DEVELOPMENT OF AN ACHIEVEMENT TEST FOR THE GENERAL EDUCATION COURSE

Sureeporn Panyangam[1], Nutthapat Kaewrattanapat[2], Marndarath Suksanga[3]
[1,2] Faculty of Humanities and Social Sciences, Suan Sunandha Rajabhat University, Thailand
[3] College of Politics and Governance, Suan Sunandha Rajabhat University, Thailand, E-Mail:
E-mail: [1]sureeporn.pa@ssru.ac.th, [2] nutthapat.ke@ssru.ac.th, [3] marndarath.su@ssru.ac.th

ABSTRACT

Abstract—This research study focuses on the development of an achievement test for the general education course GEN0117 (Survival Science). The development process involved the following steps: 1) Content Validity Analysis using the IOC (Index of Item Objective Congruence) value calculated from the opinions of 3 experts on 80 initial items; 2) Item Analysis to determine the difficulty index (*p*) and the discrimination index (*r*); and 3) Reliability Analysis using the KR-20 (Kuder-Richardson Formula 20) to select items meeting standard criteria for the item bank. A trial run was conducted on a sample group of 400 students using a 60-item test. Data analysis, based on comparing the top 27% (approximately 108 students) as the High Group (H) and the bottom 27% (approximately 108 students) as the Low Group (L), revealed that the difficulty index (*p*) ranged from 0.31 to 1.00, indicating the items ranged from moderately difficult to moderately easy. The discrimination index (*r*) ranged from 0.00 to 0.83, suggesting the items had low to acceptable discriminatory power. The KR-20 reliability coefficient for the achievement test was found to be 0.646, which means the test set is at an acceptable level for measuring student learning achievement.

Keywords—Achievement Test, Survival Science Course (GEN0117)

INTRODUCTION

The development of student learning is a core aspect of instructional management. To effectively drive this development, assessment is essential to determine the extent to which students' knowledge and abilities meet the specified learning objectives. The achievement test is a widely used instrument for measuring and evaluating learning outcomes (Paungsombat, et al., 2020, p. 12)[1]. However, in the current era where computer technology plays a crucial role in education, it has led to the development of new testing concepts and theories, such as Item Response Theory (IRT), coupled with the use of technology to create and manage assessment tools. This includes the implementation of an Item Bank, which serves as a repository for a large number of high-quality items designed to measure the knowledge, abilities, and skills resulting from student learning, aligned with course objectives and curriculum requirements.

The Item Bank is a necessary tool for administering accurate and reliable multiple-choice examinations. It allows for the convenient storage and retrieval of test items, ensuring they can be used rapidly and precisely for various instructional assessment goals, thus eliminating the need for instructors to create entirely new tests for every examination (Mcdonald, 2002, p. 200 cited in Charoenkaensai, et al., 2021, p. 32)[2]. Storing test items in a computerized database further facilitates retrieval, usage, and subsequent item additions. Nevertheless, the process of analyzing items to ensure their quality before inclusion in the item bank remains a demanding task for instructors. Generally, they face challenges in managing tests, often having to restart the item creation, storage, and retrieval process each time, which hinders the opportunity for continuous and systematic improvement of test quality.

Consequently, a robust item bank system addresses these issues by providing essential convenience, enabling instructors to manage test items efficiently. It allows for the systematic storage and accumulation of high-quality items and facilitates the quick and targeted retrieval of these items for measuring learning outcomes, consistently resulting in high-quality tests. This systematic development of item quality is beneficial for the overall teaching and learning process, granting instructors more time to dedicate to the creative and qualitative aspects of instruction. Therefore, the researcher is interested in studying the development of an achievement test for the Survival Science course (GEN0117) offered by the School of General Education and Electronic Learning Innovation at Suan Sunandha Rajabhat University. The expected research outcome is the acquisition of quality test items to be stored in an online item bank for the aforementioned course.

RESEARCH OBJECTIVE

To develop a psychometrically sound achievement test and prepare the resulting high-quality items for inclusion in an online item bank.

LITERATURE REVIEW

1. Concepts and Principles of Achievement Tests

An achievement test is an instrument used to evaluate the knowledge, abilities, skills, and learning acquired by students after the completion of an instructional process. The primary objective is to verify the extent to which students have met the Learning Objectives stipulated in the curriculum or course. The essential characteristics of a good test must include consistency with the content taught (Content Validity), appropriate difficulty (Difficulty Index, p), and the ability to differentiate high-ability students from low-ability students (Discrimination Index, r) (Ebel & Frisbie, 1991)[3]. Furthermore, Reliability is a crucial property indicating the consistency and stability of the measurement tool. In this research, the KR-20 (Kuder-Richardson Formula 20)

is used, which is suitable for estimating the reliability of tests scored dichotomously (0 or 1 for incorrect/correct).

$$KR-20 = \left(\frac{k}{k-1}\right)\left[\frac{\sum_{i=1} p_i q_i}{\sigma^2}\right]$$

Where k is the number of items, p_i is the proportion of examinees answering item i correctly, q_i is the proportion of examinees answering item i incorrectly, and σ^2 is the variance of the total scores.

2. Characteristics of Quality Items and Item Analysis

Creating a quality achievement test requires a process of Item Analysis to select sound items for inclusion in the item bank. Key criteria for consideration are:

- Content Validity: Verified by experts to calculate the Index of Item Objective Congruence (IOC), which should ideally be $\geq$ 0.50 to indicate that the item measures the intended content.
- Difficulty Index (p): This value shows the proportion of examinees who answered the item correctly. An optimal range is typically between 0.20 and 0.80 (some sources suggest 0.30 to 0.70). A value approaching 1.00 indicates the item is too easy, and a value approaching 0.00 indicates it is too difficult.
- Discrimination Index (r): This value demonstrates the item's ability to distinguish between high-scoring (High Group, H) and low-scoring (Low Group, L) students, often calculated using the top and bottom 27% of total scores. A good discriminating item should have an r value of $\geq$ 0.20; the closer the value is to 1.00, the better. Items with values below 0.20 or negative values should be revised or discarded.

3. Development and Management of Item Banking

The development of an item bank is a significant concept under the application of Item Response Theory (IRT) and advances in computer technology. An Item Bank is a database system that collects a large number of test items that have been analyzed for quality according to standard criteria. The benefits of maintaining an item bank include:

- Efficiency in Management: Reduces the instructor's burden of creating new tests each time, allowing for quick searching, selection, and retrieval of required items.
- Quality and Reliable Testing: Items in the bank have undergone quality analysis, ensuring that the assembled tests possess high validity and reliability.

- Computerized Adaptive Testing (CAT): At an advanced level, the item bank is the essential foundation for CAT, which can select items tailored to the ability level of each individual examinee (McDonald, 2002 cited in Charoenkaensai, et al., 2021)[2].

Storing test items in a computerized database and an online format is therefore a crucial approach to elevating the quality of assessment in the current era, especially for general education courses with a large number of students, such as the GEN0117 Survival Science course, which requires efficient test management.

RESEARCH METHODOLOGY

Population and Sampling

The main target group is students enrolled in the GEN0117 Survival Science course in the 2/2566 academic year at Suan Sunandha Rajabhat University, which has a total of 2,065 registrations. We selected two sample groups using Simple Random Sampling (like drawing names from a hat) to ensure fair selection:

1. Tool Quality Group: 100 students used for the preliminary testing to check if the test items we created are of good quality.

2. Actual Data Collection Group: 400 students used for collecting the actual test scores for the main analysis.

Instrument Used: The Survival Science Achievement Test, consisting of 80 items. Data was collected from the students' examination results via an online system.

Test Analysis Methods

We will analyze the test results in three main steps to select the best items for the bank:

1. Checking Content Match (Content Validity)

- 3 Experts will review each test item to ensure it measures the content that was actually taught.
- We use the IOC (Index of Item Objective Congruence) score as a criterion. If an item gets an IOC $\geq$ 0.50, it passes the content check.

2. Analyzing Item Quality (Difficulty and Discrimination)

- **Difficulty Index (*p*):** Measures how many students answered the item correctly.
  - Passing Criterion: The item should not be too easy or too difficult; the *p* value must be between 0.20 and 0.80 (values outside this range need revision).

- **Discrimination Index** (***r***): Measures how well the item separates high-performing students from low-performing students.
    - Passing Criterion: The item must have an *r* value of □0.20 (items below this value are poor discriminators and should be discarded).

Table 1 Interpretation criteria, test difficulty (p), and discrimination power (r)

p **Value (Difficulty)**	**Interpretation**	***r*** **Value (Discrimination)**	**Interpretation**
0.20 - 0.39	Rather Difficult	0.20 - 0.39	Low Discrimination
0.40 - 0.59	Moderate	0.40 - 0.59	Acceptable Discrimination
0.60 - 0.80	Rather Easy	0.60 - 1.00	Good Discrimination

3. Finding Overall Test Reliability

- We use the **KR-20 Formula** to calculate if all 60 items are internally consistent and give stable results.
- A high reliability score (*r*) means the test set is highly trustworthy and can be confidently used for assessment.

RESULT

The research began with the development of 80 achievement test items, followed by the initial quality verification step:

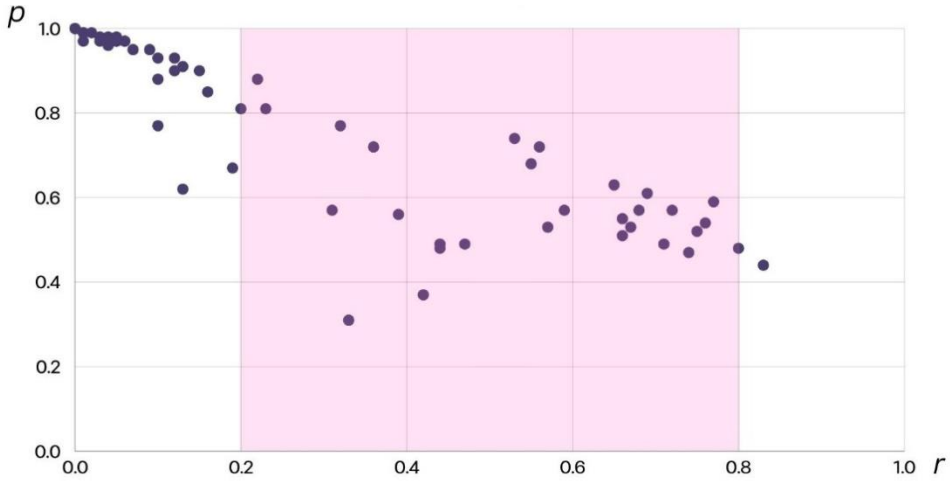
1. Content Validity Analysis.

This was conducted by calculating the Index of Item Objective Congruence (IOC) based on the judgments of 3 experts. The evaluation results indicated that 60 items met the selection criterion (IOC □ 0.50), accounting for 75% of the initial item pool. The IOC values for these approved items ranged from 0.60 to 1.00, suggesting a high degree of congruence between the majority of the test items and the defined course content. Conversely, 20 items were found not to meet the evaluation criterion and were subsequently removed at this stage. Following this successful content validation, the researcher administered the remaining 60 items to a sample group of 400 students for the subsequent quality verification steps, including the analysis of the difficulty index, discrimination index, and overall test reliability.

2. Analysis of Item Difficulty, Discrimination, and Reliability

The subsequent quality verification step involved administering the 60 content-validated items to the sample group of 400 students to analyze the Difficulty Index (*p*) and the Discrimination Index (*r*). For the discrimination analysis, the top scoring group (H) and the bottom scoring group (L) were selected, each consisting of 27% of the total sample, or approximately 108 students per group. The analysis revealed that the Difficulty Index (*p*) for the 60 items ranged from 0.31 to 1.00. This range suggests that while many items are within the acceptable range (especially those 0.31), some items (those approaching 1.00) are potentially too easy and require review or revision. The Discrimination Index (*r*) ranged from 0.00 to 0.83, indicating that while most items demonstrated good to very good discriminatory power, some items (those approaching 0.00) lacked the ability to differentiate between high and low achievers. Furthermore, the Internal Consistency Reliability of the entire test was calculated using the KR-20 (Kuder-Richardson Formula 20) , yielding a coefficient of 0. 646. This value indicates that the achievement test possesses an acceptable level of reliability.

Figure 1
The graph shows the difficulty (*p*) and discrimination (*r*) values of the test items



3. Summary of Item-by-Item Quality Analysis

The quality analysis of the 60 achievement test items, based on the established psychometric criteria (Difficulty Index 0.20 $\leq p \leq$ 0.80 and Discrimination Index $r \geq$ 0.20), yielded three primary categories for item selection:

3.1 Items Approved for the Item Bank: A total of **28 items** (e.g., items 7, 11, 12, 16, 22, 23, 28, 29, 30, 36, 37, 38, 41, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60) successfully

met both the acceptable difficulty and discrimination criteria. These items are deemed suitable for immediate inclusion in the item bank for future use in measurement.

3.2 Items Recommended for Revision: There are **6 items** (e.g., 4, 8, 9, 10, 13, 35) that failed to meet at least one of the established criteria (either *p* or *r*). These items possess potential but require further refinement and psychometric adjustment to achieve the desired quality standards before being utilized.

3.3 Items Recommended for Deletion: A significant number, **26 items** (e.g., 1, 2, 3, 5, 6, 14, 15, 17, 18, 19, 20, 21, 24, 25, 26, 27, 31, 32, 33, 34, 39, 40, 42, 43, 44, 45), clearly failed to meet both minimum criteria (very low discrimination and/or inappropriate difficulty). These items are considered unusable and should be permanently removed from the item pool.

Note: The final decision regarding item quality must also consider external factors, such as the motivation and effort level of the student sample during testing, as well as ensuring that the selected items collectively cover all learning objectives and the entire scope of the course content.

CONCLUSION

This research aimed to develop an achievement test for the Survival Science course, and the quality analysis results are presented as follows. In the initial phase of Content Validity Analysis, conducted using the Index of Item Objective Congruence (IOC), it was found that out of the 80 initial test items, 60 items successfully passed the selection criterion (with IOC values ranging from 0.60 to 1.00), while 20 items did not meet the requirement. These 60 content-validated items were subsequently administered to a sample of 400 students for further item analysis. The analysis of item characteristics revealed that the Difficulty Index (*p*) ranged from 0.31 to 1.00, indicating a wide range of difficulty from rather difficult to rather easy (or potentially too easy for some items). The Discrimination Index (*r*) ranged from 0.00 to 0.83, showing that the items possessed a discriminatory power ranging from minimal to very good. Furthermore, the overall test reliability, calculated using the KR-20 Formula, yielded a coefficient of 0.646, which signifies that the test set achieves an acceptable level of reliability for measuring student learning achievement.

DISCUSSION

The findings from the development of the achievement test for the GEN0117 Survival Science course can be discussed based on the quality of the measurement instrument:

Content Validity and Item Quality

The expert selection process (IOC) yielded 60 items with strong content congruence, forming a solid base for quantitative analysis. However, the item analysis showed a wide range for the Difficulty Index (p from 0.31-1.00) and the Discrimination Index (r from 0.00-0.83). Although 28 items met the high-quality standards, 26 items had to be discarded. Similarly, Amédée (2025)[4] found that the difficulty index of 60 psychology items in the DRC secondary school exams ranges from 0.15 to 0.80, with the easiest items being the highest index and the most difficult items being the lowest index. Item analysis can improve the quality of educational tests by identifying items with average difficulty and high discriminating power, with functional distractors (Quaigrain & Arhin, 2017)[5]. And Distractor efficiency (DE) in multiple-choice items impacts both difficulty index and discrimination power, with efficient distractors associated with lower difficulty index and discriminating items (Rezigalla et al, 2024)[6]. Item analysis of multiple choice questions can help identify good and ideal questions for future assessments, while identifying those that need revision (Kumar et al, 2021)[7]. This result is consistent Farooq & Mashood (2023)[8] with the Test analysis is a valuable assessment tool, helping to identify better multiple-choice questions for memorization and eliminate weak ones, helping to improve teachers' writing skills. Item analysis is a valuable assessment tool that identifies better multiple-choice questions to be retained while discarding or reviewing the weak ones. Faculty development programs should be organized for improving item writing skills of faculty.

Test Reliability

The overall test reliability (KR-20) was 0.646, which falls within the acceptable range. This suggests the test has a certain degree of measurement consistency but could be further improved (reliability should ideally be $\geq$0.70). The moderate reliability may be attributed to the inclusion of 26 items with low discrimination and inappropriate difficulty in the test set before final item culling, this research successfully generated 28 high-quality test items suitable for inclusion in the online item bank for future assessment, thus fulfilling the objective of instrument development, KR-20 is theoretically and empirically related to Cronbach's alpha, but KR-20 is specifically designed for dichotomous data, while Cronbach's alpha applies more generally; both coefficients tend to correlate highly but can differ depending on test length and item characteristics (Uyanah & U, 2023)[9]; (Anselmi et al., 2019)[10]. KR-20 can be used alongside item analysis metrics such as item difficulty and discrimination to evaluate and improve test quality (Ntumi et al., 2023)[11]. supporting the findings of Boopathiraj & Chellamani (2013)[12] that some items were rejected due to their poor discrimination index. And consistent Karim, Sudiro, & Sakinah (2021)[13] found that 16 items out of 50 test items were rejected due to the poor and worst quality level of difficulty and discriminating index. Meanwhile, 12 items need to be reviewed due to their mediocre quality. This demonstrates that the quality of achievement assessment is directly contingent upon the quality of the assessment instruments, which must be verified as psychometrically sound and reliable tools (Siridet Suchiva, 2003, cited in Eakwannang, 2022, p. 99)[14].

Suggestions for Application

The 28 high-quality items derived from this study should be stored in an **online item bank system**. They should be categorized according to specific Learning Objectives for efficient retrieval and use in future achievement tests.

Suggestions for Future Research

Continuous Item Development: The 6 items categorized for revision should be modified, focusing on clarifying ambiguous distractors and improving question structure, and then retested.

Analysis using Item Response Theory (IRT): In the next phase, the 28 quality items should be subjected to IRT analysis to determine precise item parameters (difficulty and discrimination), which is essential for creating equivalent parallel tests or implementing Computerized Adaptive Testing (CAT).

ACKNOWLEDGEMENTS

The authors would like to thank Suan Sunandha Rajabhat University, Bangkok, Thailand to provide funding support to attend the dissemination of research on this and thank family, friends, colleagues, students in Suan Sunandha Rajabhat University and The Office of General Education and Innovative e-Learning for cooperation and provide the dataset in research, all of you.

REFERENCES

[1] Paungsombat, K., Senarat, S., & Senarat B. (2020). QUALITYASSESSMENTOFTESTING AND DIAGNOSTIC ASSESSMENTFORPROBLEM-SOLVING SKILLS. *Journal of Graduate School Sakon Nakhon Rajabhat University*, 16(75), 7–15. retrieved from https://so02.tci-thaijo.org/index.php/SNGSJ/article/view/123666

[2] Charoenkaensai, R., Senarat, S., & Senarat, B. (2021). APPLYING 4-PARAMETER ITEM RESPONSE THEORYFOR DEVELOPING A TEST ITEM BANK OF THAI MUSIC. *Journal of Graduate School Sakon Nakhon Rajabhat University*, 18(81), 31–40. retrieved from https://so02.tci-thaijo.org/index.php/SNGSJ/article/view/243527

[3] Ebel, R.L. and Frisbie, D.A. (1991). *Essentials of Educational Measurement*. 5th Edition, Prentice-Hall, Englewood Cliffs.

[4] Amédée, M. (2025). ANALYSIS OF THE DIFFICULTY INDEX OF SIXTY PSYCHOLOGY ITEMS IN THE END OF SECONDARY SCHOOL EXAMS IN THE DRC: CASE OF THE GOMA FINALISTS. *IJRDO- Journal of Educational Research*. https://doi.org/10.53555/er.v11i1.6244

[5] Quaigrain, K., & Arhin, A. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, 4. https://doi.org/10.1080/2331186x.2017.1301013.

[6] Rezigalla, A., Eleragi, A., Elhussein, A., Alfaifi, J., Alghamdi, M., Ameer, A., Yahia, A., Mohammed, O., & Adam, M. (2024). Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, 24. https://doi.org/10.1186/s12909-024-05433-y

[7] Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Medical journal, Armed Forces India*, 77 Suppl 1, S85-S89 . https://doi.org/10.1016/j.mjafi.2020.11.007

[8] Farooq, M., & Mashood, S. (2023). Quality Assurance of Multiple-Choice Questions Test Through Item Analysis. *Life and Science*. https://doi.org/10.37185/lns.1.1.315

[9] Uyanah, D., & U., N. (2023). The Theoretical and Empirical Equivalence of Cronbach Alpha and Kuder-Richardson Formular-20 Reliability Coefficients. *International Research Journal of Innovations in Engineering and Technology*. https://doi.org/10.47001/irjiet/2023.705003

[10] Anselmi, P., Colledani, D., & Robusto, E. (2019). A Comparison of Classical and Modern Measures of Internal Consistency. *Frontiers in Psychology*, 10. https://doi.org/10.3389/fpsyg.2019.02714

[11] Ntumi, S., Agbenyo, S., & Bulala, T. (2023). Estimating the Psychometric Properties (Item Difficulty, Discrimination and Reliability Indices) of Test Items using Kuder-Richardson Approach (KR-20). *Shanlax International Journal of Education*. https://doi.org/10.34293/education.v11i3.6081

[12] Boopathiraj, C., & Chellamani, K. (2013). ANALYSIS OF TEST ITEMS ON DIFFICULTY LEVEL AND DISCRIMINATION INDEX IN THE TEST FOR RESEARCH IN EDUCATION. *International Journal of Social Sciences & Interdisciplinary Research*, 2, 189-193

[13] Karim, S., Sudiro, S., & Sakinah, S. (2021). Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. *EduLite: Journal of English Education, Literature and Culture*. https://doi.org/10.30659/e.6.2.256-269

[14] Eakwannang, W. (2022). Developing the Competency Assessment Instrument for Learning Management During Covid 19 Period of the Student Teacher, Faculty of Education, SuanSunandha Rajabhat University. *Journal of Graduate MCU KhonKaen Campus*, 9(1), 89–101. retrieved from https://so02.tci-thaijo.org/index.php/jg-mcukk/article/view/250820