Estimating the Psychometric Properties *(Item Difficulty, Discrimination and Reliability Indices)* of Test Items using Kuder-Richardson Approach (KR-20)

OPEN ACCESS

Manuscript ID:
EDU-2023-11036081

Volume: 11

Issue: 3

Month: June

Year: 2023

P-ISSN: 2320-2653

E-ISSN: 2582-1334

Received: 11.02.2023

Accepted: 18.05.2023

Published: 01.06.2023

Citation:
Ntumi, Simon, et al. "Estimating the Psychometric Properties (*Item Difficulty, Discrimination and Reliability Indices*) of Test Items Using Kuder-Richardson Approach (KR-20)." *Shanlax International Journal of Education*, vol. 11, no. 3, 2023, pp. 18–28.

DOI:
https://doi.org/10.34293/education.v11i3.6081

Simon Ntumi
University of Education, Ghana
https://orcid.org/0000-0001-7874-4454

Sheilla Agbenyo
Bia Lamplighter College of Education, Ghana
https://orcid.org/0000-0003-0307-7348

Tapela Bulala
Botswana University of Agriculture and Natural Resources (BUAN), Botswana
https://orcid.org/0000-0003-4084-1501

Abstract
There is no need or point to testing of knowledge, attributes, traits, behaviours or abilities of an individual if information obtained from the test is inaccurate. However, by and large, it seems the estimation of psychometric properties of test items in classroomshas been completely ignored otherwise dying slowly in most testing environments. In the quest to obtain sound and efficient test results, it is imperative that assessorsrely on some psychometric properties to make informed classrooms decisions. These psychometric properties can be estimated using Kuder-Richardson20 Formula. In this study, 30 multiple-choice items were administered and used for the study. The strength of each item was analysed by looking at their difficulty level and how theydiscriminated among the students. Reliability tests were also conducted in addition to the item analysis to observe the quality of the test as a whole. With lucid prose, KR-20 was used to estimate the psychometric properties of 30 set integrated science test items (which werescored dichotomously)to serve as a primer for assessorsin higher institutions. The procedure produced coefficient value of 0.6915which is approximately 0.7 implying that the reliability of the test was high.The procedure we used to arrive at the obtained coefficient is extensively outlined in the paper. We concluded thatthe suggested procedure (KR-20) for estimating psychometric properties may have a paradigm shift in classroom testing situations where it will communicate to teachers on the efficiency and process of teachermade tests. In essence, this could enhance the quest of obtaining the real knowledge, attributes, traits, behaviours or abilities of students by using test items that are reliable and dependable.

Keywords: Reliability, Kuder Richardson (KR-20), Coefficient, Psychometrics

Introduction

Despite the critical nature of psychometric properties with respect to the precision of test items in classrooms, the explanation of reliability remains contextual and contested.In the field of educational measurement and assessment, many authors (eg. Meijer, et al, 2013; Wanous & Reichers, 2016; Ginns & Barrie, 2014; Merrigan & Huston, 2019) have explained test reliability as the consistency of scores students would receive on alternate forms of the same test. In the work of Petters, et al (2015) they further explained reliability as the consistency with which a measuring instrument yields certain results when the entity being measured has not changed.

Drawing inferences, it must be asserted that due to differences in the exact content being assessed on the alternate forms, environmental variables such as fatigue, stress, anxiety, lighting condition, student error in responding, it is obvious that no two tests will consistently produce identical results. All these conditions are factors or conditions that could contribute to measurement errors or variations (Impara & Plake, 2012; Wombacher, 2018; Platukus, 2020). Combining all these prepositions on reliability, we could draw the inferences that a test result is said to be reliable if there is relative absence of measurement errors or variations (Merrigan & Huston, 2019; Ginns & Barrie, 2014; Andrich, 2014).

In the majority of studies on educational measurement and assessment, the reliability of test instruments or items are assessed by a coefficient, such as a Pearson Product-Moment Correlation Coefficient or Cronbach alpha are mostly employed. However, it is suggested and espoused in the literature that a Pearson correlation coefficient or Cronbach alpha are not appropriate or preferred for assessing reliability of classroom test items. This is to say that, more robust tool or procedure like Kuder-Richardson 20 is more appropriate, because in its index systematic variability is also treated as error (Wombacher, 2018; Adeleke & Joshua, 2015; Platukus, 2020).

Kuder-Richardson Formula 20 was derived on the assumption that the average covariance between items on different forms is the same as the average covariance on the same form. This formula is considered anestimate of the parallel form reliability because the average covariance of items with identical difficulty is expected to be higher than the average covariance of items of different difficulty.The Kuder-Richardson Formula 20, often abbreviated as KR-20 is a measure of internal consistency for measures that feature dichotomous items. As these are measures of internal consistency, they measure the extent to which all the items measure the same characteristic (Kuder & Richardson, 1937; Wombacher, 2018; Adeleke & Joshua, 2015; Platukus, 2020).

The Kuder and Richardson Formula 20 test checks the internal consistency of measurements with dichotomous choices (Mohajan, 2017; Ginns & Barrie, 2014; Tan, 2019; Platukus, 2020). It is equivalent to performing the split-half methodology on all combinations of questions and is applicable when each question is either right or wrong. A correct question scores 1 and an incorrect question scores 0. The most common use for the KR-20 is for the analysis of tests of ability or learning of students. These tests feature one correct answer for each item, meaning that even if the question presents the respondent with multiple options, only one of them would be considered correct and all the others incorrect. The fact that answers can be split into two categories, correct and incorrect, is what makes these types of questions dichotomous in nature, even when the item itself has more than two potential responses (Wombacher, 2018; Adeleke & Joshua, 2015; Saupe, 2017).

In other explanations, it is asserted thatthe KR-20 can even be used to analyze fill-in-the-blank questions, where there are no potential responses offered to respondents. It is almost meant to be employed when the questions vary in difficulty (Adeleke & Joshua, 2015; Heale & Twycross, 2015). The KR-20 cannot be used if the test allows for some responses to earn partial credit, as this would mean that the item is no longer dichotomous since responses could be grouped as correct, incorrect, or partially correct (Wombacher, 2018; Platukus, 2020; Adeleke & Joshua, 2015).

Based on the above assumptions, it can therefore be inferred that if a classroom teacher is looking to assess the reliability of a test that has a number of different formats of questions, the KR-20 is a good choice and more appropriate (Frey, et al, 2000). If the teacher is mixing multiple question formats, like true/false, multiple choice and fill-in-the-blank, the KR-20 would still be a good choice as it is likely that these questions range in difficulty. A fill-in-the-blank question is typically more difficult than a true/false question as it relies on the respondent to use unaided recall to remember the answer (Frey, et al, 2000; Platukus, 2020).

Matlock-Hetzel (2017) and Jackson (2012) in write up emphasized the advantage of using discrimination coefficient instead of discrimination index. Discrimination coefficients includes every single person taking the test even though only the upper (27%) and lower scorer (27%) are included in the discrimination index calculation process. According to Instructional Assessment Resources (IAR, 2011), Le (2012), El-Uri and Malas (2013), discrimination coefficients is a measure using point biserial correlation. The correlation, commonly known as Pearson product-moment correlation is computed to determine the relationship between student's performance in each item and their overall exam scores.

In Ghana, for one to be able to lecture any course at the tertiary level, educational assessment is one of basic courses to be undertaken. This course exposes assessors to basic rudiments in assessment practices and its principles which include how reliability coefficients can be estimated. However, it appears that most assessors in the tertiaryinstitutions are not abreast with the procedures in estimating the reliability of test items. This paper used Kuder-Richardson Formula (KR-20) to guide assessorson how they can estimate the reliability of test items.

Methods and Materials

Test Construction Procedures

In our quest to use the KR-20 formula to estimate the psychometric propertiesof test items, a test in Integrated Science was conducted to determine the extent to which students had mastered the content and behavioural outcomes required in the syllabus of Integrated Science. The test items were developed in accordance with the Bloom's (1956) taxonomy of educational objectives (see Table 1). The test consisted of 30 multiple-choice items which were based on the content in the syllabus for which students had been taught already thereby ensuring content validity of the test. The test items were constructed by science experts (integrated science examiners who construct test items for West Africa Examination Council).

Test Administration Procedures

The test developers selected four topics from the integrated science syllabus which include measurement, density, mass and atom. The test was conducted under a specified examination rules to help control some confounding or extraneous measurement errors. This was also to guide against the violation of validity and reliability assumptions.

The time allotted for the test was 30 minutes and this agreed with Alexander and Brown (2017) who arguedthat multiple-choice item which are based on factual thoughts must have the duration range 40-60 seconds per item. The test started 9:00 am prompt and so at exactly 9:30 am, the scripts were collected from the examinees. The scripts were marked and the results compiled and proceed for analysis. The data from the item difficulty and item discrimination analysis were each conveyed as mean and standard deviation of the total number of items.

Table 1 Test Blueprint or Table of Specification of how the Items were Developed

Content	Knowl	Compreh.	Appl.	Ana.	Syn	Eva.	Total
Measurement	4	4	2	0	0	0	10
Density	3	2	1	0	0	0	6
Mass	2	2	2	1	0	0	7
Atom	3	3	1	0	0	0	7
Total	12	11	6	1	0	0	30

Note: Scores of Administered Integrated Science Test Items, 2022, n=30

Data Analysis Procedure

The obtained data was analysed using SPSS, v25 and Itema software and was reported in descriptive statistics. Using the descriptive statistics, mean, variance and standard deviation were used tocompute value that were used to estimate the psychometric properties. The itema software was used to compute difficulty and discrimination indexes. The last property that is reliability coefficient was estimated using KR-20 formula. Similarly, SPSS v25 was employed in verifying the relationship between the item difficulty index and discrimination coefficient for each test item.

Results

To perform the analysis and report accordingly, item statistic was employed to evaluate the performance of individual test items utilizing student's responses to each test items on the integrated science test. The accrued results are presented in the subsequent Tables.

Table 2 Computations of Mean, Variance and Standard Deviation for the Test Items

Source	N	Min.	Max.	Mean	Std. Deviation	Variance
Candidates	30	1.00	30.00	15.5000	8.80341	77.500
Scores	30	12.00	27.00	18.5667	3.98863	15.909 (used for the computations reliability index)
Valid N	30					

Note: Scores of Administered Integrated Science Test Items, 2022, n=30

Computations of Item Difficulty and Item Discrimination

Item Difficulty

Item difficulty, commonly known as p-value refers to the proportion of examinees that responded to the item correctly. The p-value is calculated using the following formula:

$$p = R / T$$

where p = item difficulty index

R = the number of correct responses to the test item

T = the total number of responses comprises both correct and incorrect responses

The item difficulty index (p) ranges from 0.0 to 1.00. A high p-value indicates an easy item. Instructional Assessment Resources (IAR) acknowledged values of difficulty index and their evaluation as tabulated in Table 3.

Table 3 Evaluation of Item Difficulty for Item Analysis

Item Difficulty Index (p)	Item Evaluation
Above 0.90	Very easy item
0.62	Ideal value
Below 0.20	Very difficult item

Source: Instructional Assessment Resources (IAR, 2011)

Item Discrimination

Item discrimination index (D) is estimated by the formula, D = (UG-LG)/n. Where D = discrimination index, UG = the number of students in the upper group 27% who responded correctly, LG = the number of students in the lower group 27% who responded correctly and n = number of students in the upper or lower group. The value of discrimination index ranges between -1.0 to 1.0. The items were classified accordingly to their discrimination index with reference to Ebel's (as cited in El-Uri & Malas, 2013) guidelines.

Table 4 Evaluation of Discrimination Indexes for Item Analysis

Index of Discrimination	Item Evaluation
0.40 and above	Very good items; accept
0.30 – 0.39	Reasonably good but subject to improvement
0.20 – 0.29	Marginal items usually need and subject to improvement
Below 0.19	Poor items to be rejected or improved by revision

Source: Adopted fromEbel (as cited inEl-Uri & Malas, 2013)

The computed values for the Item difficulty (ρ) and Item discrimination (D) are presented in Table 5 and Table 6.

Table 5 Calculated Values of ItemsDifficulty (ρ)

Items	Item difficulty (ρ)	Remarks
#1	$\rho 1 = 0.60$	functioned well*
#2	$\rho 2=0.80$	Easy item
#3	$\rho 3=0.50$	functioned well*

#4	ρ4=0.30	functioned well*
#5	ρ5=0.40	functioned well*
#6	ρ6=0.80	functioned well*
#7	ρ7=0.30	functioned well*
#8	ρ8 =0.80	Easy item
#9	ρ9 =1.00	Very easy
#10	ρ10 =0.80	Easy item
#11	ρ11 =0.90	Easy item
#12	ρ12 =0.40	functioned well*
#13	ρ13=0.30	functioned well*
#14	ρ14=0.50	functioned well*
#15	ρ15=0.10	Very Difficult
#16	ρ16=0.60	functioned well*
#17	ρ17=0.90	Easy item
#18	ρ18 =0.40	functioned well*
#19	ρ18=0.30	functioned well*
#20	ρ20=0.90	Easy item
#21	ρ21 =0.80	Easy item
#22	ρ22=0.80	Easy item
#23	ρ23=0.70	functioned well*
#24	ρ24=0.50	functioned well*
#25	ρ25=0.80	Easy item
#26	ρ26=0.80	Easy item
#27	ρ27=0.40	functioned well*
#28	ρ28=0.90	Easy item
#29	ρ29=0.30	functioned well*
#30	ρ30=0.90	Easy item

Note: Scores of Administered Integrated Science Test Items, 2022, n=30

Item difficulty (ρ) is the proportion of examinees who score an item correctly in relation to the number of examinees who attempted the item (Impara & Plake, 2012). Impara and Plake further pointed out that the smaller the P-value the more difficult the item and when the ρ-value is large the item is easy. The ρ- index ranges between 0 and 1. Generally the recommended item ρ-value ranges between 0.3 and 0.7 to maximize test information and differences among the examinees (Iacobucci & Duhachek, 2013). Based on the above criteria, sixteen (16) of the items functioned well. These are items 1, 3, 4, 5, 6, 7, 12, 13, 14, 16, 18, 19, 23, 24, 27 and 29 with ρ- indices between 0.3 and 0.7. One of the items were found to be difficult (item, 15) with *P*-indices less than 0.3 and twelve (12) of the items were easy (2, 8, 10, 11, 17, 20, 21, 22, 25, 26, 28 and 30) with P-indices greater than 0.7. Notwithstanding, one cannot use item difficulty only to determine the effectiveness of items therefore, the need to find out how items discriminate between examinees.

Table 6 Calculated Values of how the Items Discriminated (D)

Items	Item difficulty (ρ)	Remarks
#1	D1= 0.40*	discriminated well*
#2	D2= 0.20	discriminated satisfactorily
#3	D3= 0.24	discriminated satisfactorily
#4	D4= 0.60*	discriminated well*
#5	D5= 0.54*	discriminated well*
#6	D6= 0.20	discriminated satisfactorily
#7	D7= 0.40*	discriminated well*
#8	D8= 0.00	did not discriminate
#9	D9= 0.00	did not discriminate
#10	D10= 0.10	low discriminating
#11	D11= 0.10	low discriminating
#12	D12= 0.40*	discriminated well*
#13	D13= 0.40*	discriminated well*
#14	D14= 0.60*	discriminated well*
#15	D15= 0.20	discriminated satisfactorily
#16	D16= 0.50*	discriminated well*
#17	D17= 0.30	discriminated satisfactorily
#18	D18= 0.40*	discriminated well*
#19	D19= 0.30	discriminated satisfactorily
#20	D20= -0.10	low discriminating
#21	D21= -0.10	low discriminating
#22	D22= 0.60*	discriminated well*
#23	D23= 0.00	did not discriminate
#24	D24= 0.60*	discriminated well*
#25	D25= 0.40*	discriminated well*
#26	D26= 0.10	low discriminating
#27	D27= 0.30	discriminated satisfactorily
#28	D28= 0.20	discriminated satisfactorily
#29	D29= 0.30	discriminated satisfactorily
#30	D30= 0.20	discriminated satisfactorily

Note: Scores of Administered Integrated Science Test Items, 2022, n=30

Item discrimination is a measure of the degree to which an item discriminates between students with high performance and students with low performance (Heale & Twycross, 2015). A discrimination index of 0.5 is of average discrimination power for standardised tests, a discrimination index of 0.4 or better is satisfactory. Items with discrimination index below 0.2 must either be discarded or rewritten (Mohajan, 2017; Adeleke & Joshua, 2015). To determine the D-value according to Whitney and Sabers (2014), when the total number of students taking the test is between 20 and 40, select the 10 highest-scoring and the 10 lowest-scoring papers but, one would have to keep the middle-scoring group intact. The D-index is obtained by subtracting the proportion of the low scoring group that responded to the item correctly from the proportion of the high scoring group that scored the item correctly.

A negative D-value indicates that more of the low achievers got the item correct than the high achievers. Such items are ambiguous or mis-keyed and therefore, must be either discarded or reviewed. Based on the above computed information, it is evidence that twelve (12) of the items discriminated well. That is their D-indices were 0.4 or greater. The items were 1, 4, 5, 7, 12, 13, 14, 16, 18, 22, 24, and 25. Ten (10) of the itemsindices were between 0.2 and 0.39 indicating that items discriminated satisfactorily. The items are 2, 3, 6, 15, 19, 27, 28, 29 and 30. Again, three (3) of the items had low discriminating indices (D < 0.2). These items are 2, 10, 11and 26.

Two of the items, thus items 20 and 21 had negative indices and three (3) items, thus item 8, 9 and 23 also had 0.0 indices meaning these items did not discriminate between high and low achievers. Regarding the two statistics above, (item difficulty and discrimination indices) it could be seen that seven items (1, 3, 4, 5, 7, 12, 13, 14, 16, and 18) functioned properly meeting the acceptable levels of both P and D. Items 2, 10, 11, and 26 although was easy, discriminated so well therefore require little revision. Items 8, 9, 15 must be discarded because they were too easy and did not discriminate. For item 20 which had negative discrimination, the key must be rechecked or must be replaced.

Table 7 Summary of the Functional State of the Items

State of items	Items	Number of items
Effective items	1, 3, 4, 5, 7, 12, 13, 14, 16, and 18	10
Require minor revision	2, 10, 11, and 26	4
Re-examination of keys or clarity	6, 17, 19, 20, 21, 22, 24, 25, 27, 28,	
29 and 30	13	
To be discarded	8, 9, and 23	3
Total		30

Note: Scores of Administered Integrated Science Test Items, 2022, n=30

Estimating of the Reliability Coefficient Using KR-20

Far back in 1987, Boyle and Radocy proposed using Kuder Richardson formula for analysing test with dichotomous items. Data from string instruments were divided into two sections. Kuder-Richardson 20, a formula which is based on item difficulty was used to analyse internal consistency of section A in the string instrument comprehensive test. The value of KR20 range between 0 to 1. The closer the value to 1 the better the internal consistency. The KR20 formula is commonly used to measure the reliability of achievement test with dichotomous choices. According to Wallen and Fraenkel (2013), one should attempt to generate a KR20 reliability coefficient of .70 and above to acquire reliable score. To estimate the reliability, the below formula was used.

The KR-20 is given as:

$$\rho_{KR20}=K/(K-1)\ (1-\Sigma_{\rho iqi}/\sigma^2)$$

Where;

K= number of questions

ρi= number of people in the sample who answered question correctly

q_i= variance for each of the item

σ^2 =Variance of the entire scores.

Σ= indicates to sum

Table 8 Estimating Reliability Coefficient (ERC) Using KR-20

Question No	Total number of students who answered the item correctly (R)	Proportion correct (item difficulty) ρ=R/T	Variance of each of the items q=1-ρ	ρq
#1	18	0.60	0.40	0.24
#2	24	0.80	0.20	0.16
#3	15	0.50	0.50	0.25
#4	8	0.30	0.70	0.21
#5	12	0.40	0.60	0.24
#6	23	0.80	0.20	0.16
#7	10	0.30	0.70	0.21
#8	23	0.80	0.20	0.16
#9	30	1.00	0.00	0.00
#10	24	0.80	0.20	0.16
#11	27	0.90	0.10	0.09
#12	12	0.40	0.60	0.24
#13	10	0.30	0.70	0.21
#14	15	0.50	0.50	0.25
#15	3	0.10	0.90	0.09
#16	19	0.60	0.40	0.24
#17	27	0.90	0.10	0.09
#18	7	0.40	0.60	0.24
#19	10	0.30	0.70	0.21
#20	27	0.90	0.10	0.09
#21	25	0.80	0.20	0.16
#22	24	0.80	0.20	0.16
#23	20	0.70	0.30	0.21
#24	16	0.50	0.50	0.25
#25	24	0.80	0.20	0.16
#26	25	0.80	0.20	0.16
#27	12	0.40	0.60	0.24
#28	26	0.90	0.10	0.09
#29	9	0.30	0.70	0.21
#30	27	0.90	0.10	0.09
			σ^2 =15.909	∑ρq=5.27

Note. Entries are scores of Item Difficulty and Item Discrimination on n=30

From the computation in Table 8, the obtained figures are substituted into the formula

$$\rho_{KR20}=K/(K-1)\ (1-\Sigma\rho iqi/\sigma 2)$$

$$\rho_{KR20}=30/(30-1)\ (1-5.27/15.909)$$

$$\rho_{KR20}=30/29\ (1-0.3312)$$

$$\rho_{KR20}=1.034\ (0.6688)$$

$$\rho_{KR20} = 0.6915$$

$$\rho_{KR20} = 0.70$$

In estimating the reliability coefficient, the Kuder-Richardson reliability coefficients was used. The Kuder-Richardson reliability was deemed appropriate because the items were dichotomously scored either correct or wrong. Specifically, KR-20 was employed. The KR-20 was again used because the items differed in difficulty level. The reliability estimate obtained was 0.6915 which is approximately

0.7 which means that reliability of the test was high since it was more than 0.6 (r> 0.6).

Discussion

The paper is discussed in the context of KR-20 have been used to estimate reliability of text items. In the work of Wombacher (2018), it is opined that the values for the KR-20 can range from 0.00 to 1.00,where the author asserted that higher values indicate a higher level of internal consistency. Scores from .70 and above are often considered to be acceptable; however, scores above .80 are typically preferable. Scores above .90 indicate excellent consistency. According to the estimation of KR-20, any scores below .70 indicate that the measure has poor internal consistency and that the test should not be used for further decisions and placements of students. Consequently, if the measure falls below .70, the teacher may wish to perform a factor analysis to learn more about potential issues in the measure.

From our study, we could infer that item difficulty lends a hand in distinguishing easy item from difficult ones. By and large, we can settle that there was a good distribution of difficulty throughout the test conducted for the students. The results from the current study lend support to the study of Mitra et. al (2019) who reportedsimilarly that 40% of the multiple-choice questions of pre-clinical semester 1 multidisciplinary summative tests had the difficulty level over 0.8. Similarly, Sim,et al.,(2016) in a study analysing year two examinations of a medical school found that 40% of the multiple-choice question (MCQ) surpassed the difficulty level of 0.7. 20% of the items with difficulty level of 0.2 and over were classified as easy items with three questions acquires difficulty index of 1.0 and only 2% were determined to be difficult questions.

The results from the present study further placed in the context ofSabri (2013) findings. Specifically,the findings of Sabri (2013) indicated that forty four percent of the total test items exceed the difficulty index of 0.8 suggesting easy items. Fifty nine percent (59%) of items obtained acceptable range of discrimination index. In the work of Sabri (2013), the distractor analysis reveals that some distractors were not effective. The quality of the item indicates a reliable value Kuder-Richardson 20 (KR20) value of 0.717 and Kuder-Richardson 21(KR21) value of 0.703.

The obtained results from the present study are similar to those of a study conducted by El-Uri and Malas (2013) who analyse undergraduate examination in obstetrics and gynaecology. The study reported that 38% of the test items had the discrimination coefficient less than 0.2 with 23 questions obtained negative discrimination. This implied that items with poor and negative discrimination coefficient should be highlighted for reviewing purpose. A poor discriminating power might signify confusing items which were ambiguously worded or indicates a mis keyed item. Ultimately, our study asserted that items with negative coefficient should be removed from the comprehensive test.

Similarly, Adeleke and Joshua (2015) coincide in the reasoning of the negative value in item analysis. In their study, theyaverred that student in the low achievement group often make a guess in answering the easy question and by chance come up with the correct answer. Contradictory, students in the upper achievement group embark upon the easy question too vigilantly and end up choosing the wrong answer. Items with negative discrimination coefficient should be eliminated from the test as put forward by El-Uri and Malas(2013). The reason is that item with negative discrimination coefficient indicates students with low score got the item right and students with high score answer the item incorrectly.

Corroborating with further empirical evidence, our results leans on a classical book of Boyle and Radocy(1987)which highlighted the importance of conducting item analysis. The authors advocated that item analysis facilitates test developer in the process of test quality enhancement which typically involves assessment cycles of preserve, eliminate, improve or endorse particular item. Problematic items specifically items with ambiguous wording and wrongly keyed be reviewed based on the calculated difficulty index and discrimination coefficient values to improve the quality of the test. To this end, Boyle and Radocy advocated that in constructing test items, content expert should be consulted to improve items identified as problematic in terms of language and content appropriateness.

Limitations of Using KR-20

Admittedly, there are several limitations of the KR-20. One issue with the KR-20 is that it can only analyze dichotomous variables. Cronbach's alpha, another test of internal reliability, is able to analyze both dichotomous and continuous variables, which can be seen as an advantage. Another potential issue with the KR-20 is that they do not allow for awarding partial credit. Some question formats, like true/false, do not lend themselves to partial credit, but others such as fill-in-the-blank can be much more difficult to score in a dichotomous way.

This is especially problematic if a teacher is attempting to assess learning, as being able to partially recall the information would indicate more learning than being able to recall none of the information. However, the KR-20 would score both results in the same way, which would make the items a less valid measure of learning than if they were able to award partial credit. Another issue with the KR-20 is that they only assess reliability at a single point in time. The KR-20 look at a single instance of the measure and do not compare how someone responded to an item at two different times to see if their response has changed. Many other tests of reliability only require a single instance, but some scientists prefer a testretest method as it allows you to compare how a person answered a question on two separate occasions to see if the person answered consistently each time.

Conclusions and Recommendations

In achieving or determining the reliability of test items in the teacher made test, Kuder-Richardson Formula 20 could be needful and helpful toassessors in higher institutions. Amidst its limitations, it is worth noting that for assessorsin higher institutionsto understand that Kuder-Richardson Formula 20 is one of the powerful tools for estimating the magnitude in assessing the reliability of measurements for specific test items that are scored dichotomously. The suggested estimates are all based on consistent statistics, so teacher should satisfactorily useKR-20 in large samples in their classroom to determine of estimate reliability of their test items. To this end, the researchers believe that the suggested procedure for estimating reliability of test items may impart some efficiency into the process, and the computation and results in Table 6 support such a notion.

Nonetheless, the derivation of formal estimation procedures would be useful for assessorsin higher institutionsin their quest to estimate the psychometric properties of test items. What we have described here is the efficient use of KR-20 to circumvents the unrealistic assumption that every test item in the classroom is automatically reliable and as such do not need any proof of an index of coefficient. In the main, we conclude that item analysis alleviates test developer in developing an ideal achievement test which functions as tools to evaluate learners' progress and instructional quality in tertiary institutions. Hence, estimating the psychometric properties of test items using Kuder-Richardson Formula (KR-20) should be given the needed attention and priority among assessors in higher institutions in Ghana and beyond.

Abbreviations

KR-20: Kuder-Richardson Formula; WAEC: West Africa Examination Council; IAR: Instructional Assessment Resources; ERC: Estimating Reliability Coefficient; SPSS: Statistical Package for Social Science.

Data Availability

The data (primary) used to support the findings of this study are available from the corresponding authors upon reasonable request.

Declaration

Conflicts of Interest

No conflict of interest exists in the study. We wish to state categorically that there are no known conflicts of interest associated with this publication, and there has been no any financial support for this work that could have influenced the results.

Ethics Approval and Consent to Participate

Not applicable

Funding

No funding was received for this study

Acknowledgements

We would like to thank our anonymous reviewers for their intellectual inspiration and constructive criticisms throughout the development of the manuscript. We also express our profound gratitude to our research assistants for assistance offered during data administration and scoring of the test.

References

Adeleke, A. A., & Joshua, E. O. (2015). Development and validation of scientific literacy achievement test to assess senior secondary school students' literacy acquisition in Physics. *Journal of Education and Practice,* 6(7), 28-42.

Andrich, D. (2014). An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. *Education Research and Perspectives,* 9(1), 95-104.

Alexander, P. A., & Brown, G. T. (2017). *Assessment of student achievement*. Routledge.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. White Plains, NY: Longman.

Boyle, J.D., & Radocy, R.E. (1987). *Measurement and evaluation of musical experiences*. New York: Macmillan.

El-Uri, F. I., & Malas, N. (2013). Analysis of use of a single best answer format in an undergraduate medical examination. *Qatar Medical Journal*, 3(1), 1-12.

Frey, L., Botan, C., & Kreps, G. (2000). *Investigating communication: An Introduction to Research Methods*. Boston, MA: Allyn and Bacon.

Ginns, P., & Barrie, S. (2014). Reliability of single-item ratings of quality in higher education: A replication. *Psychological Reports*, *95*(3), 1023-1030.

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence Based Nursing*, 18(4), 66-67.

Iacobucci, D., & Duhachek, A. (2013). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology,* 13(4), 478-487.

Impara, J. C., & Plake, B. S. (2012). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement,* 35(1), 69-81.

Instructional Assessment Resources. (2011). *Item Analysis*.

Jackson, T. R., Draugalis, J. R., Slack, M. K., Zachry, W. M., & D'Agostino, J. (2012). Validation of authentic performance assessment: A process suited for Rasch modeling. *American Journal of Pharmaceutical Education,* 66(3), 233-242.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika,* 2(3), 151–160.

Le, L.T. (2012). *Item Point-biserial Discrimination.* Australian Council of Educational Research.

Matlock-Hetzel, S. (2017). Basic concepts in item and test analysis*. Annual Meeting of the Southwest Educational Research Association*.

Meijer, R. R., Sijtsma, K., & Molenaar, I. W. (2013). Reliability estimation for single dichotomous items based on Mokken's IRT model. *Applied Psychological Measurement, 19*(4), 323-335.

Merrigan, G., & Huston, C. L., (2019). *Communication Research Methods*. Oxford, UK: Oxford University Press.

Mitra, N. K., Nagaraja, H. S., Ponnudurai, G., & Judson, J. P. (2019). The levels of difficulty and discrimination indices in type A multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. *IeJSME,* 3(1), 2-7.

Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University Economic Series,* 17(4), 59-82.

Nitko, A. J. (1996). *Educational assessment of students*. Englewood, NJ: Prentice Hall.

Petters, J. S., Asuquo, P. N. & Eyo, M. (2015). Psychosocial variables in occupational aspirations of secondary school students in Calabar, Nigeria. *Advances in Social Sciences Research Journal,* 2(7), 89-94.

Platukus, G. L. (2020). *The Relationship between Critical Thinking and Information Literacy* *in Community College Students: A Mixed Methods Study*. Drexel University.

Sabri, S. (2013). Item analysis of student comprehensive test for research in teaching beginner string ensemble using model-based teaching among music students in public universities. *International Journal of Education and Research,* 1(12), 1-14.

Saupe, J. L. (2017). Some useful estimates of the Kuder-Richardson Formula Number 20 reliability coefficient. *Educational and Psychological Measurement,* 21(1), 63-71.

Sim, S. M., & Rasiah, R. I. (2016). Relationship between item difficulty and discrimination indices in true/false type multiple choice questions of a Para-clinical Multidisciplinary Paper. *Annals Academy of Medicine, Singapore,* 3(5), 67-71.

Tan, S. (2019). Misuses of KR-20 and Cronbach's alpha reliability coefficients. *Education and Science,* 34(152).

Wallen, N. E., & Fraenkel, J. R. (2013). *Educational research: A guide to the process*. Routledge.

Wanous, J. P., & Reichers, A. E. (2016). Estimating the reliability of a single-item measure. *Psychological Reports,* 78(2), 631-634.

Whitney, D. R., & Sabers, D. (2014). *Improving essay examinations III: Use of item analysis, Iowa City: University Evaluation and Examination Service*. The University of Iowa.

Wombacher, K. (2018). Reliability, Kuder-Richardson Formula. In M. Allen, (Ed.), *The SAGE Encyclopedia of Communication Research Methods*. Sage Publications.

Author Details

Simon Ntumi, *University of Education, Ghana,* ***Email ID:*** *sntumi@uew.edu.gh*

Sheilla Agbenyo, *Bia LamplighterCollege of Education, Ghana,* ***Email ID:*** *sheilla.agbenyo001@stu.ucc.gh*

Tapela Bulala, *Botswana University of Agriculture and Natural Resources (BUAN), Botswana,*
Email ID: *tabulala@buan.ac.bw*