INTERNAL CONSISTENCY IN ACHIEVEMENT TESTS: ROLE OF CRONBACH'S ALPHA AND GUTTMAN PATTERN

M.J. Ermamatov, A.B. Normurodov, A.R. Sattiyev

Scientific and Educational Practical Center Under the Agency for Assessment of Knowledge and Competences, a.normurodov@uzbmb.uz

Abstract. This paper explores the significance of Cronbach's alpha in evaluating the internal consistency and reliability of test scores, particularly in educational assessments. Through a comprehensive analysis, we highlight how closely the response matrix of test results aligns with the Guttman pattern, emphasizing the importance of the number of items when making inferences using Cronbach's alpha. We conducted a field test with 423 students to evaluate a newly developed biology assessment tool, ensuring the reliability and validity of the test through robust statistical methods. The study presents a detailed examination of the Guttman pattern, its implications for test analysis, and the interpretation of Cronbach's alpha in different test scenarios. We also compare the descriptive statistics and reliability measures of simulated data sets with normal distribution and actual test data, providing insights into the practical applications and limitations of Cronbach's alpha in real-world settings. The findings underscore the importance of careful item selection and test design to enhance the reliability and interpretability of educational assessments

Keywords: assessment, classical test theory, achievement tests

1. Kirish

Analyzing test results is essential for comprehending information, making informed decisions, and deriving meaningful conclusions. This analysis can help teachers assess the reliability of their tests, evaluate student achievement, and identify areas for improvement in future exams. By accurately interpreting test scores, teachers can draw valid conclusions from the data and better prepare for post-exam discussions with students. Furthermore, this process helps evaluate the quality of the tests, assess their readiness for release, and track their developmental progress.

It is imperative to develop instruments that measure accurately. To advance this goal, numerous investigations have been conducted [1-4], and various statistical formulas have been introduced to quantify measurement properties. Among these, Cronbach's alpha [5] stands out as a

metric that specifically assesses the reliability of tests, particularly their internal consistency. For example, Beckmann and Jastrowski Mano [6] demonstrated excellent internal consistency for their measure with Cronbach's alpha ranging from 0.94 to 0.97, while Bharara and Duncan [7] reported a total scale alpha of 0.943, indicating strong interrelatedness of items. Chen et al. [8] similarly documented good internal consistency across interventions with alpha values between 0.67 and 0.91.

Reliability refers to the extent to which a test consistently measures the same concept across different instances or sets of items. One key aspect of reliability is internal consistency, which measures the degree to which the individual items of a test are consistent with one another in assessing the same construct. High internal consistency indicates that the test items are all measuring the same underlying trait or ability, contributing coherently to the overall assessment. For example, in a math test designed to assess algebraic skills, high internal consistency would suggest that the test items focus on algebra rather than unrelated topics like geometry or calculus.

Internal consistency is usually evaluated by examining the correlations among test items. Strong correlations between items suggest that they all measure the same construct. In contrast, low correlations could signal that the items assess different constructs or that the test lacks focus.

Reliability, and particularly internal consistency, is crucial because it indicates the consistency of the test results. A test with high reliability consistently produces the same outcomes when administered under similar conditions, reflecting a stable measurement of the intended construct. If a test lacks internal consistency, its results may be unreliable, meaning the test could produce inconsistent outcomes, and the scores may not accurately reflect the test-taker's ability in the targeted area. In contrast, validity concerns whether the test measures what it is intended to measure. It refers to the accuracy of the test in assessing the specific concept, trait, or ability it is designed to evaluate. Therefore, while reliability (including internal consistency) ensures that the test consistently measures a construct, validity confirms that the test actually measures that construct accurately. During the test development process, ensuring reliability involves selecting and crafting items that consistently measure the same construct. To ensure validity, test developers must also confirm that the items effectively assess the intended concept without measuring irrelevant or unrelated constructs.

Internal consistency is essential in educational testing because it ensures that the test items collectively measure the same underlying construct, contributing to the reliability of the assessment. Reliability refers to the degree to which test results are free

from errors – unwanted variations that can influence the outcome of a test. Errors can arise from various sources, such as differences in test-takers' performance due to external factors (e.g., anxiety, fatigue) or inconsistencies in how test items are interpreted. These errors can undermine the accuracy and consistency of test scores.

By ensuring internal consistency, test developers can reduce the likelihood of such errors and enhance the reliability of the test. When test items are internally consistent, they are less likely to introduce random errors that could distort the test results. This allows educators and test developers to confidently interpret test scores, knowing that the results reflect the specific ability or knowledge area the test is designed to measure, with minimal error.

One of the most common measures of internal consistency is Cronbach's alpha [5]. This statistic provides a numerical estimate of how well the test items measure the same construct. A high Cronbach's alpha (typically above 0.7) indicates good internal consistency, meaning the items reliably assess the same concept. For example, a reading comprehension test with a Cronbach's alpha of 0.85 suggests that the items consistently measure the construct of reading comprehension. The correct interpretation of Cronbach's alpha is crucial for accurately assessing the reliability of scales, a point thoroughly discussed by Cortina [4]. Cronbach's alpha is often misunderstood as a simple measure of internal consistency; however, Cortina emphasizes that it is influenced by multiple factors, including both the number of items on a scale and the inter-item correlations. A higher number of items can inflate alpha even if the inter-item correlations are relatively low, which may give a false sense of reliability.

Moreover, the shape of the score distribution and the pattern of inter-item correlations significantly influence the interpretation of Cronbach's alpha [1]. Alpha assumes tau-equivalence, meaning all items have equal true-score variances and measure the same underlying construct. When this assumption is violated, alpha can still appear high, which may be misleading. As Schmitt [3] cautions, a high alpha does not guarantee unidimensionality. Multidimensional tests – those measuring multiple constructs – can also produce inflated alpha values due to the number of items or item redundancy. While our analysis focuses on alpha, we recognize that additional evidence (e.g., dimensionality analysis) would be necessary to confirm the internal structure of the test. We note this as a limitation and encourage future work to incorporate such methods for a more comprehensive evaluation.

Given these limitations, researchers have explored alternative reliability coefficients that address the shortcomings of alpha. One such measure is McDonald's Omega, which

incorporates factor analysis into the estimation of reliability [9]. Omega is particularly advantageous when the assumption of tau-equivalence is violated, as it provides a more accurate representation of the reliability in scales where items have differing factor loadings. Omega distinguishes between general and specific factors, making it more appropriate for scales that may be multidimensional in nature.

Recent research, such as the work of Orcan [10], has examined and compared the estimates provided by Cronbach's alpha and McDonald's Omega. These studies suggest that while alpha remains a popular metric due to its simplicity, omega offers a more nuanced and potentially more reliable estimate, particularly when factor structure is a concern. As a result, many researchers are now advocating for the use of McDonald's Omega as a complement or even an alternative to Cronbach's alpha, especially in scales where the unidimensionality assumption may not hold.

In psychometric literature, reliability refers to the consistency or stability of measurement. It encompasses several forms, including test–retest reliability (the stability of scores over time), inter-rater reliability (the consistency of scores across different raters), and internal consistency reliability, which is the focus of this paper. Internal consistency refers to the degree to which items within a test measure the same underlying construct. It is typically assessed using statistical indices such as Cronbach's alpha or McDonald's omega. While all types of reliability relate to measurement consistency, internal consistency specifically evaluates the coherence among items within a single test administration.

In this paper, we aim to demonstrate that Cronbach's alpha indeed reflects how closely the matrix of responses from test results aligns with the Guttman pattern [11]. Guttman's contribution to the classical test theory is thoroughly examined and discussed in reference [12]. Additionally, we will explore why the number of items is crucial when making inferences using Cronbach's alpha. To this end, we will analyze the results from a field test in biology to assess the quality of the test before its formal application.

2. Method

2.1. Study Design

The study was designed to assess the internal consistency and reliability of a newly developed biology assessment tool using Cronbach's alpha and the Guttman pattern. The focus was on evaluating how well the test results aligned with the Guttman pattern, particularly in educational assessments.

In addition to reliability, the study also aimed to assess the validity of the tool. To determine validity, the test was evaluated using several established methods. First, content validity was assessed by reviewing the test items for relevance and coverage of the intended biological concepts, with input from subject matter experts. Next, construct validity was examined by analyzing how well the test scores correlated with other measures of similar constructs, such as performance on related biology assessments. Finally, the test's criterion-related validity was evaluated by comparing its results with real-world outcomes, such as student grades or performance in biology-related tasks.

2.2. Sample or Study Group

The study was conducted with a total of 423 students from various educational backgrounds. The participants included 179 students from academic lyceums, 185 from secondary schools, and 59 from training centers. This diverse sample was chosen to ensure that the test results would be representative of different student populations.

2.3. Data Collection Tools

The primary data collection tool was a 48-item biology assessment test specifically designed for this study. The items in the test were selected to represent varying levels of difficulty, aligning with the theoretical Guttman pattern. While the Guttman pattern offers an idealized model for how test items should correlate based on their difficulty levels, it is important to note that achieving perfect alignment with this pattern is rare in practice. The literature suggests that, in reality, tests that attempt to follow a Guttman pattern typically exhibit only moderate levels of alignment due to various factors, such as individual differences in test-taker performance and the complexity of real-world assessments. Therefore, while the theoretical Guttman pattern guided the design of this test, the expectation was that the data would show some deviation from this idealized model, reflecting the inherent complexities of educational assessments.

2.4. Data Collection Process

The test was administered across the three different educational settings – academic lyceums, secondary schools, and training centers. The process involved distributing the test to the students and collecting their responses, which were then used for statistical analysis. The students' responses were scored, with correct answers marked as '1' and incorrect answers as '0'.

2.5. Data Analysis Procedures

The data collected from the test was analyzed using Cronbach's alpha to determine the internal consistency of the test. The study also examined the alignment of the response matrix with the Guttman pattern. Descriptive statistics such as mean, median, mode, variance, skewness, and kurtosis were calculated to understand the distribution of the test scores. The analysis also involved comparing the real test data with simulated datasets to evaluate the reliability and interpretability of the results across different groups. This phase was crucial for evaluating the test's reliability, as indicated by measures such as Cronbach's alpha, and its validity across a representative student population. The scale of the test allowed for robust statistical insights, ensuring that the findings were well-grounded and reflective of the test's performance in realistic settings.

3. Results

3.1. Guttman pattern as a foundation

Louis Guttman [11], a prominent psychometrician, conceptualized the ideal test as one in which a person succeeds on all items up to a certain level of difficulty, and then fails on all items that are more difficult. When individuals and items are ordered by their raw scores, this produces a dataset characterized by a "Guttman pattern" [11]. Such data sets are not typically analyzable using Rasch analysis because each person or item in turn becomes an extreme score. In simpler terms, items in a Guttman pattern are arranged such that a person who agrees with or correctly answers a particular item will also agree with or answer all less difficult items correctly. This pattern suggests a unidimensional underlying trait.

The use of the Guttman response pattern in this study serves a primarily illustrative and pedagogical purpose. As a theoretical ideal, the Guttman pattern represents a perfectly ordered item-response matrix, where each item strictly increases in difficulty and each respondent answers all easier items correctly before missing harder ones. While such a structure can demonstrate the upper boundary of internal consistency (e.g., Cronbach's alpha approaching 1), it is not a realistic or desirable goal in practical test design. Real-world assessments rarely exhibit such deterministic patterns due to variability in human performance, guessing behavior, and multidimensional traits. Thus, the Guttman matrix in this study is used not as a test design model but as a

benchmark to highlight how item ordering and inter-item correlations can influence reliability coefficients.

To illustrate, consider a hypothetical scenario where 20 students each answer 20 questions. Correct answers are scored as '1' and incorrect ones as '0'. The resulting data, presented in Table 1, exemplifies a perfect Guttman pattern. Although such patterns are not analyzable using the Rasch model, they can be assessed through various other methods. It is crucial to recognize that interpreting extreme scores within classical test theory is challenging. Extreme scores suggest either an infinite ability level or extreme difficulty. For instance, a student who fails to answer any of the items on the test might succeed on a hypothetically easier 21st question, not included in the original test. Conversely, a student who answers all the provided questions correctly might fail an additional, more challenging question. Moreover, if all students answer a particular question correctly or if no students answer it, does this make the question infinitely easy or difficult, or are there potential respondents not yet tested who could affect this outcome? Can a scenario in which each successive student answers one more question than the last accurately represent a real class, or is it possible to structure the questions such that the first is answered by one student, the second by two students, and so on? Such considerations underscore the importance of the number and distribution of both questions and test-takers.

Table 1
Hypothetical response pattern of 20 students (Guttman pattern)

Item / person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

8	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
9	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
10	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
12	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
13	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
14	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
21	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Let's examine the components of the Cronbach's alpha formula, detailed at [13]:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{S^2}{\sum s^2}\right)$$

where k is the number of items, S^2 is variance of total score, $\sum s^2$ is the sum of variances of the students' responses to each item.

In the matrix of responses, units appear in the upper triangle, and zeros appear in the lower triangle when the abilities of the respondents and the difficulties of the questions are arranged in increasing order. This structured arrangement highlights the predictable response patterns aligned with the Guttman scale. Notably, the mean and median of the total scores across this dataset are both 10, although the mode of the total score remains undefined due to the diversity of response patterns.

As we analyze the variances in the responses of respondents to each item (denoted as s^2), we observe that these variances increase towards items of average difficulty and decrease symmetrically thereafter, as illustrated in Figure 1. This variance pattern reflects the concentration of errors around the median difficulty level, where respondent performance begins to diverge.

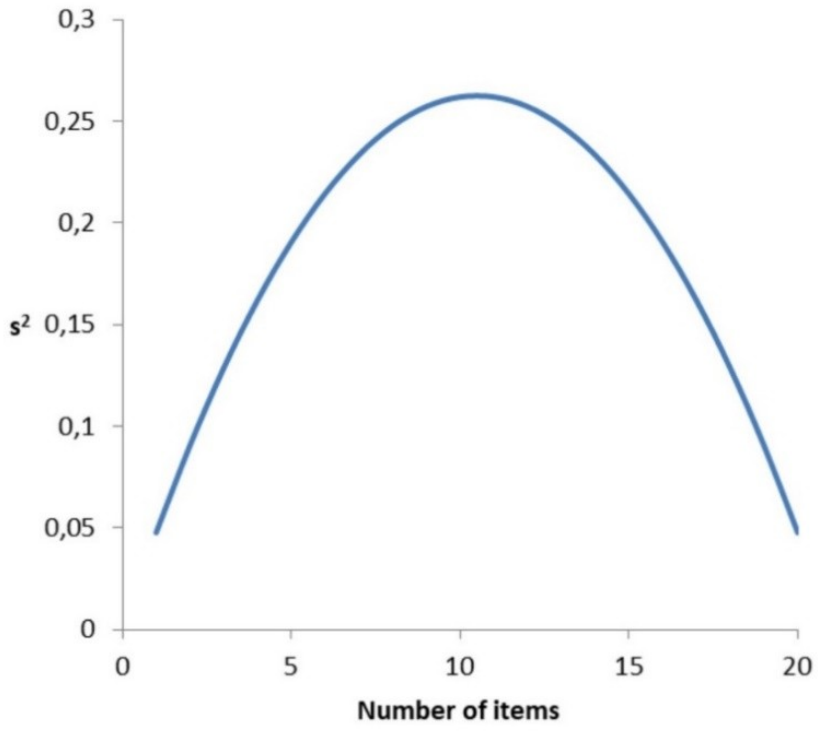

Figure 1. Variation of s^2 (item variance) with items arranged in increasing order of difficulty

The variance of the total scores (S^2) for such a pattern is calculated to be 38.5. Given these statistical properties, the calculated Cronbach's alpha for this pattern is exceptionally high at 0.95, indicating a high level of internal consistency within the test. This high reliability suggests that the test consistently measures some underlying construct. However, it is important to note that reliability does not necessarily guarantee validity. While the test is reliable in terms of internal consistency, further steps were taken to evaluate its validity. Validity concerns whether the test measures what it is intended to measure - i.e., the specific biological concepts it is designed to assess. The validity of the test was assessed through expert reviews of the content (content validity), as well as by examining the correlation between the test scores and other established measures of biological knowledge (construct validity).

These assessments, alongside the high reliability, support the validity of the test in measuring the intended biological constructs. This high alpha value underscores the reliability of the test in assessing the trait it is intended to measure, albeit within the confines of a structured and idealized response pattern as depicted. Let's see the level interpretation usually people use for the Cronbach's alpha in Table 2.

Upon examining formula, it becomes apparent that the number of questions, k, plays a critical role in

determining the value of Cronbach's alpha.

Table 2
Level interpretation for Cronbach's alpha

Alpha	Level interpretation
< 0.5	Unacceptable
> 0.5	Poor
> 0.6	Questionable
> 0.7	Acceptable
> 0.8	Good
> 0.9	Excellent

This component influences all other variables in the formula, affecting the overall calculation of reliability. Specifically, a smaller variance in responses to each individual question (s^2) and a larger variance of the total scores (S^2) contribute to an increase in Cronbach's alpha. This dynamic indicates that a diverse range of total scores across the test, combined with consistent individual responses, enhances the reliability measure.

Moreover, it is essential to consider the implications of an exceptionally high Cronbach's alpha, approaching 1. Such a high value typically suggests an ideal response pattern, as depicted in Table 1. While this might indicate excellent internal consistency, it can also hint at redundancy among items or a lack of sufficient challenge across the test's scope. In practical settings, an alpha value very close to 1 might not always signify an optimal test but could point to potential issues in test design that need further investigation.

As mentioned previously, achieving response patterns like those depicted in Table 1 is generally unrealistic in practical settings. Even with carefully selected groups and test items, it is highly improbable to achieve such a perfect pattern. A Cronbach's alpha value of 0.95, as calculated for the scenario involving 20 questions and 21 respondents, might itself be regarded as suspicious in real-life situations. Such a high value typically suggests an unusually consistent pattern of responses that may not reflect the diversity of an actual test-taking population.

In more realistic scenarios, one might encounter patterns that approximate but do not exactly match the Guttman pattern. For instance, the upper triangle of the response matrix may contain a few 'ones', while the lower triangle predominantly contains 'zeros'. Consider a more idealized scenario where, instead of having one student scoring '0', another '1', and so on, we have 100 students scoring '0',

another 100 scoring '1', etc. In this situation, the variance of the total score would be significantly higher, potentially leading to an even larger Cronbach's alpha. This occurs because the presence of larger groups of students scoring at the same level results in a pattern where 'zeros' and 'ones' occupy larger blocks within the matrix. Mathematically, this configuration suggests a higher probability of achieving a smaller sum of item variances Σs^2 and a larger total score variance S^2, thereby inflating the Cronbach's alpha value. Similar reasoning can apply to the selection and number of items included in the test.

Another aspect of the ideal patterns discussed so far is that the total scores from such patterns are uniformly distributed. However, in real-world situations, the distribution of total scores often approximates a normal distribution, especially when the sample size is large and the test measures a trait or ability that naturally varies within the population, as described by the Central Limit Theorem. This phenomenon is common with standardized tests and various types of assessments, where individual variations and a large number of questions can average out to produce a bell-shaped curve. Nevertheless, it's important to recognize that not all real-world situations will perfectly follow a normal distribution. The actual distribution can be influenced by the nature of the group being tested, the type of test, and its alignment with the abilities being assessed. For instance, if a test is too easy or too hard for the group being tested, a skewed distribution might be observed instead.

To understand these nuances in distribution, the concepts of skewness and kurtosis are employed. Skewness measures the degree of asymmetry in a distribution; a distribution with most data concentrated on the left and a longer right tail is said to be right-skewed or positively skewed, while one with the peak toward the right and a longer left tail is left-skewed or negatively skewed. Kurtosis, on the other hand, describes the 'tailedness' of the distribution. For kurtosis, we refer to the definition provided by Kevin P. Balanda and H.L. MacGillivray in "Kurtosis: A Critical Review," published in The American Statistician in May 1988 [14], which can also be found at [15]. In the context of this paper, we consider excess kurtosis as defined in the aforementioned reference. For the Guttman pattern discussed, the skewness is 0, indicating no asymmetry, and the kurtosis is 1.2, suggesting a relatively low level of peakedness.

Let's now examine a simulated pattern for 500 individuals and 20 items, where the total scores are normally distributed. The frequency distribution of the total scores is presented in Figure 2, with bins chosen at a step size of 1. The solid line represents the theoretical normal distribution, which has the same mean, median, and mode as the total scores from the simulated data.

Descriptive statistics for this dataset are presented in Table 3. In a perfect normal distribution, the mean, median, and mode are identical, with both skewness and kurtosis equaling zero. The total score from the simulated data closely approximates this normal distribution profile. In contrast, analysis of the Guttman pattern data reveals a variance of 38.5, with both mean and median at 10, skewness at 0, and kurtosis at -1.2; the mode was undetermined. Additionally, the Cronbach's alpha for the Guttman pattern stands at 0.95, signifying excellent reliability.

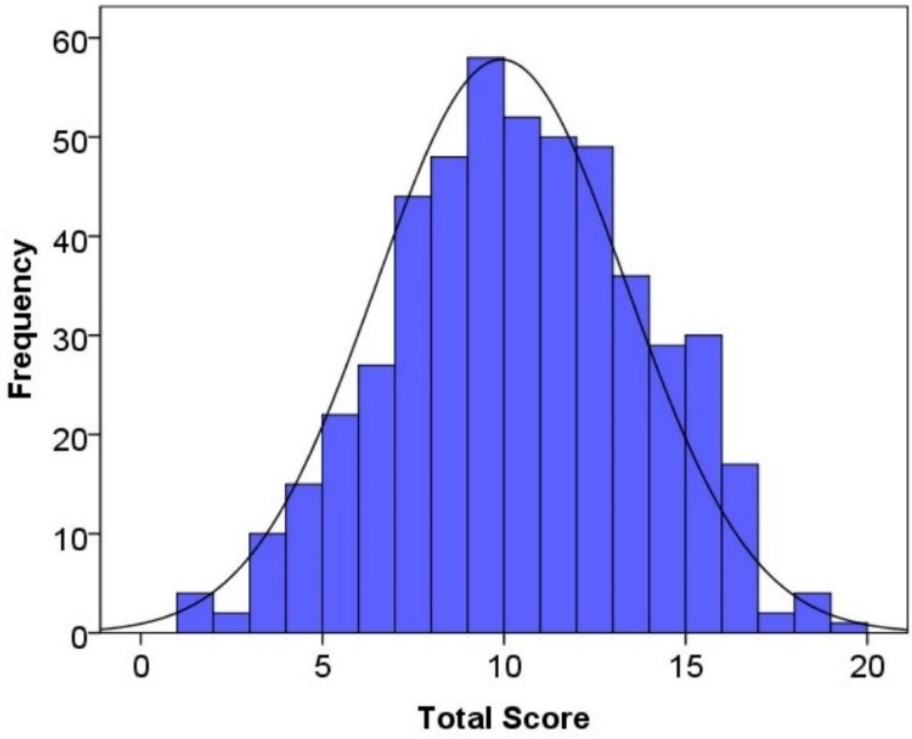

Figure 2. Frequency Distribution of Total Scores from the Simulated Data for 500 Individuals and 20 Items

The alpha for the simulated dataset is 0.75, indicating that the reliability of the test results from this 20-item test form is 'acceptable'. With careful selection of items and targeted learning strategies, there is potential to significantly improve this reliability score. Typically, however, a Cronbach's alpha between 0.75 and 0.8 is expected for a test form consisting of 20 items. Descriptive statistics for this dataset are presented in Table 3.

While a higher Cronbach's alpha is often interpreted as an indicator of greater internal consistency, it is important to recognize that alpha can also increase due to factors that may not reflect improved test quality. As noted by Cortina [4], alpha is sensitive not only to inter-item correlations but also to the number of items, meaning that adding similar or overlapping items can artificially inflate alpha. In such cases, a high alpha may suggest item

redundancy rather than meaningful coherence.

Table 3
Descriptive statistics for the simulated dataset

Number of students	500
Number of Items	20
Mean	9.93
Median	10
Mode	9
Variance	11.90
Standard deviation	3.45
Skewness	-0.061
Kurtosis	-0.41
Range	18
Minimum	1
Maximum	19

Furthermore, if added items are not conceptually relevant to the construct, this can lead to construct irrelevance, where the scale appears statistically reliable but lacks substantive validity. Therefore, alpha should be interpreted with caution, alongside considerations of content representativeness and unidimensionality. For comparison with Guttman pattern, we present the variation of s^2 (item variance) with items arranged in increasing order of difficulty in Figure 3. This visualization illustrates how the dependency changes when transitioning from a Guttman pattern to a normal one.

As shown in Figure 3, the s^2 for the simulated normal distribution, and consequently the sum of s^2 (3.38), has slightly decreased compared to that of the uniform distribution (3.67), a change attributable to the increase in the number of students from 21 to 500. However, the decrease in Cronbach's alpha for the simulated normal distribution, which is 11.90, as opposed to 38.5 for the uniform distribution, is due to the smaller S^2.

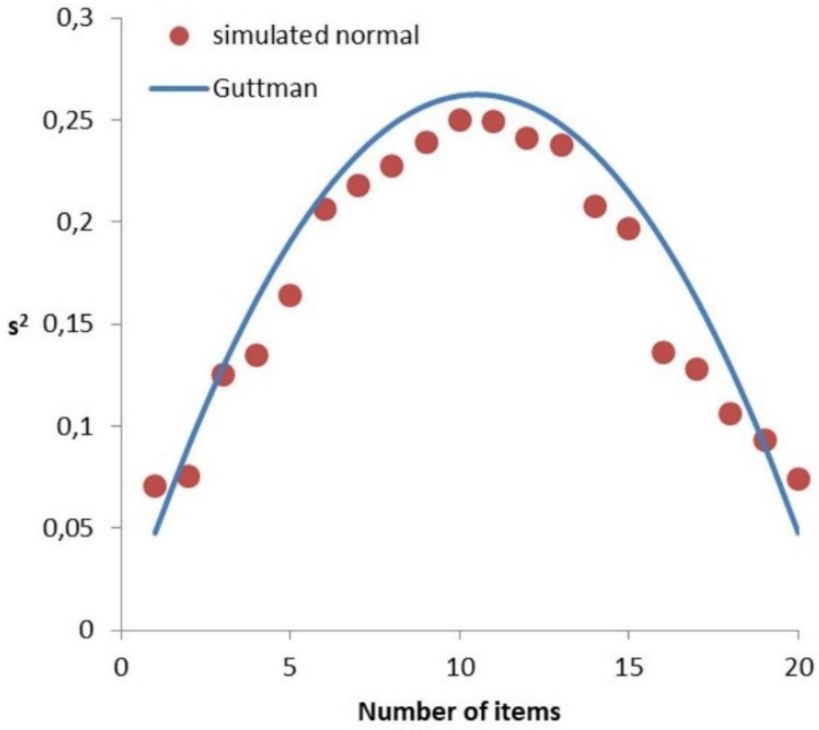

Figure 3. The variation of s^2 (item variance) with items arranged in increasing order of difficulty for simulated normal distribution

3.2. Normal distribution with 48 items

Before delving into the analysis of the test results, we think it is necessary to clarify what constitutes a typical Cronbach's alpha for scores from a test form with 48 items that are normally distributed. Let's now examine a simulated pattern for 500 individuals using 48 items, where the total scores are normally distributed. The frequency distribution of the total scores is presented in Figure 4, with bins set at a step size of 1. The solid line represents the theoretical normal distribution curve, which has the same mean, median, and mode as the total scores from the simulated data.

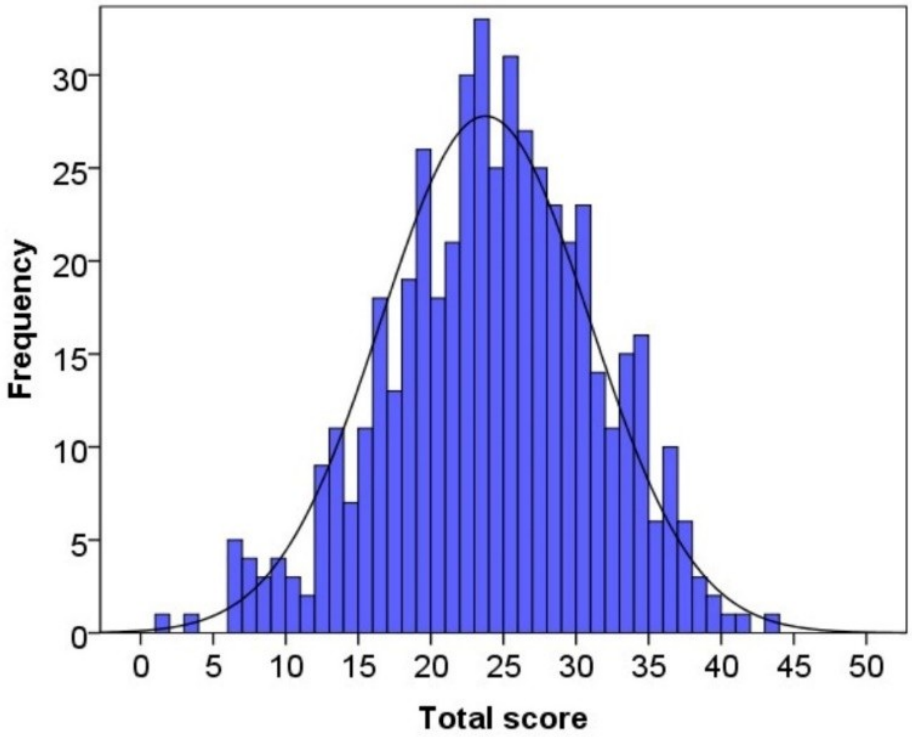

Figure 4. Frequency Distribution of Total Scores from the Simulated Data for 500 Individuals and 48 Items

Descriptive statistics for this dataset are presented in Table 4. As with the earlier 20-item dataset, the total scores from this simulated data closely approximate a normal distribution profile. The alpha for this simulated dataset is 0.86, indicating that the reliability of the test results from this 48-item test form is very close to 'excellent'. However, with careful selection of items and targeted learning strategies, there is potential to significantly improve this reliability score. Typically, a Cronbach's alpha around 0.9 is expected for a test form consisting of 48 items. Notably, the s^2 for the simulated normal distribution, and consequently the sum of s^2 (8.02), has increased compared to the 20-item dataset due to the larger number of items. However, the increase in Cronbach's alpha is attributed not to this individual item variance but rather to the enhanced effect of the total variance (S^2) which is a function of the larger number of items (48).

Table 4

Descriptive Statistics for the Simulated Dataset for 500 Individuals and 48 Items

Number of students	500
Number of Items	48
Mean	23.76
Median	24
Mode	24

Variance	51.50
Standard deviation	7.18
Skewness	-0.23
Kurtosis	-0.07
Range	42
Minimum	1
Maximum	43

3.3. Comparative analysis of descriptive statistics for real test results

Having established an understanding of the descriptive statistics of the Guttman pattern and the datasets simulated for 20 and 48 items, we now turn our attention to the analysis of a biology field test designed for 9^{th} – grade students. We prepared a test form consisting of 48 items, anticipating a typical Cronbach's alpha for this number of items. The test was administered to 179 students from academic lyceums, 185 students from secondary schools, and 59 students from training centers.

The frequency distribution of the total scores is presented in Figure 5, with bins set at a step size of 1. The solid line represents the theoretical normal distribution curve, which has the same mean, median, and mode as the total scores from the biology test data.

Descriptive statistics for this dataset are presented in Table 5. Although the mean (27.32) and median (29) of the distribution are relatively close, the mode of this distribution (12) is significantly lower. Visually, as observed in Figure 4. This multifaceted pattern is quantitatively supported by Table 5, where the skewness value indicates a symmetrical distribution, while the kurtosis value resembles that of the uniform distribution discussed earlier in relation to the Guttman pattern.

This distribution pattern may be linked to the three distinct groups (academic lyceums, secondary schools, and training centers) to which the test was administered. Comparing this with simulated data, the variances s^2 (item score variance) and S^2 (total score variance) show notable differences: s^2 increased to 10.45 and S^2 to 132.39, compared to 7.18 and 51.50, respectively, in the simulated data for the same number of items. Comparing this with simulated data, the variances s^2 (item score variance) and S^2 (total score variance) show notable differences: s^2 increased to 10.45 and S^2 to 132.39, compared to 7.18 and 51.50, respectively, in the simulated data for the same number of items.

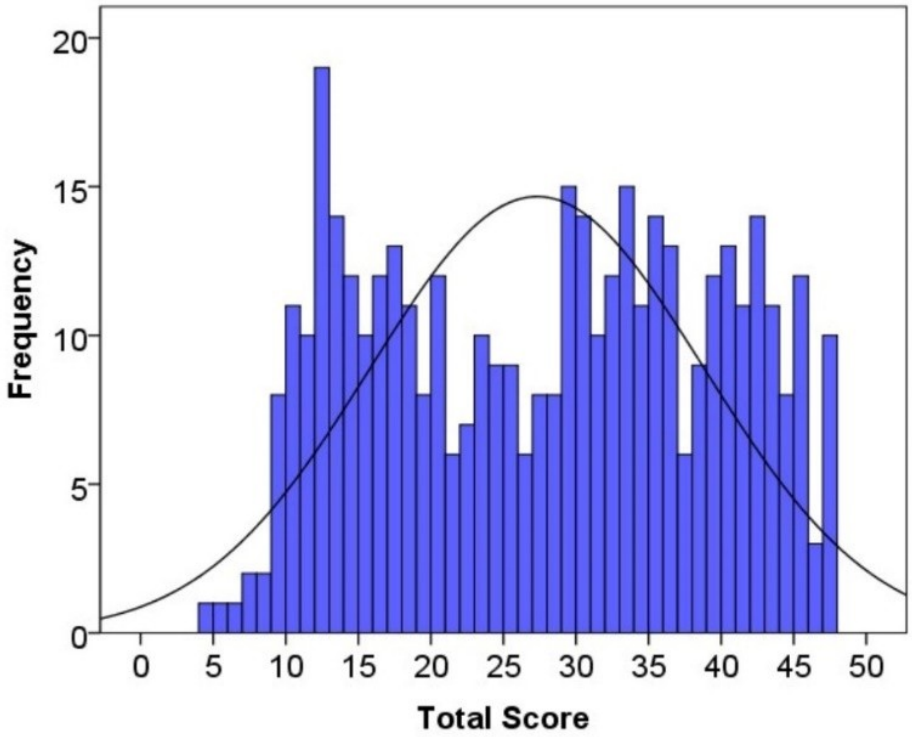

Figure 5. Frequency Distribution of Total Scores for the Dataset from the biology test for 423 Individuals and 48 Items

Table 5
Descriptive Statistics for the Dataset from the biology test

Number of students	423
Number of Items	48
Mean	27.32
Median	29
Mode	12
Variance	132.39
Standard deviation	10.45
Skewness	-0.04
Kurtosis	-1.24
Range	43
Minimum	4
Maximum	47

These changes suggest that s^2 is moving in a direction that could slightly decrease Cronbach's alpha, which is not desirable. Ideally, we would prefer a smaller variance among items, given the sufficient number of items available, as this typically stabilizes the alpha estimate. A higher item variance indicates inconsistency in how individual items measure the underlying ability, which could destabilize the reliability measure provided by Cronbach's alpha. As observed in Fig. 1 and 3, item variance

values are larger in the middle part of the graph where the items with medium difficulty are located. This implies that the test form used in the real test contains fewer items with “stronger and weaker” difficulty and more items with medium difficulty compared to both the simulated normal and Guttman distribution. Conversely, excessively small item variances may indicate that the test form contains an imbalance of “easy”, “hard” or a majority of “easy” and “hard” items with few “medium” difficulty items. Table 6 presents the percentage of correct answers for each item in the simulated normal distribution from 500 individuals and 48 items, the Guttman pattern similar to that in Table 1 for 48 items, and the real test administered to 423 individuals with 48 items. Analysis of this table reveals distinct patterns in item difficulty across the different testing scenarios. Notably, the real test dataset shows variations in item difficulty compared to the simulated distributions, with a greater proportion of items falling within the medium difficulty range.

Table 6

Percent of correct answer to each item for simulated normal distribution from 500 persons and 48 items, Guttman pattern like in Table 1 for 48 items and real test for 423 persons and 48 items

No	Guttman	Real	Normal	No	Guttman	Real	Normal
1	97,96	88,89	93,6	25	48,98	60,76	45
2	95,92	82,03	92,2	26	46,94	59,57	44,6
3	93,88	81,09	89,6	27	44,90	59,57	44,2
4	91,84	79,91	89,4	28	42,86	59,34	43,4
5	89,80	76,60	88,2	29	40,82	58,63	37,2
6	87,76	74,23	87,2	30	38,78	58,16	33
7	85,71	72,10	87	31	36,73	57,92	32,2
8	83,67	70,45	84,6	32	34,69	52,96	30,8
9	81,63	69,98	83,4	33	32,65	51,77	26
10	79,59	69,74	80,4	34	30,61	49,65	26
11	77,55	68,09	79	35	28,57	49,17	25,8
12	75,51	66,67	78,2	36	26,53	47,99	24,2
13	73,47	66,67	76,8	37	24,49	47,04	22,8

14	71,43	66,19	76,6	38	22,45	45,86	18,6
15	69,39	66,19	72,6	39	20,41	43,97	18
16	67,35	65,25	71	40	18,37	39,95	15,2
17	65,31	65,01	70,4	41	16,33	36,64	14,4
18	63,27	65,01	64,8	42	14,29	34,99	12,6
19	61,22	64,78	64,2	43	12,24	34,52	10,6
20	59,18	63,83	63,2	44	10,20	29,08	10,2
21	57,14	62,88	58,6	45	8,16	28,13	9,6
22	55,10	62,17	55,6	46	6,12	21,75	9,4
23	53,06	61,23	54,8	47	4,08	20,80	6,2
24	51,02	60,99	48,8	48	2,04	13,48	6

Though the significant increase in S^2 favourably impacts the alpha value, consequently yielding a Cronbach's alpha of 0.94, slightly higher than that of the simulated normal distribution, this discrepancy may be attributable to the three distinct groups, potentially leading to a more dispersed total score distribution. To explore this assumption further, we can examine the total score distributions of these distinct groups separately. Figures 6-8 display the total score frequencies, with bins set at a step size of 1 for these groups. The solid lines represent the theoretical normal distribution curve, which have the same means, medians, and modes as the total scores from the biology test data obtained from lyceums, schools, and training centers, respectively. To address potential heterogeneity among test-takers from different educational backgrounds, we conducted a stratified analysis by institution type - academic lyceums, secondary schools, and training centers – as shown in Figures 6–8 and Tables 7–9.

Descriptive statistics for the datasets from lyceums, schools and training centers separately are presented in Table 7-9.

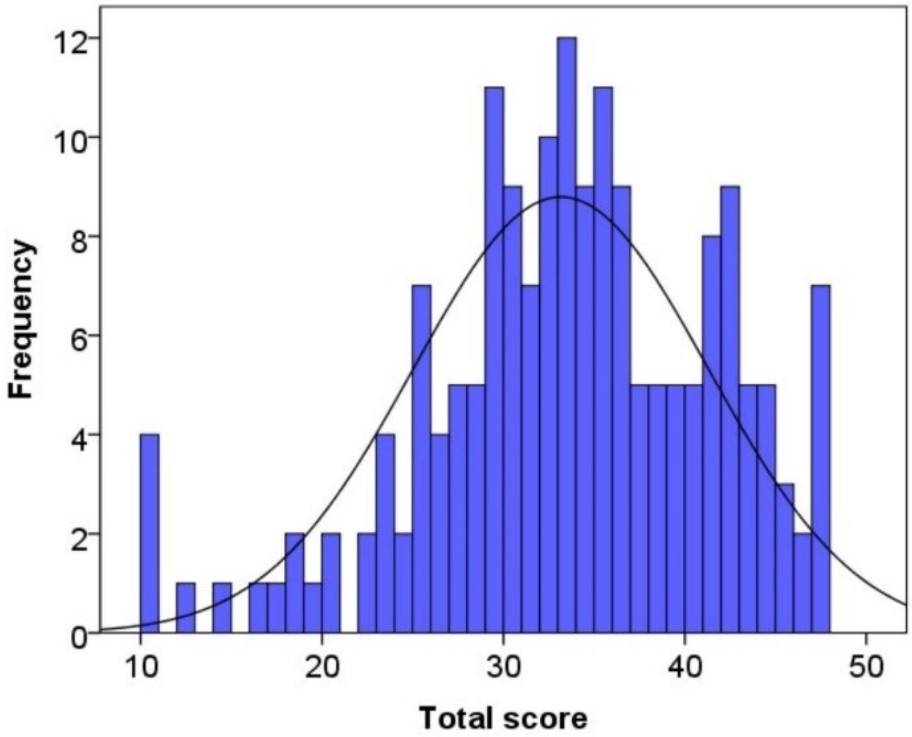

Figure 6. Frequency Distribution of Total Scores for the lyceum for the Dataset from the biology test for 179 Individuals and 48 Items

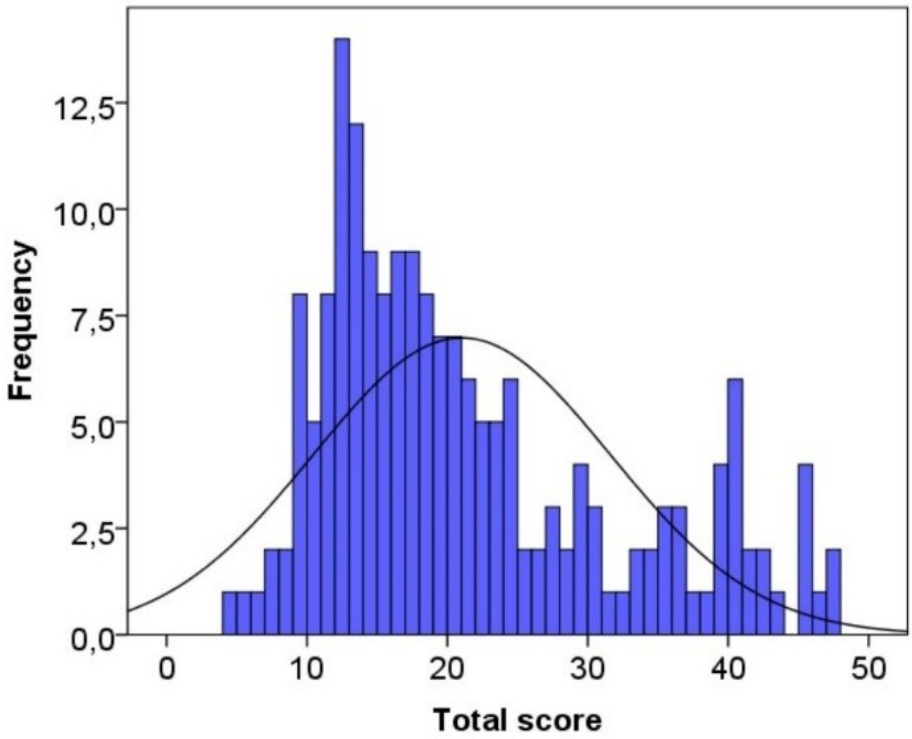

Figure 7. Frequency Distribution of Total Scores for the schools for the Dataset from the biology test for 185 Individuals and 48 Items

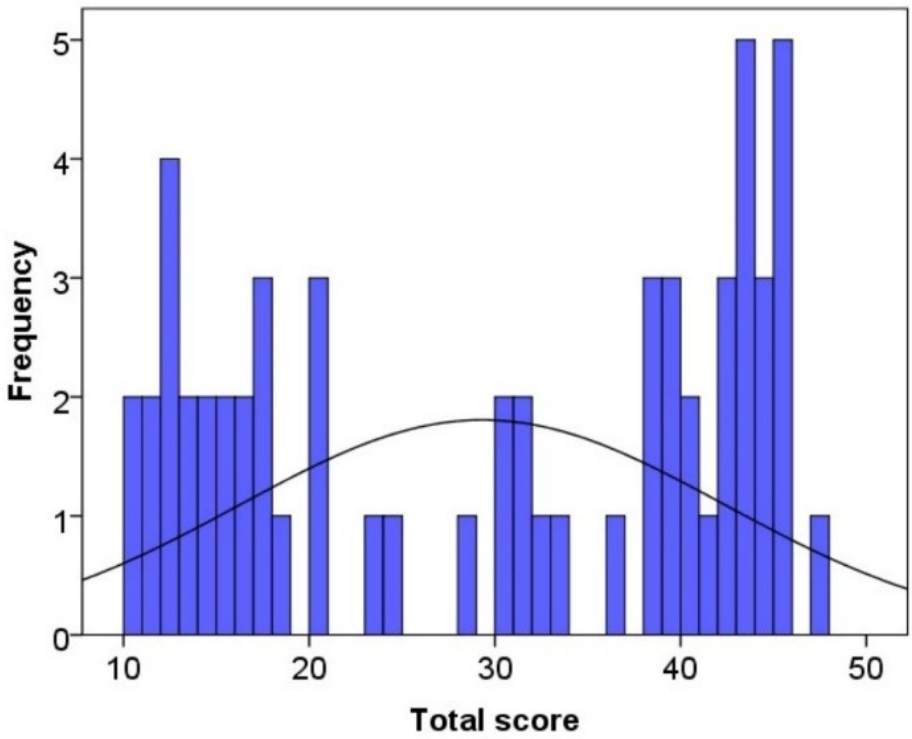

Figure 8. Frequency Distribution of Total Scores for the training centers for the Dataset from the biology test for 59 Individuals and 48 Items

Table 7
Descriptive Statistics for lyceums for the Dataset from the biology test

Number of students	179
Number of Items	48
Mean	33.16
Median	33
Mode	33
Variance	65.97
Standard deviation	8.12
Skewness	-0.57
Kurtosis	-0.40
Range	37
Minimum	10
Maximum	47

Table 8
Descriptive Statistics for schools for the Dataset from the biology test

Number of students	185
Number of Items	48
Mean	21.02
Median	18

Mode	12
Variance	111.92
Standard deviation	10.58
Skewness	0.86
Kurtosis	-0.28
Range	43
Minimum	4
Maximum	47

Table 9

Descriptive Statistics for training centers for the Dataset from the biology test

Number of students	59
Number of Items	48
Mean	29.32
Median	31
Mode	43
Variance	169.81
Standard deviation	13.03
Skewness	-0.175
Kurtosis	-1.67
Range	37
Minimum	10
Maximum	47

As is seen from Fig. 6 and Table 7 for the lyceums the total score distribution is more likely to the normal distribution, since the mean, median and mode for this distribution are almost the same. Variance (65.97) and standard deviation (8.12) more looks like to those of (51.50, 7.18, respectively) simulated data for 48 items and 500 persons. As a result, Cronbach's alpha 0.89 is also close to that of the simulated normal distribution for 48 items and 500 persons. From these one can imagine "zeros" and "ones" at lower and upper triangles as compared to the Guttman pattern for 48 items – number of "ones" in the lower triangle are more than "zeros" at the upper triangle. Minimum of the total score for the lyceums is 10, maximum is 47 and the range is 37. The students, who would answer less than 10 items correctly, do not exist in this group. This is understandable since the lyceums are specialized in biology.

Slightly higher absolute value of -0.57 of the skewness as compared to the simulated normal distribution (-0.24) shows slightly left skewed distribution - more data concentrated on the right and a longer tail on the left. This also mean for this group the difficulty of test form is a little bit "easier" than those of the "medium"

difficulty of the test form for the whole groups. Absolute value of kurtosis -0.40 is much larger than that of the simulated normal distribution (-0.07) and much less than that of the Guttman pattern (-1.2). The frequency distributions at the total score 10 and 47 affect to the value of kurtosis. These are seen visually in Fig. 6, too.

As seen from Figure 7 and Table 6, the total score distribution for the schools is more likely to be right-skewed. The mean, median, and mode for this distribution are not the same. The variance (169.18) and standard deviation (13.03) are significantly larger than those of the simulated data for 48 items and 500 persons (51.50 and 7.18, respectively). Consequently, Cronbach's alpha is 0.96, higher than that of the simulated normal distribution for 48 items and 500 persons. In this case, the increased kurtosis leads to greater dispersion in the total scores, resulting in a higher Cronbach's alpha value. The minimum total score for the schools is 4, the maximum is 47, and the range is 43, which is larger than in the case of the lyceums. There are no students who scored less than 4, indicating that the schools are not specialized in biology. However, some students answered up to 47 questions correctly, similar to the lyceums.

A higher positive skewness value (0.86) compared to the simulated normal distribution (-0.24) indicates a right-skewed distribution, with more data concentrated on the left and a longer tail on the right. This suggests that the test form is slightly more challenging for this group compared to the overall medium difficulty. The absolute value of kurtosis (-0.28) is larger than that of the simulated normal distribution (-0.07) but less than that of the Guttman pattern (-1.2). This is visually evident in Figure 7.

As seen from Figure 6 and Table 7, the total score distribution for the lyceums is closer to a normal distribution, with the mean, median, and mode being nearly the same. The variance (65.97) and standard deviation (8.12) are more similar to those of the simulated data for 48 items and 500 persons (51.50 and 7.18, respectively).

Consequently, Cronbach's alpha is 0.89, close to that of the simulated normal distribution for 48 items and 500 persons. This suggests that there are more "ones" in the lower triangle compared to the Guttman pattern for 48 items. The minimum total score for the lyceums is 10, the maximum is 47, and the range is 37. No students scored less than 10, which is understandable given that the lyceums specialize in biology.

A slightly higher absolute value of skewness (-0.57) compared to the simulated normal distribution (-0.24) indicates a slightly left-skewed distribution, with more data concentrated on the right and a longer tail on the left. This suggests that the test form is slightly easier for this group compared to the overall medium difficulty. The absolute value of kurtosis (-0.40) is much larger than that of the

simulated normal distribution (-0.07) and much less than that of the Guttman pattern (-1.2). This is visually evident in Figure 6.

As seen from Figure 8 and Table 7, the total score distribution for the training centers is closer to a uniform distribution than a normal distribution. The distribution contains three sub-distributions that also resemble uniform distributions, likely due to the presence of beginners, intermediate, and advanced groups, as training centers accept students throughout the year. The mean, median, and mode are not the same. The variance (111.92) and standard deviation (10.58) are larger than those of the simulated data for 48 items and 500 persons (51.50 and 7.18, respectively). Consequently, Cronbach's alpha is 0.92, higher than that of the simulated normal distribution for 48 items and 500 persons. The increased skewness contributes to a higher dispersion in the total scores, resulting in a higher Cronbach's alpha value. The minimum total score for the training centers is 10, the maximum is 47, and the range is 37. No students scored less than 10, which is understandable given that the training centers specialize in biology.

A smaller absolute value of skewness (-0.176) compared to the simulated normal distribution (-0.24) indicates a sufficiently symmetric distribution, with slightly more data concentrated on the left. The absolute value of kurtosis (-1.67) is larger than that of the simulated normal distribution (-0.07) and even larger than that of the Guttman pattern (-1.2), indicating a much flatter peak and thinner tails compared to a normal distribution. This is visually evident in Figure 8.

Discussion and conclusion

The comparative analysis of the different test patterns reveals several critical insights regarding the reliability and validity of assessment tools. Firstly, the Guttman pattern exemplifies an ideal, yet unrealistic, response distribution that results in a high Cronbach's alpha. This ideal pattern serves as a useful benchmark for understanding the upper limits of internal consistency within a test. However, real-world tests rarely achieve such perfection, reflecting more complex and varied patterns of student responses. The real test data, derived from a biology field test, shows a distribution that deviates from the idealized Guttman pattern but still maintains a significant degree of internal consistency as indicated by its Cronbach's alpha. This suggests that while perfect patterns are unattainable, a well-designed test can still exhibit strong reliability. The mean, median, and mode discrepancies, along with the multimodal distribution observed, hint at underlying differences in the test-taker population, likely attributable to the distinct groups (academic lyceums,

secondary schools, and training centers) involved in the study.

Furthermore, the analysis underscores the importance of item difficulty distribution in shaping the reliability of a test. The real test's item difficulties, predominantly falling in the medium range, suggest that balanced item selection is crucial for maintaining internal consistency without inflating Cronbach's alpha artificially. This balance ensures that the test measures a broad spectrum of abilities while avoiding redundancy or overly homogeneous item difficulties.

The observation that medium-difficulty items exhibit higher variance suggests that such items contribute most to test discriminability, particularly among test-takers of average ability. This pattern aligns with psychometric principles, where items located near the population mean offer the greatest information. These findings can inform item selection strategies by highlighting the value of including a balanced distribution of item difficulties, with emphasis on moderately challenging items to enhance measurement precision.

Comparative statistics from simulated normal distributions with 20 and 48 items reinforce the understanding that increasing the number of items generally enhances reliability, provided the items are well-calibrated. The significant increase in Cronbach's alpha observed in the 48-item simulated dataset confirms this, aligning with the expectations for tests of this length.

The study highlights the importance of analyzing test patterns and understanding the components influencing reliability metrics like Cronbach's alpha. While ideal patterns like the Guttman scale offer theoretical benchmarks, practical test design must account for the inherent variability in real-world data. By carefully selecting items and considering their difficulty distribution, educators and test designers can develop reliable assessment tools that accurately measure student abilities and provide meaningful insights into their performance.

The findings from the biology field test, supported by simulated data, highlight the importance of test structure in achieving high reliability. A well-designed test can produce strong reliability metrics, but excessively high Cronbach's alpha values can raise concerns beyond those discussed by Cortina [4] and Nunnally & Bernstein [1]. While a high alpha is often seen as a positive indicator of internal consistency, values that are too high may also signal potential issues such as item redundancy or imbalanced item difficulty.

Redundancy occurs when multiple items are essentially measuring the same construct in highly similar ways, leading to inflated alpha scores without necessarily improving the overall quality of the test. This lack of diversity in item content can diminish the test's

ability to capture the full breadth of the underlying construct. Similarly, an imbalance in item difficulty – where too many items are either too easy or too difficult for the target population – can lead to an artificially high alpha, masking deeper issues with the test's discriminatory power across different ability levels.

Therefore, excessively high alpha values should prompt a careful review and adjustment of the test. It is essential to ensure that the items are varied enough to cover the complexity of the domain being assessed while avoiding unnecessary overlap that may inflate reliability metrics without adding value to the measurement process.

Ultimately, the goal is to design balanced and reliable assessments that not only achieve strong psychometric properties but also reflect the diversity and complexity of the population being tested. This requires a thoughtful approach to test development, ensuring that the test captures a wide range of abilities and nuances while avoiding overemphasis on specific subdomains or redundant item content. Achieving this balance enhances the test's ability to provide meaningful and accurate information about the constructs of interest, leading to more valid inferences about the tested population.

Future research could further explore the impact of different test-taker populations on reliability metrics and investigate methods for optimizing item selection to enhance both reliability and validity. Additionally, applying these findings across various subjects and educational levels would help generalize the conclusions and support the development of robust assessment tools in diverse educational contexts.

References

1. J. Nunnally & L. Bernstein. Psychometric theory (3rd ed.). McGraw-Hill. 1994.
2. R. Cohen & M. Swerdlik. Psychological testing and assessment. McGraw-Hill Higher Education. 2010.
3. N. Schmitt. Uses and abuses of coefficient alpha. Psychological Assessment, 8(4), 350-353, 1996.
4. J. M. Cortina. What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78(1), 98-104, 1993.
5. L. J. Cronbach. Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334, 1951.
6. E. A. Beckmann & K. E. Jastrowski Mano. Initial development and validation of the School Anxiety Inventory–College Version (SAI-CV). Psychology in the Schools, 60, 2540–2563, 2023. https://doi.org/10.1002/pits.22879
7. G. Bharara & S. Duncan. Preliminary development and validation of the positive school transition readiness survey (PSTRS). Psychology in the Schools, 61, 1217–1237, 2023. https://doi.org/10.1002/pits.23108
8. F. Chen et al. Enhancing children's well-being using a Malaysian-adapted version of Super Skills for Life. Psychology in the Schools, 61, 3894–3907, 2024. https://doi.org/10.1002/pits.23258
9. R. P. McDonald. Test theory: A unified treatment. Lawrence Erlbaum Associates, 1999.
10. F. Orcan. Comparison of Cronbach's alpha and McDonald's omega for ordinal data: Are they different? International Journal of Assessment Tools in Education, 10(3), 709–722, 2023. https://doi.org/10.21449/ijate.1271693
11. L. Guttman. The basis for scalogram analysis. In S. A. Stouffer et al. (Eds.), Measurement and prediction, Princeton University Press, pp. 60-90, 1950.
12. D. W Zimmerman, R. H. Williams, B. D. Zumbo & D. Ross. Louis Guttman's Contributions to Classical Test Theory. International Journal of Testing, 5(1), 81–95, 2009. https://doi.org/10.1207/s15327574ijt0501_7
13. B. Griffin. Calculating Cronbach's alpha in Excel. Psychology of Business. 2023. https://psychologyofbusiness.beehiiv.com/p/calculating-cronbachs-alpha-excel.
14. K. P. Balanda & H. L. MacGillivray. Kurtosis: A critical review. The American Statistician, 42(2), 111–119, 1998.

15. S. Brown. Measures of shape: Skewness and kurtosis. 2011. https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/Simon?action=AttachFile&do=get&target=kurtosis.pdf. 2011.

O‘ZLASHTIRISH TESTLARIDA ICHKI MUVOFIQLIK: KRONBAX ALFA VA GUTTMAN NAMUNASINING ROLI

M.DJ. Ermamatov, A.B. Normurodov, A.R. Sattiyev

Bilim va malakalarni baholash agentligi huzuridagi Ilmiy-o‘quv amaliy markazi, 100084, Toshkent sh., Bog‘ishamol k., 12, a.normurodov@uzbmb.uz

Qisqacha mazmuni. Ushbu maqolada test ballarining ichki muvofiqligi va ishonchliligini ta’minlashda, baholashda Cronbach alfa ko‘rsatkichining ahamiyati o'rganildi. Keng qamrovli tahlil orqali test natijalarining javoblar matritsasi Guttman namunasiga qanchalik mos kelishi ko‘rsatilib, Kronbax alfa yordamida xulosa chiqarishda topshiriqlar sonining muhimligi ta’kidlandi. Ilmiy tadqiqot uchun biologiya fanidan 423 nafar o‘quvchi ishtirokida test sinovi o‘tkazildi va testning validligi hamda ishonchliligi statistik usullar bilan tahlil qilindi. Tadqiqotda Guttman namunasining batafsil tahlili, uning test tahlilidagi nimani anglatishi hamda Kronbax alfani turli testlar uchun qanday talqin qilinishi keng yoritilgan. Shuningdek, normal taqsimotga ega simulyatsiya qilingan va haqiqiy test ma’lumotlarining tavsif statistikasi hamda ishonchlilik qiymatlari taqqoslanib, Kronbax alfaning amaliy qo‘llanilishi va cheklovlari muhokama qilingan. Natijalar topshiriqlarni ehtiyotkorlik bilan tanlash va testni puxta loyihalash baholashning ishonchliligini hamda talqin qilinish imkoniyatlarini oshirishda naqadar muhim ekanini ko‘rsatdi.

Kalit so‘zlar: baholash, klassik test nazariyasi, o’zlashtirish testlari