Understanding Kuder-Richardson Formula 20 (KR-20): Definition and Calculation

Authored by

Mohammed looti

November 1, 2025

Introduction to Reliability and Dichotomous Items

In the field of psychometrics, assessing the quality of measurement tools, such as tests or surveys, is paramount. A crucial aspect of quality is **reliability**, which refers to the consistency of a measure. If a test is reliable, it should yield similar results when administered repeatedly under the same conditions. When dealing specifically with tests composed of items that can only be scored as right or wrong (known as **dichotomous items**), standard measures like Cronbach's Alpha are often used, but a more specialized measure is available: the **Kuder-Richardson Formula 20 (KR-20)**.

The KR-20 formula, developed by Frederic Kuder and M. W. Richardson in 1937, provides a robust estimate of the internal consistency reliability for tests where all items are scored dichotomously. This means each question must have exactly two possible outcomes, typically represented as 1 (correct/yes) or 0 (incorrect/no). This formula is widely applied in educational and psychological testing to ensure that the test items are functioning consistently together to measure a single underlying construct. Understanding and applying the KR-20 is fundamental for test developers seeking to validate their instruments.

While conceptually similar to Cronbach's Alpha, the KR-20 is tailored precisely for this binary response format, offering a more direct and accurate reliability assessment under these specific constraints. A high KR-20 value indicates that the items within the test are highly correlated and consistently measure the same characteristic, whereas a low value suggests inconsistency or heterogeneity among the items.

Understanding the Kuder-Richardson Formula 20 (KR-20)

The **Kuder-Richardson Formula 20 (KR-20)** is a specific method used to determine the internal consistency reliability of a test. Internal consistency refers to how well the items within the test measure the same underlying concept. The index is derived by considering the number of test items, the proportion of test takers who answered each item correctly, and the overall variability of the total scores. It is often preferred over simple split-half methods because it effectively calculates the average of all possible split-half reliabilities, offering a more comprehensive and stable estimate.

The application of KR-20 is restricted exclusively to tests where items are scored as **dichotomous items**. This limitation is crucial; if any item allows for partial credit, or if the scoring scale is continuous (e.g., Likert scale responses), the formula becomes inappropriate, and alternatives such as Cronbach's Alpha must be utilized instead. This strict requirement ensures that the statistical assumptions underlying the formula--particularly those related to item variance based on success and failure proportions--are met.

By providing a single coefficient ranging from 0 to 1, the KR-20 allows researchers and educators to quickly gauge the quality of their assessment tool. This coefficient summarizes the degree to which individual items contribute to a cohesive whole. Interpreting this value correctly is essential for making informed decisions about whether a test is suitable for its intended purpose, or if revisions to the item pool are necessary to improve measurement precision.

Components and Mathematical Breakdown of the KR-20 Formula

To fully appreciate the utility of the KR-20, one must examine its mathematical structure. The formula integrates several key components of test performance into a single, comprehensive reliability estimate. This formula is mathematically represented as:

KR-20 = (k / (k-1)) * (1 - Σpjqj / σ2)

Understanding what each variable represents is necessary for accurate calculation and interpretation:

k: Represents the **total number of questions** (or items) included in the test. The term k / (k-1) acts as a correction factor related to the number of items.
pj: Represents the **proportion of individuals** who answered question *j* correctly. This is essentially the item difficulty index for that specific item.
qj: Represents the **proportion of individuals** who answered question *j* incorrectly. Since the item is dichotomous, qj is simply calculated as 1 - pj.
Σ**pjqj**: This summation term is the sum of the item **variances** across all *k* items. For a dichotomous item, the variance of that item's score distribution is calculated as pj multiplied by qj. This aggregate measure reflects the total inconsistent variance across all items.
σ**2**: Represents the **variance of scores** for all individuals who took the test. This is the variance of the total raw scores obtained by the group of test-takers. This term captures the total observed variability in the test scores.

The core logic of the formula relies on the ratio of the sum of item variances (Σpjqj) to the total score variance (σ2). If the item variances are small relative to the total score variance, it suggests that the items are consistent and the test is highly reliable. Conversely, if the item variances account for a large portion of the total score variance, it signals internal inconsistency and low reliability.

Interpreting the KR-20 Coefficient

The resultant value for **KR-20** is a coefficient that always falls within the range of 0.00 to 1.00. This boundary is critical for interpretation: a value of 1.00 signifies perfect internal consistency (meaning all items measure the exact same construct with zero measurement error), while a value close to

0.00 suggests virtually no relationship among the items, indicating that the test is highly unreliable and essentially measuring random noise rather than a coherent trait.

In practical applications, achieving a perfect score of 1.00 is highly unlikely, as some degree of measurement error is always present. Therefore, benchmarks are used to judge the acceptability of the coefficient. Generally, coefficients above 0.70 are considered acceptable for research purposes, and values above 0.80 or 0.90 are often required for high-stakes testing, such as clinical assessments or certification exams. The required threshold, however, depends heavily on the context and the stakes associated with the test results.

It is important to note that the KR-20 coefficient is influenced by both the quality of the items and the length of the test. All else being equal, a longer test tends to yield a higher reliability coefficient. Furthermore, if the test is administered to a population with a very narrow range of abilities (low score variance), the KR-20 estimate may be artificially suppressed, even if the items are individually well-constructed. Therefore, interpretation must always be contextualized by considering the test length, the heterogeneity of the sample group, and the intended use of the assessment.

Detailed Practical Example: Calculating Kuder-Richardson Formula 20

To solidify the theoretical understanding of the **Kuder-Richardson Formula 20**, let us proceed through a detailed, step-by-step practical example. This scenario demonstrates how raw test data is processed to derive the reliability coefficient, utilizing statistical software features often found in programs like Excel or specialized statistical packages. Suppose an educator administers a short quiz consisting of seven binary-scored items (k=7) to a class of ten students (N=10). A score of 1 indicates a correct response, and 0 indicates an incorrect response.

The first step involves organizing the raw data, calculating individual student totals, and determining the performance statistics for each item. The results, showing the 10 students' responses across the 7 questions, are structured as follows:

	A	B	C	D	E	F	G	H	I	J
1	**Student**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Total Correct**	
2	**1**	0	1	1	0	1	1	1	5	
3	**2**	1	1	1	1	0	0	0	4	
4	**3**	1	1	1	1	0	1	1	6	
5	**4**	1	1	0	0	1	1	0	4	
6	**5**	0	1	1	1	1	0	1	5	
7	**6**	1	0	1	0	1	1	0	4	
8	**7**	1	1	0	0	0	0	0	2	
9	**8**	1	1	0	1	0	1	0	4	
10	**9**	0	0	1	1	0	0	0	2	
11	**10**	1	1	1	0	1	0	1	5	
12										
13										
14										
15										
16										
17										
18										
19										

From this raw data, we must calculate three critical components for the KR-20 formula: the number of items (k), the item variances (pjqj) for all items, and the total score variance (σ2).

The subsequent calculations involve four main stages:

Item Analysis (Calculating pj and qj): For each question (j=1 to 7), we calculate pj (proportion correct) by dividing the sum of correct answers by the total number of students (10). We then find qj (proportion incorrect) as 1 - pj.

Calculating Item Variances (pjqj): The variance for each item is computed by multiplying pj by qj.

Summing Item Variances (Σpjqj): We sum the variances calculated in the previous step across all seven items. This result forms the numerator term within the formula's parentheses.

Calculating Total Score Variance (σ2): We calculate the variance of the total scores (column I in the image, representing the sum of correct answers for each student) for the entire group of 10 students. Note that depending on the software, the sample variance (VAR.S) or population variance (VAR.P) might be used, though VAR.S is standard in psychometric applications unless

the sample represents the entire population.

The following screenshot illustrates how these calculations are typically performed in a spreadsheet environment, leading directly to the final KR-20 value:

	A	B	C	D	E	F	G	H	I
1	**Student**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Total Correct**
2	**1**	0	1	1	0	1	1	1	5
3	**2**	1	1	1	1	0	0	0	4
4	**3**	1	1	1	1	0	1	1	6
5	**4**	1	1	0	0	1	1	0	4
6	**5**	0	1	1	1	1	0	1	5
7	**6**	1	0	1	0	1	1	0	4
8	**7**	1	1	0	0	0	0	0	2
9	**8**	1	1	0	1	0	1	0	4
10	**9**	0	0	1	1	0	0	0	2
11	**10**	1	1	1	0	1	0	1	5
12									
13	**p**	0.7	0.8	0.7	0.5	0.5	0.5	0.4	
14	**q**	0.3	0.2	0.3	0.5	0.5	0.5	0.6	
15	**pq**	0.21	0.16	0.21	0.25	0.25	0.25	0.24	
16									
17	**k**	7.0000							
18	**Σpq**	1.5700							
19	**σ^2**	1.6556							
20	**KR-20**	0.0603							
21									
22									
23									

Here are the formulas used in various cells:

B13 (pj for Question 1): `=SUM(B2:B11) / 10` (Calculates the proportion of correct answers for Question 1.)
B14 (qj for Question 1): `=1-B13` (Calculates the proportion of incorrect answers.)
B15 (pjqj for Question 1): `=B13*B14` (Calculates the variance for Question 1.)
B17 (k - Total Items): `=COUNTA(B1:H1)` (Counts the number of items, k=7.)
B18 (Σpjqj - Sum of Item Variances): `=SUM(B15:H15)` (Aggregates the variances of all 7 items.)
B19 (σ2 - Total Score Variance): `=VAR.S(I2:I11)` (Calculates the sample variance of the total

scores.)
B20 (KR-20): `=(B17/(B17-1))*(1-B18/B19)` (Applies the main KR-20 formula using the derived components.)

Analysis of the Example Results and Implications

Following the precise calculation outlined above, the resulting **KR-20** value for this particular seven-item test administered to ten students turns out to be **0.0603**. This coefficient is exceptionally close to zero, necessitating a careful analysis of its practical implications for the test's quality.

A KR-20 value of 0.0603 indicates that the test possesses extremely low internal consistency reliability. In practical terms, this suggests that the individual items are not measuring a common construct; they are highly heterogeneous, or the measurement error is overwhelming the true score variance. This result implies that if the students were to take a parallel form of the test, their relative scores would likely change dramatically, rendering the current scores untrustworthy for ranking or making significant decisions about their knowledge or ability.

There are several reasons why such a low reliability coefficient might occur. The most common issues include:

Poor Item Quality: Some items might be confusing, ambiguous, or poorly worded, leading to random guessing rather than a measure of knowledge.
Multidimensionality: The test might be measuring more than one distinct construct. For instance, if four questions measure algebra skills and three measure verbal reasoning, the items are not internally consistent with respect to a single underlying trait.
Restricted Variance: If the sample (N=10) is too small or too homogeneous, the total score variance ($\sigma 2$) may be very low, artificially suppressing the KR-20 coefficient. However, even considering the small sample, a value this low usually points to fundamental problems with the test design itself.

The conclusion drawn from this analysis is unequivocal: the assessment tool, in its current form, is unreliable and should not be used for any formal evaluation. The test developer must revise the test, potentially removing non-performing items, rewriting ambiguous questions, or ensuring that all items align with a single, clearly defined construct before retesting and recalculating the KR-20.

Limitations and Alternatives to KR-20

While the **Kuder-Richardson Formula 20** is a powerful tool for specific types of assessments, it operates under stringent assumptions which also define its limitations. The primary limitation, as repeatedly stressed, is its strict dependence on **dichotomous items**. Any deviation from the 1/0

scoring structure invalidates the result. Furthermore, KR-20 assumes that all items are equally difficult and measure the underlying construct equally well, an assumption often violated in real-world test design.

For scenarios where these assumptions are not met, alternative measures of internal consistency are required. The most well-known alternative is **Cronbach's Alpha** (often simply called Alpha), which is a generalization of the KR-20. In fact, if all items on a test are scored dichotomously, the value of Cronbach's Alpha will be mathematically identical to the value of the KR-20. However, Cronbach's Alpha is designed to handle polytomous scoring (e.g., Likert scales ranging from 1 to 5) and continuous data, making it far more versatile for surveys and instruments that allow for graded responses.

Additionally, the Kuder-Richardson family includes the **Kuder-Richardson Formula 21 (KR-21)**. This simpler version is used when there is the even stronger assumption that all items have exactly the same difficulty (i.e., all pj values are equal). Because this assumption is rarely met in practical testing, the KR-20 is generally preferred as it accounts for varying item difficulty by using the item variance term (pjqj) for each question.

Ultimately, selecting the appropriate reliability coefficient--be it KR-20, KR-21, or Cronbach's Alpha--depends entirely on the nature of the assessment data and the scoring method employed. For robust, binary-scored tests, the KR-20 remains the definitive and authoritative measure of internal consistency reliability.

Additional Resources

The following tutorials provide explanations of terms commonly used when assessing the validity of tests and questionnaires: