalence assumption of classical measurement theory.

Dunn, Baguley, & Brunsden (2014) concluded that Omega represents a practical and comprehensive solution to the problem of estimating internal consistency, calling for its adoption as a primary criterion in psychological and educational research.

Hayes & Coutts (2020) also provided a balanced perspective in their comprehensive critical review. While emphasizing Omega's theoretical superiority, they noted that practical differences between the two coefficients may be minimal in many research applications, especially when items are homogeneous and follow a normal distribution. However, they recommended using Omega as the primary criterion while continuing to report alpha for comparability with previous literature and to ensure methodological transparency.

In the context of multidimensional testing, Yang & Green (2015) demonstrated that using Cronbach's alpha on scales with a complex factor structure leads to misleading reliability estimates. Hierarchical Omega, on the other hand, offers an elegant solution to this problem by allocating variance between the general factor and subfactors, ensuring a more accurate estimate of overall reliability. Trizano-Hermosilla & Alvarado (2016) compared six different measures of internal consistency and found that Omega was the most stable and least affected by violations of statistical assumptions.

Omega comes in two main forms:

- Total Omega (ω_{total}): It does not assume complete homogeneity of the items, but rather uses factor loadings from a single-factor model to estimate reliability, making it more accurate when the contributions of items to construct measurement vary (Dunn et al., 2014). It is calculated according to the following formula:

$$\omega_{total} = \frac{\left(\sum_{i=1}^{K} \lambda_i\right)^2}{\left(\sum_{i=1}^{K} \lambda_i\right)^2 + \sum_{i=1}^{k} \theta_{ii}}$$

Where:

- $\left(\sum_{i=1}^{K} \lambda_i\right)^2$ is the sum of the factor loadings for all items.
- $\sum_{i=1}^{k} \theta_{ii}$ is the sum of the unique variance (error) for the item.

- Hierarchical omega (ω_h): Designed specifically for multidimensional tests, it divides the variance into the portion attributable to the general factor (the main construct) and the portions attributable to the subfactors. This is the optimal method for estimating the reliability of the overall score in tests with a complex factorial structure (Zinbarg et al., 2005). Many recent references recommend calculating and reporting omega alongside alpha, or considering it the main criterion when possible (Hayes & Coutts, 2020). It is calculated according to the following formula:

$$\omega_h = \frac{\left(\sum_{i=1}^{k} \lambda_{gi}\right)^2}{\left(\sum_{i=1}^{k} \lambda_{gi}\right)^2 + \sum_{i=1}^{m}\left(\sum_{i=1}^{kj} \lambda_{sj}\right)^2 + \sum_{i=1}^{k} \theta_{ii}}$$

Where:

- $\left(\sum_{i=1}^{k} \lambda_{gi}\right)^2$ is the factor loading of item i on the general factor.
- $\left(\sum_{i=1}^{kj} \lambda_{sj}\right)^2$ is the factor loading of item i on the subfactor.
- $\sum_{i=1}^{k} \theta_{ii}$ is the unique variance of item i.

2.2. Item Response Theory (IRT)-Based Measures: As discussed earlier, IRT introduces the concept of a Test Information Function (TIF) as a more sophisticated alternative to reliability. This function describes how accurately a test measures individuals at different levels of a latent trait (θ). This allows for a more accurate understanding of the strengths and weaknesses of a test across the entire ability spectrum (Embretson & Reise, 2000).

2.3. Stratified Alpha: This is a practical alternative to multidimensional tests, in which alpha is calculated separately for each dimension (stratum) and then combined using a formula that takes into account the variance of each dimension to yield a more accurate overall reliability (Cronbach & Shavelson, 2004).

PRACTICAL RECOMMENDATIONS FOR ADDRESSING ALPHA LIMITATIONS

To overcome alpha limitations, researchers can follow the following procedures:

1. Factor Analysis First: Always conduct an exploratory factor analysis (EFA) followed by a confirmatory factor analysis (CFA) before calculating alpha to verify the factorial structure of the instrument. If the instrument is multidimensional, reliability (alpha or omega) should be calculated for each dimension separately.

2. Reporting Associated Measures: When reporting alpha, always report:

• Average item correlation (should be between 0.15 and 0.50).

• Correlation of each item with the corrected total score (should typically be >0.30).

• Number of items used in the calculation.

For example, if we have a scale consisting of (10) items to measure anxiety, and the value of a = 0.88 while the value of ω-total = 0.79, the difference between the two indices reveals that a may have provided a higher

estimate of reliability due to its sensitivity to the number of items, while ω provides a more accurate estimate that takes into account the differences in factor loadings between items (Hayes & Coutts, 2020; Dunn et al, 2014).
3. Calculating Omega: In modern research, it is highly recommended to calculate the total omega coefficient (ω) as a primary or complementary estimate of alpha, especially with the availability of statistical programs that facilitate its calculation (such as the 'psych' package in R).
4. Focus on the quality of items, not their quantity: Tests should be designed with high-quality items with high discriminatory power, rather than relying on increasing the number to increase reliability.

PRACTICAL RECOMMENDATIONS FOR RESEARCHERS
In light of the above, the following practical recommendations can be made for researchers when using Cronbach's alpha:
1. Checking Unidimensionality: Before calculating alpha, an exploratory factor analysis (EFA) or confirmatory factor analysis (CFA) should be conducted to verify that the items belong to a single factor. If multiple factors appear, alpha should be calculated for each factor separately.
2. Examining Item-Total Correlation: The correlation coefficients of each item with the total score should be reviewed (after excluding that item). Items with a correlation coefficient less than 0.30 are typically candidates for deletion, as they do not contribute to measuring the common construct.
3. Comprehensive Reporting: When reporting alpha in research, the coefficient value and the number of items used in its calculation should be included. It is also useful to report the mean inter-item correlation, with a value between 0.15 and 0.50 generally considered acceptable (Piedmont, 2014).
4. Considering Alternatives: In modern research, it is recommended to calculate and report both McDonald's alpha and omega (ω) to provide a more complete picture of the instrument's reliability, especially if there are doubts about item homogeneity (Hayes & Coutts, 2020).

Future Prospects for Reliability Estimation
With the development of measurement methodologies, relying on a single coefficient is no longer sufficient to estimate reliability. Recent trends have begun to focus on:
1. Bayesian Models: These allow for more flexible estimates of reliability with small samples or non-normal data (Zickar, 2020).
2. Reliability in multilevel models: especially when using hierarchical data such as students within classes or patients within clinics (Raykov & Marcoulides, 2017)
3. Integration of artificial intelligence (AI) in measurement: Machine learning algorithms are used to detect weak items and improve the accuracy of reliability estimates (Zickar, 2020).

CONCLUSION

For more than seven decades, Cronbach's alpha has been a cornerstone of assessing the quality of measurement tools in the social sciences. Its ease of calculation and interpretation has made it an indispensable tool for researchers. However, advances in research methodology and psychological measurement have highlighted the limitations of this coefficient and its strict requirements. It is no longer sufficient for a researcher to calculate an alpha value and compare it to a threshold of 0.70 to claim that their instrument is "reliable." Proper assessment requires a deep understanding of the nature of the construct being measured, the factorial structure of the instrument, and more powerful statistical alternatives such as omega. Cronbach's alpha remains a valuable tool in the researcher's toolbox, but its use must be critical and aware of its context and limitations, ensuring an accurate and responsible assessment of the quality of the measure.

FUNDING
The author acknowledges that no funding source supported the article, its writing, or its publication.

CONFLICT OF INTEREST
THERE WAS NO CONFLICT OF INTEREST IN PUBLISHING THE ARTICLE.

REFERENCES

1. American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing. American Educational Research Association*.
2. Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Marcel Dekker.
3. Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98
4. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

5. Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957
6. Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391-418. https://doi.org/10.1177/0013164404266386
7. Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046
8. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates Publishers.
9. Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.
10. Furr, R. M. (2018). *Psychometrics: An introduction* (3rd ed.). Sage Publications.
11. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991*). Fundamentals of item response theory*. Sage Publications.
12. Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures, 14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629
13. Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160. https://doi.org/10.1007/BF02288391
14. Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
15. Nunnally, J. C. (1978*). Psychometric theory* (2nd ed.). McGraw-Hill.
16. Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill.
17. Piedmont, R. L. (2014). *Inter-item correlations.* In A. C. Michalos (Ed.), Encyclopedia of quality of life and well-being research (pp. 3303-3304). Springer. https://doi.org/10.1007/978-94-007-0753-5_1493
18. Raykov, T., & Marcoulides, G. A. (2017). Thanks coefficient alpha, we still need you!. *Educational and Psychological Measurement, 77*(2), 205–211. https://doi.org/10.1177/0013164417727687
19. Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0
20. Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18
21. Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International *Journal of Medical Education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd
22. Travisson, D. H. (2001). *Introduction to test theory.* Johns Hopkins University Press.
23. Zickar, M. J. (2020). *Measurement theory and research*: Toward a new era. In Measuring and analyzing behavior in organizations (pp. 1-22). Routledge.
24. Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ωH: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7
25. Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*(1), 21-29. https://doi.org/10.22237/jmasm/1177992180