The Modus Operandi of National Benchmark Test Project in South Africa: A Systematic Review

Ayanwale Musa Adekunle
University of Johannesburg

Ndlovu Mdutshekelwa
University of Johannesburg

Ramdhany Viren
University of Johannesburg

National Benchmark Test Project (NBTP) was conceived as far back as 2004 by the public higher education leadership in South Africa and formally launched in 2005. NBTP was initiated to provide additional information to higher education in South African institutions in the admission and placement of students, establish the relationship between entry-level proficiencies and school-level exit, and inform the nature of foundation courses and curriculum responsiveness. The NBTP assessed the academic proficiency of candidates in three core domains. This test is criterion-referenced, providing additional information to complement the norm-referenced National Senior Certificate (NSC) results. About 22 studies dealing with various NBT titles are examined in a systematic review from multiple databases. There is no doubt that the end-product of NBTP will be an invaluable addition to the available current tools used for programme placement at higher education institutions (HEIs) in South Africa. In this fourth industrial revolution (4IR) era, we also recommend NBTP to include Digital Literacy (DL) in the existing core domains.

Keywords: digital literacy, education 4.0, mathematics, National Benchmark Test Project, proficiency bands, quantitative literacy, standardised assessment, standard setting, testing

INTRODUCTION

The national benchmark test project was conceived as far back as 2004 by the public higher education leadership. NBTP was initiated to assess the relationship between entry-level proficiencies and school-level exit outcomes (Griesel, 2006, p. 4). The NBTs' conceptualisation, design, and implementation make it well placed to provide information for placement and curriculum development (Prince, 2016; Prince, 2017, p.134) and build a responsive enrolment system for the South African institutions (O'Connell, 2006). In 2005, the NBTP was formally launched by the Higher Education South Africa (HESA), now called Universities South Africa (USAf), to assess all prospective students' academic competency and proficiencies in three core areas: academic literacy, quantitative literacy, and mathematics, respectively (Parliamentary Monitoring Group, 2009). A test provides a sample of behaviour or a content domain (

Foxcroft & Roodt, 2005). From this test, educational assessment experts made inferences regarding the level of performance of an individual or a group. A test is often administered under standardised (controlled) conditions, and systematic procedures are used to score and interpret test performance. Thus, the NBT's are criterion-referenced tests, which is different from National Senior Certificate (NSC) examination that is norm-referenced. Although, the NBT's still provide a complementary mechanism to school leaving certificates to help select and place students into apt courses and programmes in higher education and provide information that will contribute significantly to curriculum development (le Roux & Sebolai, 2017; Prince, 2017).

It is noteworthy to stress that those students write mathematics proficiency of NBT with mathematical related entering disciplines (such as engineering, mathematics, statistics, etc.). According to Frith and Prince (2018), the mathematics test is designed to assess students' abilities to do the mathematics learned at (K-12) school and transfer it to higher education first-year mathematics courses. The quantitative literacy test is entrenched in the NSC Mathematics curriculum but aligned with first-year mainstream courses in higher education. Academic literacy test items illuminate the challenges of first-year students' ability to read standard literary texts they might encounter in their studies. Generally, NBTP doesn't have any curriculum or syllabus that guides the development of test items; assessment experts from various institutions and high schools in academic literacy, quantitative literacy, and mathematics are nominated by the Centre for Educational Testing for Access and Placement (CETAP) to develop the pool of items for the test (Griesel, 2006).

The establishment of NBTP hinged on four goals, as stated by (Griesel, 2006). These include assessing students' entry-level academic quantitative literacy and mathematics proficiency, assessing the relationship between entry-level proficiencies and school-level exit outcomes, providing a service to successful engagement with the demands of higher education in the admission and placement of students, and informing the nature of foundation courses and curriculum responsiveness. The USAf believed that students' competency assessed in critical areas using the NBTP test should provide minimum proficiency levels (Prince & Frith, 2017). Students' competency is categorised as proficiency score band, intermediate score band, and basic score band. Also, each score band category description and recommendations for the kind of educational provision apt for students whose scores are in that category will be discussed in the subsequent session of the paper. This article will examine a systematic review of the modus Operandi adopted by NBTP, such as the need for benchmarking, NBT in quantitative literacy and mathematics, test specification framework, test development, test administration, test scoring, setting cut-off and validity and reliability, respectively.

To achieve this feat, the scientific search of literature with the keywords "National Benchmark test project" as well as "domains (quantitative literacy and mathematics)" in the title, abstract, and keywords are linked with the occurrence of the keywords "proficiency band and standardised testing" in the title of the papers for our search. A database such as Scopus, Department of Higher Education and Training (DHET), Directory of Open Access Journals (DOAJ), and Scientific Electronic Library Online (SciELO) is used due to its high quality and inclusivity. The search results are a list of 27 papers, out of which three are conference proceedings, 14 journal articles, two reports, two book chapters, and one hearing—the search results of NBT's show a significant increase in scientific papers starting from 2015.

THE NEED FOR NATIONAL BENCHMARK TESTS

The need for benchmarking entry level for prospective students into higher education was founded on three conditions as stated by (Griesel, 2006, p.12). These include (a) Situations where the NSC results are an inadequate reflection of students' potential skills and intellectual ability. It is imperative to develop additional forms of assessment to equitably and precisely select and place students. And to achieve equitability, test development and the interpretation of results must be informed by specific constructs, psychometric qualities (slope/discrimination, threshold/difficulty, and chance factor), and standards, respectively. Also, this condition has translated into the development of a range of assessment protocols in the South African higher education context. (b) For curricula to be responsive to the changing profile of students, higher education needs to have a full grasp of the nature of preparedness and the varying levels of under-preparedness of entry groups. This condition needs an accurate assessment of entry levels to inform institutions' understanding of and response to the nature of entry groups, including the varying levels of preparedness that must responsibly be addressed in first-year curricula and foundation courses. In essence, it's an opportunity to develop the kinds of graduates required by the fourth industrial revolution world. (c) Given the variability of school-leaving results and the reality of a new school curriculum and exit qualification, which is yet to be benchmarked against comparable qualifications. It seems necessary for higher education to set minimum entry cut-offs and assess proficiency levels, at least until the implementation of the new curriculum and National Senior Certificate have stabilised. This assumption is imperative because if the national senior certificate can be used to gauge the levels of proficiency and achievement, higher education may not have needed to develop alternative forms of assessment (Frith & Prince, 2018; Griesel, 2006). And suppose the proposed National Senior Certificate (NSC) had been benchmarked against comparable qualifications. In that case, there indeed may also not have been the need for higher education to develop entry-level benchmarks.

THE NATIONAL BENCHMARK MATHEMATICS TEST

The mathematics test is explicitly designed to probe higher education competencies within the context of the NSC curriculum. Also, the mathematics achievement test (MAT) content is embedded in the NSC Mathematics curriculum but aligned with first-year mainstream needs. It is not the intention of the MAT tests to replicate the NSC (Frith, et al., 2003; Frith, et al., 2004a; Hughes-Halett, 2001). It assumed that if a student achieves a competent pass in the NSC, a certain level of content and procedural competence will be reached when the student writes the first MAT test. The MAT test specification comprises 60 multiple-choice items with four options for each item. The NBT mathematics achievement test is designed to assess candidates' ability to several mathematical topics, such as problem solving and modeling, requiring the use of algebraic processes and understanding and using functions represented in different ways. Basic trigonometry, including graphs of trigonometric functions, problems requiring solving trigonometric equations, and applying trigonometric concepts. Spatial perception (angles, symmetries, measurements, etc.) includes representing and interpreting three-dimensional objects; analytic geometry, data handling, and probability (Placement, 2017a). The fact that mathematics requires learners to integrate many different skills and concepts in each problem implies that individual test items will assess the range of mathematical competencies. For example, an item dealing with a function's graphical representation will also assess spatial and algebraic competence. More so, the MAT subdomains Number sense and Geometric reasoning are associated with the Q.L. subdomains quantity, number and operation, Shape, dimension, and space, but are essentially different, especially in the sense that for Q.L., no specific school curriculum knowledge is required, whereas the MAT subdomains are integrally related to the Curriculum and Assessment Policy Statement (Prince & Frith., 2017; le Roux & Sebolai, 2017; Prince & Simpson, 2016).

THE NATIONAL BENCHMARK QUANTITATIVE LITERACY TEST

Quantitative literacy can manage situations or solve problems in practice and involves responding to quantitative (mathematical and statistical) information presented verbally, graphically, in tabular or symbolic form (Prince & Frith, 2017). The tests assess proficiency in Q.L. for all students, consisting of 50 multiple-choice items with four options for each item. The importance of Q.L. for higher education is widely recognised (Steen, 2004). There is also an increasing awareness that many academic disciplines make complex quantitative demands that are often very different from those focusing on conventional mathematics courses (Prince & Frith, 2017). In developing the Q.L. test, Frith and Prince (2006); Prince and Archer (2008); Prince and Simpson (2016) described the test construct in terms of the contexts in which Q.L. is practiced; the mathematical and statistical content that is required in this practice; and the reasoning and behaviours that are integral to Q.L. practice. Thus, these three dimensions to the test construct (contexts, content, and reasoning/behaviours) are entrenched in the meaning of Q.L. stated earlier.

In practice, the NBTP QL test assesses students' ability to competently interpret and reason with quantitative information presented in various modes. For instance, the construct informing test specifications of Q.L. test as outlined by Frith and Prince (2009); Prince (2017) include that they must understand and use a range of quantitative terms and phrases, read, and interpret tables, graphs, charts, diagrams, and texts and integrate information from different sources. The test also assessed the ability to apply quantitative procedures in various situations, do simple calculations and estimations that may involve multiple steps, and formulate and apply simple formulae. Students are also required to identify trends and patterns in various situations, interpret representations of two-dimensional and three-dimensional structures, and reason logically. The questions are designed to assess Q.L. practices and do not assume that students know any school subject (Frith & Prince, 2018; Nel, 2020). Tables 1-3 presented the Q.L. specifications/description in terms of these three factors, which explain in detail the competencies, mathematical and statistical content, and cognitive level according to (Frith & Prince, 2016).

TABLE 1
QUANTITATIVE LITERACY TEST SPECIFICATION BASED ON COMPETENCIES CLASSIFICATION

Competence area		**Description/specifications**	**Percentage of test**
Comprehending: Vocabulary identifying or locating	Vocabulary	The ability to understand the meanings of commonly encountered "quantitative" terms and phrases (such as percentage increase, rate, approximately, representative sample, compound interest, average, order, rank, category, expression, equation) and the mathematical and statistical concepts (including basic descriptive statistics) that these words refer to. This includes knowledge of systems of units of measurement	15-20%
	Representations of numbers and operations	The ability to understand the conventions for representing (whole numbers, fractions, decimals, percentages, ratios, scientific notation, measurements, variables) and simple operations (+, -, ×, ÷, positive exponentiation, square roots) on them.	5-10%
	Conventions for visual representations	The ability to understand the conventions for the representation of data in tables (several rows and columns and with data of different types combined) and charts (pie, bar, compound bar, stacked bar, "broken" line, scatter plots). Also, graphs and diagrams (tree diagrams, scale and perspective drawings, and other visual representations of spatial entities).	20 - 25 %

Acting, interpreting, and communicating	**Using representations of data**	**The ability to derive and use information from representations of contextualised data and to make meaning from this information. For example, Reading values off a chart or observing trends or relationships in tabulated data, using observations of the slope of a graph to derive information about rates, reading off maximum and minimum values.**	**20 - 25 %**
	Computing	Computing is the ability to identify the necessary simple calculations required by a problem in context and perform the analysis.	15-20%
	Conjecturing	The ability to formulate appropriate questions and conjectures to make sense of quantitative information and recognize conjecture's tentativeness based on insufficient evidence.	0-5%
	Interpreting	The ability to interpret quantitative information (its embedded context) and translate between different representations of the same quantitative information. This interpretation includes synthesising information from more than one source. For example: identifying the correct algebraic formula or graphical representation from a verbal description of a relationship, interpreting the results of a calculation in the original context, deriving and using data from more than one representation to solve a problem.	10-15%
	Reasoning	The ability to identify whether the available evidence supports a claim, formulate conclusions that can be made given specific evidence, or place the evidence necessary to support a claim.	5-10%
	Representing quantitative information	The ability to represent quantitative information verbally, graphically, diagrammatically, and in tabular form using appropriate representational conventions and language. For example: choosing right/correct representations of quantitative data.	5-10%

TABLE 2
QUANTITATIVE LITERACY TEST SPECIFICATION BASED ON MATHEMATICAL AND STATISTICAL CONTENT CLASSIFICATION

Mathematical and Statistical Content	**Description/specifications**	**Percentage of test**
Quantity, number, and operations	The ability to order quantities, calculate and estimate the answers to computations required by a context using whole numbers, fractions, decimals, percentages, ratios, scientific notation, etc., and simple operations (+, -, ×, ÷, positive exponentiation) on them. The ability to express the same decimal number in alternative ways (such as by converting a fraction to a percentage, a common fraction to a decimal fraction, and so on). The ability to interpret the words and phrases used to describe ratios between quantities within a context, convert such phrases to numerical representations, perform calculations with them, and interpret the result in the original context—the ability to work similarly with ratios between quantities represented in tables and charts and scale diagrams.	25-30%
Shape, dimension, and space	The ability to understand the conventions for the measurement and description (representation) of 2- and 3-dimensional objects, angles, and direction. The ability to perform simple calculations involving areas, perimeters, and volumes of simple shapes such as rectangles and cuboids.	10 - 15%
Relationships, pattern, and permutation	The ability to recognize, interpret and represent relationships and patterns in various ways (graphs, tables, words, and symbols.) The ability to manipulate simple algebraic expressions using simple arithmetic operations.	10-15 %
Change and rates	The ability to distinguish between changes (magnitudes) expressed in absolute terms and those described in relative terms (such as percentage change). The ability to quantify and reason about changes or differences. The ability to calculate average rates of change and recognise that the steepness of a graph represents the rate of change of the dependent variable to the independent variable. The ability to interpret the curvature of graphs in terms of rate changes.	10-15 %
Data representation and analysis	The ability to derive and use information from representations of contextualised data in tables (several rows and columns and with data of different types combined), charts (pie, bar, compound bar, stacked bar, "broken" line, scatter plots), graphs, and diagrams (such as tree diagrams) and to interpret the meaning of this information. The ability to represent data in simple tables and charts, such as a bar or line charts.	20-25%

Chance and uncertainty	The ability to appreciate that many phenomena are uncertain and quantify the chance of uncertain events using empirically derived data. This includes understanding the idea of taking a random sample. The ability to represent a probability as a number between 0 and 1, 0 representing impossibility, and 1 expressing certainty.	5-10%

TABLE 3
QUANTITATIVE LITERACY TEST SPECIFICATION BASED ON COGNITIVE LEVEL CLASSIFICATION

Cognitive level	**Description/specifications**	**Percentage of test**
Basic Knowledge	Items functioning at the basic-knowledge level require test-takers to demonstrate mathematical and statistical facts, vocabulary, and simple algorithms. For instance: • Calculate using the basic operations including the operations +, -, ×, and ÷ and appropriate rounding of numbers; • Know and use appropriate vocabulary such as equation, formula, bar graph, pie chart, table of values, diameter, radius, mean, median and mode, maximum, probability. • Know and use simple formulae such as the rectangle area, where the required dimensions are given. • Read information directly from a table, chart, or graph. • Know the conventions for representing numbers and operations such as exponentiation or use of scientific notation.	15-20%
Applying routine procedures in familiar contexts	Items at the applying-simple-procedures-in-context level require test-takers to perform simple procedures in context. The required method is easily identified from the way the problem is posed. All the information needed to solve the problem is immediately available. Little reasoning or interpretation is required. Some examples: • Calculate the price of an article after a given percentage reduction. • Find the ratio between two values read off a chart or table • Use a given scale in a diagram to calculate the dimension of an object represented. • Identify the most appropriate type of graphic representation for a simple set of data.	25-30%
Applying multi-step procedures in a variety of contexts	Items at the applying-multistep-procedures-in-context level require test-takers to perform not immediately apparent operations and involve more than one step. Some reasoning, interpretation, or synthesis will be necessary. Some examples: • Select the appropriate data from a chart or several charts and use it to solve a problem or make comparisons. • Identify and perform calculations involving intermediate steps, estimation, or unit conversion.	25-30%

Reasoning and reflecting	**Items at the reasoning-and-reflecting level require test-takers primarily to apply higher-order thinking such as deductive reasoning, synthesizing, and evaluation. Some examples: (a) Determine the truth or falsity of statements using available evidence. (b) Evaluate the validity of arguments. (c) Identify the correct graphical representation of a given practical situation involving rates of change. (d) Generalize patterns observed in cases, make predictions based on these patterns.**	**25-30%**

NEED FOR DIGITAL LITERACY INCLUSION IN NATIONAL BENCHMARK TEST

Integration of digital literacy dimension into the initial three domains assessed by NTBP would not be out of context in today's world of the fourth industrial revolution (4IR), which comes with the use of technology such as artificial intelligence, blockchain, robot, the internet of things (IoT) and so on. Students need to master a new skill (digital literacy) that would enable them to find, evaluate, utilise, share, and create content using information technologies and the internet, necessary to succeed in higher education institutions. Also, as print mediums begin to face out, the ability to comprehend information found online becomes more and more critical. Students who lack digital literacy skills may soon find themselves at just as much of a disadvantage as those who cannot read or write. Because digital literacy is so critical, educators are increasingly required to teach students digital literacy in the classroom (Edvocate, 2017).

Most students already use digital technology, such as tablets, smartphones, and computers, at home. Many students already know how to navigate the web, share images on social media, and do a Google search to find information. Nevertheless, true digital literacy goes beyond these essential skills. One of the critical components of digital literacy is the ability to find and evaluate information. It means finding the answer to a question or a bit of needed information and then judging whether the source is reliable. The ability to weed out false information and find reliable sources is a critical component of digital literacy and a crucial life skill in the 21st century. Thus, as technology becomes part of daily life, it's more important than ever for assessment bodies such as the CETAP to integrate digital literacy into the existing academic literacy, quantitative literacy, and mathematics areas of assessing examinees. Whether they plan on going to universities, colleges or not, students will need digital literacy to succeed in their personal and professional lives.

TEST DEVELOPMENT OF NBT

A test is used to measure a person's level of skill, accomplishment, or knowledge in a particular area. Educational and training settings often used tests. For instance, tests are used frequently to determine the level of education and the subject area's content in school settings. Examinees may use such tests to determine if they are ready to progress into another grade level. Standardised tests are also utilised primarily in educational settings to determine if students have met specific learning criteria. Tanguma (2000) argues that suitable tests are not easy to come by; they employ efficiency and chronological procedure of test development. This involves four stages: (i) planning the test, (ii) preparing the test, (iii) analysis of the item (iv) marking rubric development. There are two contrasting frameworks (Classical Test, CTT, and Item Response Theory, IRT) in the measurement community where items and tests can be assessed (Ayanwale, 2019). The only difference between the two theories during test development is the nature of the item analysis to establish the test's psychometric properties. A critique of the wording of items, ambiguous use of words, options/keys, and review of items by a panel of experts is the same for the two measurement frameworks (CTT and IRT).

In NBT, the commission engaged item commission writers such as lecturers teaching year-one student courses in South Africa higher education institutions and a few practicing high school teachers to develop items during item development workshops. The teams are constructed based on the participants' expertise in test writers' proficiency at the school-leaving stage wishing to enter higher education. Also, language, disciplinary, and assessment experts (such as assessment system cooperation, USA) drawn from outside the test development teams function as reviewers of the tests in terms of their language, content and format appropriateness, construct representation, biasness, and item analysis through CTT and IRT (Placement, 2017a; Prince, et al., 2018). These items are reviewed for content representation, fairness, and sensitivity to ensure that they do not display bias or function differentially for sub-population of groups. The appropriate items are selected to be pre-tested as non-scoring items in the assembled and administered NBT AQL and mathematics tests. After pre-testing, items are reviewed by review panels for their psychometric properties such as discrimination (a), difficulty (b), and pseudo-guessing (c), respectively. Items acceptable for inclusion in the item bank are used as NBT items (Frith & Prince, 2018). Also, it is important to stress that NBTP usually organises this academic exercise from time to time.

TEST ADMINISTRATION OF NBT

Before the upsurge of the COVID-19 pandemic, the NBT test was primarily administered under a standardised condition, as spelled out in the test administration manual as pencil on paper instruments. NBTP has examination centers across all the provinces (Eastern Cape, Free State, Gauteng, Kwazulu-Natal, Limpopo, Mpumalanga, Northern Cape, North-West, Western Cape, and SADC Region) in South Africa. In the year 2020, this exam was postponed due to the need to comply with Covid-19 rules. With this development, on July 25, 2020, the Centre for Educational Testing for Access and Placement (CETAP) announced the migration of the NBT exam to the linear form of Computer Based Test (CBT)(*National Benchmark Test Project* |, n.d.). This is an electronic way of administering the test items using technology. Candidates were later allowed to choose between the two modes of administration when the issue of COVID-19 gets to a better level in the year 2020. These procedures are the same as those under which the pilot tests were administered and have remained unchanged since the tests first became operational in 2009 (Prince, 2016). These procedures are available from the CETAP at the University of Cape Town (UCT). Twenty-five test forms are assembled and administered each year and include regular test items, common (anchor) items and pre-test items in different forms. Each other document of the Q.L. test contains common items for equating purposes. According to Holland and Dorans (2006), IRT ensures that the scores on different tests are linked and equated to ensure that performance is not a function of the test version that the candidate has received.

NBT SCORING

NBT items are scored dichotomously, that is, either as of right or wrong. Since all tests are power tests, missing responses are scored as wrong. This is valid, given that piloting and the experience of several years show that sufficient time has been allocated to each domain (NBT report, 2017; Prince, et al., 2018). However, scoring multiple-choice response test items is simple, easy, and accurate with the help of machines, especially in large-scale assessments like NBTP. Ayanwale and Adeleke (2019) suggest that it is free of grading bias and efficient, broadly believed by psychometricians to be more valid and reliable. It enables a broader content sampling because numerous test items can be addressed within a specific time frame. The ease of scoring multiple-choice questions allows the assessment bodies to give examinees speedy feedback.

Optical Scanner technology is employed to scan candidates' responses to the NBT items recorded on the bubble sheet. Their responses are scored using the uni-dimensional three-parameter logistic model (a, b, c) of Item Response Theory for the AQL and Mathematics tests, and a score is generated for each candidate on a scale of 0% to 100% (Yen and Fitzpatrick, 2006). IRT is a set of models that relate the likelihood of a particular reaction by an individual examinee with a given trait level to the item's characteristics designed to elicit the level to which individual examinee possesses that trait (Ayanwale, 2019; Rupp, 2009). In addition, the model establishes the relationship between an examinee's latent abilities and the probability of the examinee responding to a specific item correctly. It estimates the parameters involved, explains the processes, and predicts the result of such an encounter (Nenty, 2015). IRT is mainly interested in whether an examinee gets an item correctly or not, rather than in the raw test scores (that is, CTT). This is referred to as the item-pattern scoring procedure. The scoring method produces a maximum likelihood trait estimate based on the pattern of item responses (Tomkowickz & Wright, 2007).

NBT CUT-SCORES

Cut-scores for candidates' assessments have always been indiscriminately determined in many institutions. In South Africa, criterion-referenced is used to report NBT scores for benchmarks set through standard-setting methods to place candidates in one of the described three score bands (proficiency, intermediate and basic) as stated by (Prince & Frith., 2017; Prince, et al., 2018). Table 4 presents the bands.

TABLE 4
TEST PERFORMANCE STANDARDS AND THEIR DESCRIPTIONS FOR BACHELORS, DIPLOMA AND HIGHER CERTIFICATE PROGRAMMES

Performance Band	Description
Proficiency	Performance in the test areas suggests that academic performance would not be seriously affected in the related domains. If admitted, students are likely to be placed on regular programmes of study. The required score range for this band is 70%-100% for Bachelor programmes and 63%-100% for Diploma and Higher Certificate programmes, respectively.
Intermediate	Identified challenges in domain areas are such that it is predicted that academic progress will be seriously affected in related domains. If admitted, students' educational needs should be met in a way deemed apt by the institution (for example, extended or augmented programmes, special skills provision). The required score range for this performance band is 38%-69% for Bachelor programmes and 34%-62% for Diploma and Higher Certificate programmes, respectively.
Basic	Serious learning challenges were identified. Without extensive and long-term support, students will not cope with university study, perhaps best provided through bridging programmes. Institutions admitting students within this score proficiency need to offer this support themselves. The score range for this performance band is 0%-37% for Bachelor programmes and 0%-33% for Diploma and Higher Certificate programmes respectively.

The bands explain their readiness for the demands of higher education and the extent to which the curricula should be responsive to their level of preparedness. Standard-setting workshops to determine the benchmark levels take place every three years, in which panels of lecturers in first-year courses in higher education take part. Several methods (such as Angoff, Ebel, Nedelsky, Bookmark, and I.D. Matching) are available for standard-setting in high stake examination (Hambleton & Pitoniak, 2006). Currently, NBTP uses the Angoff method to determine the candidates' benchmarks, which must be justified with empirical data (Frith & Prince, 2018).

Angoff's method uses a group of experts to judge how difficult each item is in an exam to determine the cut-off score (Angoff, 1971). For instance, the cut-off score or mark or benchmark is like a borderline in the sand that divides candidates into two groups: those below the cut-off and those above the cut-off.

Below the cut-off may indicate a failure, and above the cut-off may show a pass. The Angoff method calculates a cut-off mark based on the performance of candidates to a defined standard (absolute) as opposed to how they perform to their peers (relative). It involves judging exam items (test-centered) instead of exam candidates (examinee-centered). In practice, this method relies on subject-matter experts (SMEs) who examine the content of each test question (item) and then predict how many minimally qualified candidates (that is, a candidate believed to be located at the borderline of a particular proficiency category) would answer the item correctly. The average of the judges' predictions for a test question becomes its predicted difficulty. The sum of the predicted difficulty values for each item averaged across the judges and items on a test is the recommended Angoff cut score. This process is repeated for all borderlines between proficiency categories of interest. For example, Table 5 below shows a hypothetical way a cut-off mark can be estimated using the Angoff method.

TABLE 5
SCORES (%) FROM EACH SUBJECT-MATTER EXPERT (SME) AND FINAL CUT-SCORE

Question	**SME 1 (%)**	**SME 2 (%)**	**SME 3 (%)**	**SME 4 (%)**	**SME 5 (%)**	**SME 6 (%)**	**SME 7 (%)**	**SME 8 (%)**	**SME 9 (%)**	**Cut score, mean**
Question 1	55	60	60	60	55	50	50	85	80	61.67
Question 2	54	56	55	55	60	60	60	55	60	57.22
Question 3	55	55	50	50	60	65	65	65	60	58.33
Question 4	52	53	55	55	60	60	50	70	65	57.78
Question 5	65	60	60	70	65	65	60	55	58	62.00
Question 6	56	50	49	55	50	50	55	55	58	53.11
Question 7	60	58	65	60	60	50	55	53	50	56.78
Question 8	70	67	65	60	75	65	65	58	55	64.44
Question 9	56	55	60	50	50	55	52	75	60	57.00
Question 10	70	65	60	75	60	60	58	70	54	63.56
The final average cut score for minimum competency/borderline										
	59.3	57.9	57.9	59	59.5	58	57	64.1	60	51.19

In this hypothetical example, the table submitted that the total average percentage is 51.19. This can be rounded to 51, giving a cut-off rate of 51%. If the test were out of 100 marks, a borderline candidate would be expected to get 51 out of 100 marks. In essence, the SMEs estimate the borderline candidate's expected standard or cut score or benchmark on each item. Then these desired item scores are summed to obtain an estimated accurate score for the borderline candidate on the collection of test items. This process, which involves SMEs in determining the cut-scores for the proficiency bands, supports the argument for the validity of the use of the test for describing the proficiency of prospective students for the demands of higher education (Prince, et al., 2018).

USABILITY OF NBT SCORES

A large number of universities in South Africa used NBT scores to place candidates appropriately. Universities to which candidates apply for admission receive NBT results directly from the CETAP. Many South African universities use the NBTs with the National Senior Certificate (NSC) to access their programmes (Prince & Frith, 2017). The NBTs scores and NSC results help the universities in different ways to make decisions about candidate access/application to university. This is used to determine whether candidates are ready for academic study. Also, few universities use the results for placement within the institution. This implies that NBT scores are used to decide whether the candidate will need extra academic support after being accepted to university or not. And some institutions use NBT results to develop curricula within their university.

VALIDITY AND RELIABILITY OF NBT

Test developers and test users applied validity to make inferences derived from test scores valid for making decisions. As Taherdoost (2016) described, validity is a measure of what is purported to be measured. This is used to infer from test scores to a larger domain of items comparable to those utilized in the test. Validity is not an absolute state but rather a collection of evidence indicating that the scores obtained on a test are valid for their intended uses (American Educational Research Association, 2014). In this paper, types of validity employed by NBTP were examined. Content validity, construct validity (convergent and discriminant), criterion validity (concurrent and predictive), and reliability. Content validity is the degree to which items in an instrument reflect the domain area they are purported to measure. This is achieved using test specifications developed by experienced experts in teaching the domains (quantitative literacy and mathematics) that constitute the test. In addition, items for the tests are generated annually by South African higher education academics, using domains test specification as to their guide, which also verifies that the items' content is akin to the pertinent aspects of the test specifications and aligned with the construct (Prince, 2016). Furthermore, this paper suggests adopting the content validity ratio (CVR) proposed by Lawshe (1975) and test specification used. The CVR is a linear transformation of a proportional level of agreement on how many "experts" within a panel rate an item "essential" calculated in the following way:

$$CVR = \frac{n_e - \left({}^{N}\!/_{2}\right)}{\frac{N}{2}}$$

where CVR is the content validity ratio, n_e is the number of panel members indicating "essential," and N is the total number of panel members. The final evaluation to retain the item based on the CVR is on the number of panels.

Construct validity is the extent to which a test measures the theoretical construct or trait that it was purported to measure. This is usually conducted by analyzing the observed score correlations of a test with another test based on the theory underlying the constructs being measured. If the idea of the constructs predicts that the two tests correlate, then there should be a real relationship between the tests to be valid. Otherwise, the tests do not measure the constructs (Ayanwale, 2019).

According to Cohen and Swerdlik (2009), criterion validity is a judgment regarding how adequately a test score can be used to infer an individual's relative standing on some measures of interest. Okpala et al. (2007) stated that criterion-related validity is the degree to which scores gotten with an evaluation instrument are under current criterion measures. Criterion-related validity exists in two forms, namely, predictive validity and concurrent validity. Predictive validity involves using test scores to predict criterion measurement made at some point in the future. In contrast, concurrent validity is the relationship between the test scores and criterion measures when both are obtained simultaneously.

Also, reliability is the extent to which a measurement of a phenomenon provides a stable and consistent result. Testing for reliability is essential as it refers to the consistency across the parts of a measuring instrument (Huck, 2007). The NBT test is assessed for reliability as measured by Kuder Richardson 20 (KR_{20}), which is frequently used to determine the internal consistency of multiple-choice items. KR_{20} indicates how related the scores on the items "hang together" and measure the same construct. No absolute rules exist for internal consistencies. However, most agree that a minimum internal consistency coefficient of 0.70) or higher has traditionally been considered; reliability values above 0.80 are desirable (Nunally, 1978; Robinson, 2009; Whitley, 2002).

CONCLUSION AND RECOMMENDATIONS

The NBT aims to assess the school-leaving higher education applicant pool, i.e., the national cohort of school-leavers wishing to access higher education in their first year. The tests aim to address the following question: What are the academic literacy, quantitative literacy, and mathematics levels of proficiencies of the school-leaving population, who wish to continue with higher education, at the point before they enter into higher education at which they could realistically be expected to cope with the demands of higher education study? The constructs and domains of the three tests are based on testing this question, and the levels of the tests have been set with the notion of levels of proficiency as focus. There is also no doubt that the NBTP will strengthen HESA's enrolment services and, at a national systems level, increase the sector's responsiveness to the different challenges entailed in access and the consequences of a changing schooling-HE interface. NBTP will inform HE curricula and teaching and learning practices and provide second chance entry opportunities to those whose school-leaving results prevent them from gaining access to higher education study. The study recommends that DL should be included in the existing three core domains of assessment, the four-parameter logistic model of IRT should be explored for item calibration, model-data fit assessment should always be conducted to determine the best model fit, and adaptive type of computer-based test should be considered in today's world of 4IR for efficiency measurement of precision.

REFERENCES

American Educational Research Association. (2014). *Standards for Educational and Psychological Testing*.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement*. American Council on Education.

Ayanwale, M.A. (2019). *Efficacy of Item Response Theory in the Validation and Score Ranking of Dichotomous and Polytomous Response Mathematics Achievement Tests in Osun State, Nigeria*. Doctoral Thesis, Institute of Education, University of Ibadan (Issue April). https://doi.org/10.13140/RG.2.2.17461.22247

Ayanwale, M.A., & Adeleke, J.O. (2019). Invariance Person Estimate of Basic Education Certificate Examination: Classical Test Theory and Item Response Theory Scoring Perspective. In *Jiste* (Vol. 23, Issue 1).

Cohen, R.J., & Swerdlik, M.E. (2009). *Psychological testing and assessment: An introduction to tests and measurement* (4th ed.). Mayfield Publishing House.

Edvocate. (2017). *What is digital literacy?* Retrieved from https://www.theedadvocate.org/what-is-digital-literacy/

Foxcroft, G.F., & Roodt, R.C. (2005). *An Introduction to Psychological Assessment in the South African Context* (2nd ed.). Oxford University Press.

Frith, V., & Prince, R. (2006). Quantitative literacy. In H. Griesel (Ed.), *Access and entry-level benchmarks, the National Benchmark Tests Project* (pp. 47–54). Higher Education South Africa.

Frith, V., & Prince, R. (2009). A framework for understanding the quantitative literacy demands of higher education. *South African Journal of Higher Education*, *23*(1), 83–97. https://doi.org/https://doi.org/10.4314/sajhe.v23i1.44804

Frith, V., & Prince, R. (2017). Mathematical literacy of students in the first year of medical school at a South African university. In A.B. & R. Laugksch (Ed.), *Proceedings of the 12th Annual Conference of the Southern African Association for Research in Mathematics, Science and Technology Education* (Vol. 21, Issue 2, pp. 791–798). SAARMSTE.

Frith, V., & Prince, R. (2018). The National Benchmark Quantitative Literacy Test for Applicants to South African Higher Education. *Numeracy, 11*(2). https://doi.org/10.5038/1936-4660.11.2.3

Frith, V., & Prince, R.N. (2016). Quantitative Literacy of School Leavers Aspiring to Higher Education in South Africa: Lessons from the South African National Benchmark Quantitative Literacy Test.

South African Journal of Higher Education, *30*(1), 138–61. https://doi.org/https://doi.org/10.20853/30- 1-552

Frith, V., Bowie, L., Gray, K., & Prince, R. (2003). Mathematical literacy of students entering the first year at a South African university. *Proceedings of the Ninth National Congress of the Association for Mathematics Education of South Africa*, pp. 186–193.

Griesel, H. (2006). *Access and entry-level benchmarks, the National Benchmark Tests Project*. Higher Education South Africa.

Hambleton, R.K., & Pitoniak, M.J. (2006). Setting Performance Standards. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433–700). Greenwood Praeger.

Holland, P.W., & Dorans, N.J. (2006). Linking and Equating. In R.L. Brennan (Ed.), *In Educational Measurement* (4th ed., pp. 187–220.). Greenwood/Praeger.

Huck, S.W. (2007). *Reading Statistics and Research*. Allyn & Bacon.

Hughes-Halett, D. (2001). Achieving Numeracy: The Challenge of Implementation. In *Mathematics and Democracy*. Retrieved from https://scholarcom

Lawshe, C.H. (1975). A Quantitative Approach to Content Validity. *Personnel Psychology*, *28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Le Roux, N., & Sebolai, K. (2017). The national benchmark test of quantitative literacy: Does it complement the grade 12 mathematical literacy examination? *South African Journal of Education*, *37*(1), 1–11. https://doi.org/10.15700/saje.v37n1a1350

National Benchmark Test Project. (n.d.). Retrieved November 9, 2021, from https://www.nbt.ac.za/

Nel, B.P. (2020). Implications of the quantitative literacies test results of the national benchmark test project (NBTP) for teachers. *South African Journal of Education*, *40*(1), 1–8. https://doi.org/10.15700/saje.v40n1a1792

Nenty, H.J. (2015). Possible Factors that Influence Student's Pattern of Responses to Mathematics Examination Items. *Journal of Educational Assessment*, *17*(4), 47–58.

Nunally, J.C. (1978). *Psychometric Theory* (2nd ed., Issue October 2011). McGraw-Hill.

O'Connell, B. (2006). Preamble. In H. Griesel (Ed.), *Access and Entry-Level Benchmarks: The National Benchmark Tests Project*. HESA.

Okpala, P.N., Onocha, C.O., & Oyedeji, O.A. (2007). *Measurement in Education*. Stirling-Horden Publishers (Nig) Ltd.

Parliamentary Monitoring Group. (2009). *National benchmark tests project & standards for national examination & assessment systems*. Department of Higher Education. Retrieved from https://pmg.org.za/committee-meeting/10668/

Placement, C. (2017a). *National Benchmark Tests Project National Report: 2017 Intake Cycle*.

Placement, C. (2017b). *National Benchmark Tests Project National Report: 2017 Intake Cycle* (Issue March).

Prince, R. (2016). Predicting Success in Higher Education: The Value of Criterion and Norm-Referenced Assessments. *Practitioner Research in Higher Education*, *10*(1), 22–38.

Prince, R. (2017). The relationship between school-leaving examinations and university entrance assessments: The case of the South African system. *Journal of Education (University of KwaZulu-Natal)*, *70*, 133–160.

Prince, R., & Frith, V. (2017). The quantitative literacy of South African school-leavers who qualify for higher education. *Pythagoras*, *38*(1). https://doi.org/10.4102/pythagoras.v38i1.355

Prince, R., Balarin, E., Nel, B., Padayashni, R.P., Mutakwa, D., & Niekerk, A.D. (2018). *The National Benchmark Tests national report: 2018 intake Cycle* (Issue May). Retrieved from www.nbt.ac.za

Prince, R.N., & Archer, A. (2008). A New Literacies Approach to Academic Numeracy Practices in Higher Education in South Africa. *Literacy and Numeracy Studies*, *16*(1), 63–75. https://doi.org/https://doi.org/10.5130/lns.v16i1.1948

Prince, R.N., & Frith, V. (2017). The Quantitative Literacy of South African School-Leavers Who Qualify for Higher Education. *Pythagoras*, *38*(1), 355–370. https://doi.org/https://doi.org/10.4102/pythagoras.v38i1.355

Prince, R.N., & Simpson, Z. (2016). Quantitative Literacy Practices in Civil Engineering Study: Designs for Teaching and Learning. In R. Anne-Mette Nortvig, Birgitte Holm Sørensen, Morten Misfeldt (Ed.), *Proceedings of the 5th International Conference on Designs for Learning* (Vol. 6, Issue 7, pp. 73–85).

Robinson, J. (2009). *Triandis theory of interpersonal behaviour in understanding software privace behaviour in the South African context*. The University of the Witwatersrand.

Rupp, A.A. (2009). Item Response Theory modeling with Bilog-MG and Multilog for windows. *International Journal of Testing*, 3(4), 365–384.

Steen, L.A. (2004). *Achieving Quantitative Literacy: An Urgent Challenge for Higher Education* (pp. 94–110). The Mathematical Association of America. https://doi.org/10.5281/zenodo.1162967

Taherdoost, H. (2016). Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire / Survey in a Research. *International Journal of Academic Research in Management*, *3*, 28–36. https://doi.org/10.2139/ssrn.3205040

Tanguma, J. (2000). *Steps in Test Construction*. Paper Presented at the Annual Meeting of the Southwestern Psychological Association.

Tomkowickz, J.T., & Wright, K.R. (2007). *Investigation of the Effect of Test Equating and Scoring Methods on Item Parameter Estimates and Student Ability Scores*. A Paper Presented at the Annual Conference of American Educational Research Association.

Whitley, B.E. (2002). *Principals of Research and Behavioural Science*. McGraw-Hill.

Yen, W.M. & Fitzpatrick, A.R. (2006). Item Response Theory. In R.L. Brennan (Ed.), E*ducational Measurement* (4th ed., pp. 53–111). Greenwood/Praeger.