JOURNAL OF INCLUSIVE & INNOVATIVE PRACTICES (JOIIP)
ONLINE ISSN: 3005-9976

Construction and Validation of the Basic Operations Achievement Test (BOAT): An Arithmetic Achievement Test for Learners in Zimbabwe

[1] Mushangwe Edward T, and [2] Zirima Herbert
Great Zimbabwe University

ABSTRACT

This study sought to develop an Arithmetic Achievement test through a scientific process. It aimed to develop a good quality test which could be used in the Zimbabwean education system as a replacement for the Western tests which are culture-biased, with irrelevant content included. To achieve this goal, a blueprint of the scale was developed based on pertinent literature of achievement test development, with a view to measuring students' success in learning natural numbers and operations. As a developmental quantitative descriptive study carried out in four Manicaland Districts, namely Chimanimani, Chipinge, Mutare and Mutasa, it analyzed the items of the test that were tried out on 250 pupils attempting 30 open-ended questions included in the test. Their answers were analyzed in terms of validity, item difficulty, and reliability to some extent. Firstly, the validity analysis showed that all items were considered valid, with a content validity index of 0.75 considered acceptable since it falls slightly off 0.8 that is considered excellent. Then the difficulty index analysis of 40 pool test items showed that more than a half of the items were categorised as moderate but the rest were either very easy or very difficult. The average pass rate from the nine piloted schools was 40%, with insignificant score variance between boys and girls. Finally, the reliability estimation analysis showed that the consistency index of the test already met the minimum reliability index required. Based on the findings, the test developer is recommended to cascade the achievement test to other districts and provinces, with standardization of the developed test in mind.

Key Words: Achievement, Development, Validation, Reliability, Correlation

© 2025 This is an open access article distributed under the terms of the Creative Commons Attribution Licence (https:// creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution and production in any medium, provided the original work is properly cited.

INTRODUCTION

According to Kubiszyn and Borich (2024), achievement testing and evaluation can be used for the purposes of diagnosis and formalisation as well as identifying one's level, while it has been one of the leading ways to identify students' level of achievement. Psycho-educational testing, through administering achievement tests, in Zimbabwe can be understood in terms of the colonial heritage of the country and the democratisation of the education system following the country's independence in 1980 (Chimbunde & Moreeng, 2024). Regulations for the psychological practice and Western-oriented psychology curricula inherited from the colonial era are elements of continuity in the country's psycho-educational testing practices (Machingura & Kalizi, 2024). Fernandez (2019) reports that the demands of the post-independence education dispensation that extended psychological services to the native Blacks represent elements of change that challenge psychological and educational testing for the country to be more responsive to cultural diversity. A survey of test users in the educational settings in the country revealed a significant use of Western tests, although some limited progress has been made in developing local tests, particularly in the area of achievement testing. According to Oppong (2024), lack of documentation of African practices has resulted in the dominance of Western ideas through the importation of these assessment tools and models.

According to Mpofu and Nyanungo (2008), the current status of psycho-educational testing in Zimbabwe is best characterised as being in a transitional or modelling stage, entailing the application of Western concepts and technologies. Authentic testing has great potential as an alternative in this developing country, hence the need to develop local achievement standardised tests. Achievement tests are of great importance to ensure that suitable candidates are selected for education beyond primary level. According to Tho et al. (2024), major investigations directed to this need have been in the area of adapting the tests already available in the Western countries and determining the reliability and predictive validity of a battery of tests so adapted and their tryout in African schools, to determine their efficiency as tools for selection.

However, it has been extensively documented that the use of psychometric tests in a culture other than the one in which the tests were developed requires evidence of validity and reliability in the new setting, and,

CC BY

usually the development of new cultural norms must take precedence (Cakir & Ari, 2022). In Zimbabwe, the continued use of traditional outdated achievement tests designed in Britain or America, with minor or no modifications at all, is widespread (Nkoma & Kufakunesu, 2024). Zindi's 1994, research results, in his attempt to adapt WISC – R to the Zimbabwean context, were not conclusive (Mpofu et al., 1997). This study has demonstrated the possibility of achieving such an objective and even constructing a Zimbabwean achievement test.

OBJECTIVES

1. To draft quality test items to be used in the development of an achievement test for primary school pupils in Zimbabwe.

2. To pre-test the test to a small set of respondents from the population for the full scale survey.

3. To administer the test to a larger sample so as to ensure standardization.

STATEMENT OF THE PROBLEM

Achievement tests used in Zimbabwe are Eurocentric and the concept of content of the test is problematic for Zimbabweans, who live differently and have different forms of intelligence. Psychological tests are always linked to the context in which they were designed. The Western achievement tests were designed in a specific society and culture and for a specific purpose. The way in which the test performance would be interpreted would be linked to the behavioural criteria, norms or cut-scores developed in the context where this test was developed.

METHODOLOGY

In this study, the quantitative method was applied, as the researcher sought to find the correlation coefficient of the pre-test and post-test scorers. According to Clark-Carter (2024), quantitative research is a formal, objective, systematic process for obtaining quantifiable information about the world which is presented in numerical form, and analysed through the use of statistics. The study design was Quasi-Experimental. Quasi-Experimental design is a research method that seeks to evaluate the causal relationships between variables, but without the full control over the independent variable(s) that is available in a true experimental design (Mamolo, 2021). Pre-test and Post-test as a subtype of Quasi-Experimental design, administered to both the pilot or small group sample and large representative sample, was identified as the

CC BY

ideal research design since this design involves measuring the dependent variable(s) before and after an intervention or event, but without a control group.

Before deploying a large-scale assessment, pilot testing was administered to a representative sample of 50 participants to gather feedback on item difficulty, clarity, and relevance. To select the large representative sample of 200 participants, the researchers got a list of urban and rural primary schools from the four districts, with enrollments of over 500 learners for rural schools and over 1000 learners for urban schools. The researchers then used the simple random sampling technique via balloting to draw the nine primary schools. The selected schools were informed of their impending participation in the research. Permission was obtained from the relevant Ministry and local school authorities. Before data were collected, arrangements were made with each school head of the eight selected schools with regard to the preliminary visits, as well as actual days and time for administering the test. According to Osadebe (2001), contacting respondents before the research paves the way for carrying out an effective research. The test was administered to the selected pupils by the researcher, who administered and collected the response sheets soon after the children had finished writing. Marking and scoring then followed, with statistical computation of data being the last stage in determining the correlation coefficient of the pre- and post-test.

In administering the test, the amount of **time** allowed **between** measures is critical. The shorter the **time gap**, the higher the correlation; the longer the **time gap**, the lower the correlation. If the **time interval** is short, people may be overly consistent because they remember some of the questions and their responses (Jayanthi, 2014). The researchers had to settle for a short interval, 3 days between the pre-test and post test administration. The goal was to minimise test effect, where the act of taking pre-test impacts the post-test score. A short interval was used to avoid potential maturation effects like age and experience.

POPULATION AND SAMPLING

The study had a sample of 250 Grade 4 pupils, with both sexes equally represented. According to Kyriazos (2018), the group size for data collection should be at least twice the number of items to be measured. Having a study group of 1,000 individuals is considered excellent, whilst 500 is good, 300

CC BY

average, and 100 individuals in a group is seen as inadequate (Comrey & Lee, 2013). Based on this criterion, the current study's sample of 250 students was deemed appropriate for analysis. The selection utilised a probability random sampling, where the targets were randomly picked from a class teacher's mark schedule, with interest on the below average, average and the above average pupils.

DATA COLLECTION INSTRUMENTS

The data were collected through administration of an achievement test called Basic Operations Achievement Test (BOAT), for primary school pupils, developed by the researcher. The basic goal was to test pupils' ability to independently compute Mathematical problems, differentiate Math operations symbols, and be conscious of place values when arranging their work. The test scores are used to inform errors peculiar to each individual for early interventions.

Procedure for test development

In developing the test, the following steps, as adopted from Osadebe (2016), guided the researchers:

Planning the test

Planning involves identifying objectives, resources and processes for the research. Objectives must be clear and measurable. These objectives guide the design of the test.

Constructing the test

This involves deciding on the structure of the test, the type of questions to include, and how to best assess the content covered. The design should be reflective of the nature of the subject being taught. With the blueprint in hand, it will be time to start writing the actual test items. This step involves crafting clear, concise, and well-structured questions that align with the objectives. The test items should be designed to assess both knowledge and skills, according to the different levels of cognitive complexity.

Initial validation of the test

Validity refers to **how accurately a method measures what it is intended to measure.** If research has high validity that means it produces results that correspond with real properties, characteristics, and variations in the physical or social world.

Small- group testing (Pilot testing)

CC BY

Administering the test to a small group helps identify issues with instructions, timing and item clarity before wider use. This is important for identifying any problems that could affect test validity.

Item analysis

This is a process which examines student responses to individual test items (questions) in order to assess the quality of those items and of the test as a whole. Item analysis is used to eliminate ambiguous or misleading items in a single test administration. In addition, item analysis is valuable for increasing instructors' skills in test construction and identifying specific areas of course content which need greater emphasis or clarity.

Selection of good items

After the test is administered item analysis is carried out. This involves examining the difficulty and discrimination of individual test items. Based on this analysis, items that are too easy, too difficult, or do not discriminate well between high and low achievers are modified or removed.

Reliability

Reliability refers to the consistency of the measurement. Reliability shows how trustworthy is the score of the test is. If the collected data show the same results after being tested using various methods and sample groups, the information is reliable. If one's method has reliability, the results will be valid.

FINDINGS AND DISCUSSION

The findings and discussion are based on the research objectives, which include drafting of quality test items to test primary school pupils' achievement. Table 1 is the actual sample of the developed test with all problems uniquely and horizontally arranged to measure the "right" psychological constructs. Intelligence, self-esteem and creativity are examples of such psychological traits. Results of the data analysis are shown in Table 2, where BOAT reliability coefficient is evidently high due to high estimates of internal consistency which account for error due to content sampling. The discussion concludes with a comparison between the developed test and Wide Range Achievement Test (WRAT) which is outdated but is still widely used by Educational Psychologists in Zimbabwe.

CC BY

Table 1: Sample of the developed BOAT test **for** *Grade 1 to Grade 7 pupils*

O O O O O O						
O O O O O O						
O O O O O O O						
8	13	9	1	32	41	14
7	10	11	33	27	15	31
23	4	2	40	26	0	
+			-			
x			÷			
=			≠			
2 + 1 =		5 + 6 =		7 + 7 =		
4 – 2 =		7 – 5 =		11 – 8 =		
6 ÷ 2 =		8 ÷ 3 =		17 ÷ 5 =		
4 x 2 =		3 x 3 =		4 x 7 =		
46 + 68 =				124 + 13 =		
478 + 144 =						
301 – 78 =				456 - 121 =		
512 – 200 =						
75 ÷ 5 =				133 ÷ 2 =		
465 ÷ 7 =						
46 x 5 =				38 x 12 =		
78 x 15 =						
18. 6 x 9 =				33.4 x 0.5 =		
23.3 x 1. 2 =						
256 ÷13 =				766 ÷11 =		
1259 ÷ 12 =						

Grade Equivalents Scoring Scale:

Score	1 – 4	5 – 8	9 – 12	13 – 16	17 – 20	21 – 24	25 - 28	29 – 30
GE	1	2	3	4	5	6	7	7 +
Count & Read			Score 0 - 23 ECD A			Score 24 – 46 ECD B		

Content validity index (CVI) of BOAT:

The content validity index (CVI) of Basic Operations Achievement Test (BOAT) was computed based on the joint ratings of relevance of BOAT items by two content specialists. The table below shows joint ratings of the relevance of BOAT items by two content specialists. These ratings are for 40 test items with a relevance of 30.

Table 2: Showing CVI analysis

Specialist 1

		Item rated 1&2	Item rated 3&4	Total
Specialist 2	Item rated 1&2	(a) 3	(b) 12	a + b =15

CC BY

	Item rated 3&4	(c) 7	(d) 18	c + d = 25
	Total	a + c = 10	b + d = 30	a + b+ c + d =40

This analysis was carried out using a 4-point rating scale where:

4 stands for 'very relevant', 3 stands for 'quite relevant', 2 stands for 'somewhat relevant', and 1 stands for 'not relevant'.

i. Cell 'a' indicates the number of items rated 1 & 2 by the first and second content specialists.
ii. Cell 'b' indicates the number of items rated 3 & 4 by first content specialist and then 1 & 2 by the second content specialist.
iii. Cell 'c' indicates the number of items rated 1 & 2 by the first content specialist and then 3 & 4 by the second specialist.
iv. Cell'd' indicates the number of items rated 3 & 4 by both content specialists.

Thus, CVI= $b+d$ / $a+b+c+d$ = 40 / 4+8+6+32 = 30 / 40 = 0.75. This implies that 75% of test items which are equivalent to 30 items out of 40 were rated quite relevant and very relevant to the component objectives.

The reliability coefficient of BOAT:

To certify reliability, the test-retest method was used for the computation of the data, using the three statistics Po, k and Pc. Po determines the degree of agreement of decision made on two administrations of a test, k measures degree of agreement uncontaminated by chance, while Pc measures the proportion of individuals to have consistent classification. These were computed using data obtained from the 250 pupils in BOAT, based on their score.

Table 2: Showing CVI analysis

				Po	K	Pc
	Mastery	Non – Mastery	Total	0.63	0.36	0.40
Test 1	105	145	250			
Test 2	109	141	250			

Reliability

This study sought to construct a reliable achievement test. BOAT is sufficiently reliable to permit stable estimates

CC BY

of the ability levels of individuals in the target group compared to Wide Range Achievement Test (WRAT), whose content includes units of measurement like inches which are not in the syllabi. Fundamental to the evaluation of any instrument is the degree to which test scores are free from measurement error and are consistent from one occasion to another when the test is used with the target group, as BOAT has proven. Sources of measurement error, which include fatigue, nervousness, content sampling, answering mistakes, misinterpreting instructions and guessing are high when using WRAT because many of the number stories included, which may be disadvantageous to a pupil who has dyslexia, and the use of unfamiliar division sign symbols, all contribute to an individual's score and lower a test's reliability. BOAT has proven to have higher estimates of internal consistency, which account for error due to content sampling, usually the largest single component of measurement error compared to WRAT. The focus of BOAT is the ability to compute operations, whilst WRAT focuses on what the child should have learnt at a particular grade. The quality of the evidence is a critical factor in making sensible decisions after consulting theory, as evidenced by the researchers' acceptable variance of the pre- and post-test scores across all the schools tested. The average pass rate was 40%. Although the pass mark was 15, those who scored 13 and 14 retained their current Grade 4 placement. The reliability estimation analysis showed that the consistency index of the test already met the minimum reliability index required.

Item difficulty index

For the Difficulty Index, the formula is given as: D = $U + L \ / \ N\ U + N\ L$

Where D = difficulty index

U=Number in group meeting criterion who answered correctly

L= Number of group not meeting criterion but answered correctly

Nu= Number in group who met criterion

NL = Number in group not meeting criterion.

Table 3: Item difficulty index key

Item difficulty range	Level of difficulty
0.0 - 0.19	Very difficulty
0.20 - 0.39	Difficult
0.40 - 0.60	Average /moderately difficult
0.61- 0.79	Easy
0.80 – 1.0	Very easy

CC BY

Table 4: Item difficulty index table

Item	Difficulty Index	Difficulty level	Follow up Action
1	0.85	Very easy	Discarded
2	0.89	Very easy	Discarded
3	0.60	Average	Accepted
4	0.39	Difficult	Accepted
5	0.50	Average	Accepted
6	0.55	Average	Accepted
7	0.65	Easy	Accepted
8	0.43	Average	Accepted
9	0.55	Average	Accepted
10	0.66	Easy	Accepted
11	0.58	Average	Accepted
12	0.54	Average	Accepted
13	0.49	Average	Accepted
14	0.38	Difficult	Accepted
15	0.44	Average	Accepted
16	0.25	Difficult	Accepted
17	0.29	Difficult	Accepted
18	0.57	Average	Accepted
19	0.60	Average	Accepted
20	0.42	Average	Accepted
21	0.44	Average	Accepted
22	0.57	Average	Accepted
23	0.15	Very difficult	Discarded
24	0.29	Difficult	Accepted
25	0.44	Average	Accepted
26	0.58	Average	Accepted
27	0.27	Difficult	Accepted
28	0.18	Very difficult	Discarded
29	0.58	Average	Accepted
30	0.58	Average	Accepted
31	0.41	Average	Accepted
32	0.40	Average	Accepted
33	0.19	Very difficult	Discarded
34	0.22	Difficult	Accepted
35	0.15	Very difficult	Discarded
36	0.28	Difficult	Accepted
37	0.13	Very difficult	Discarded
38	0.12	Very difficult	Discarded
39	0.13	Very difficult	Discarded
40	0.10	Very difficult	Discarded

From Table 4, the value of difficulty index lies between 0.10 and 0.89. Item number 2 has the largest value (0.89) and item number 40 has the smallest value (0.10). Item difficulty indices relate to the percentage or proportion of students answering each item correctly. The difficulty index shows prepared items had an average difficulty. Only two items were found to be easy. The

CC BY

indices were used in arranging BOAT items according to their increasing order of difficulty.

CONCLUSION

The Basic Operations Achievement Test is a reliable evaluation instrument. BOAT was found to be highly reliable with three statistics Po, Pc and k. The computation shows that Po = 0.63, Pc = 0.36, and k = 0.40. The study revealed that the test is generally within the ability of the test takers. The test takers have a high probability of correctly responding to the majority of the items in the test. The study also concluded that the test can measure the objectives of the revised basic education curriculum, as the test is in line with the ability level of the students in the appropriate level of schooling.

REFERENCES

Çakır, H., & Arı, A. G. (2022). Development and validation of an achievement test in Biology. *Journal of Social Sciences and Education. Eğitim Dergisi, 5(1), 64–75. https://doi.org/10.53047/josse.1102697*

Chimbunde, P., & Moreeng, B. B. (2024). Post-colonial educational reforms in Zimbabwe: A Fake badge of decolonization of the curriculum. *Power and Education*, 17577438241257662

Clark-Carter, D. (2024). *Quantitative psychological research: The complete student's companion*. Routledge.

Comrey, A. L., & Lee, H. B. (2013). *A first course in factor analysis*. Psychology Press.

Fernandez, A. L. (2019). Modern neuropsychological tests for a diversity of cultural Contexts. *The Clinical europsychologist*, 438-445),

Jayanthi, J. (2014). Development and validation of an achievement test in mathematics. *International Journal of Mathematics and Statistics invention*, 2(4), 40- 46.

Kubiszyn, T., & Borich, G. D. (2024). *Educational testing and measurement*. John Wiley & Sons.

Kyriazos, T. A. (2018). Applied psychometrics: ample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology*, 9(08), 2207.

Machingura, F., & Kalizi, C. S. (2024). Christian Education in Colonial and Post-Independent Zimbabwe: A

CC BY

Paradigm Shift. *Religions*, 15(2), 213.

Mamolo, L. A. (2021). Development of an Achievement Test to Measure Students' Competency in General Mathematics. *Anatolian Journal of Education,* 6(1), 79-90. https://doi.org/10.29333/aje.2021.616a

Mpofu, E. & Nyanungo, K. R. L. (2008). Educational and Psychological Testing in Zimbabwean Schools: Past, Present and Future. *https://doi.org/10.1027/1015 5759.14.1.71*

Mpofu, E., Zindi, F., Oakland, T., & Peresuh, M. (1997). School Psychology practices in East and Southern Africa: Special Educators' Perspectives. *The journal of Special Education*, 31(3), 387-402.

Nkoma, E., & Kufakunesu, M. (2024). Provision of educational psychological services under a High inflationary environment in Masvingo Province, Zimbabwe. *School Psychology International*, 45(6), 659-680.

Oppong, S. (2024). *Indigenous psychology in Africa: a survey of concepts, theory, research, and Praxis*. Cambridge University Press.

Osadebe, P. U. (2001). Construction and Validation of Test. *A Seminar Paper Presented at the University Of Port Harcourt.*

Osadebe, P. U. (2016). Practical guide to item generation. *A seminar presented at Delta State University, Abrak*

Tho, C. C., Anh, N. T. Q., An, V. T., Ha, L. T., Huong, D. T. T., Cuong, D. X., & Dung, N. V. (2024). *Adaptation of 6th grade*. University of Port Harcourt.

Zindi, F. (1994). Towards the standardization of the WlSC-R for early childhood assessment in Zimbabwe. *IFE Psychologia: An International Journal*, 2(2), 19-32.