https://doi.org/10.59865/abacj.2023.59
Special Issue-Application of Structural Equation Modeling in Business, Sport, Tourism, Recreation, Education, and Language Studies

5 vs 4: A QUANTITATIVE INVESTIGATION INTO THE QUALITY METRICS OF DIFFERENT MULTIPLE-CHOICE TEST FORMATS

Sarhistthep Sukkaew[1] and Supamas Chumkaew[2,*]

Abstract

This study employed quantitative methods to address two primary objectives: 1) to compare the quality of 5-choice and 4-choice multiple-choice tests, and 2) to evaluate the discriminant power of these formats using test response theory with kernel smoothing. Data were collected from 1,966 students at Sukhothai Thammathirat Open University who took a 120-question multiple-choice exam during the second semester of 2019. Four test configurations were analyzed: the Initial Case utilized the original 5-choice format; Case 1 randomly omitted one option from the 5-choice test, excluding the correct answer; Case 2 randomly omitted one option, including the correct answer; and Case 3 adapted the options based on the test-taker's proficiency level. The study employed Cronbach's Alpha (denoted as raw_alpha) as a reliability metric, discovering varying levels of reliability across the four cases. The highest reliability was observed in Case 3 with a raw_alpha value of 0.87. There were no differences in the difficulty values or discriminatory power across all cases. The mean scores indicated that students generally performed better on the 4-choice tests in Cases 1-3 than on the original 5-choice format, referred to as the Initial Case. These findings have significant implications for test design, suggesting that 4-choice tests can achieve comparable reliability and discriminatory power to traditional 5-choice tests.

Keywords: 5-choice multiple choice test; 4-choice multiple choice test; multiple choice test; Adaptive test; Quality of test; Kernel Smoothing

1. INTRODUCTION

Educational assessment is a cornerstone of effective pedagogy, yet direct measurement of student performance often proves challenging (Greaney & Kellaghan, 2008). In light of these difficulties, multiple-choice tests have become a widely adopted indirect assessment tool (Greving, Lenhard & Richter, 2020, 2023). Such tests vary in format, most commonly employing either 4-option or 5-option choices. The chosen format impacts not only the test's difficulty and the likelihood of guessing but also its overall reliability and validity.

One vital aspect to consider in the design of multiple-choice tests is the crafting of effective distractors—incorrect options that are plausible enough to be chosen by those not fully knowledgeable on the subject. Distractors that fail to be chosen indicate their ineffectiveness and compromise the quality of the assessment (Kumar, Nayak, Shenoy & Goyal, 2023; McDaniel & Little, 2019; Rodriguez, 2005).

Prior research has been inconclusive re

[1] Dr. Sarhistthep Sukkaew, is currently a lecturer in the Office of Registration, Records and Evaluation, Sukhothai Thammathirat Open University, Thailand. He obtained an Ed.D. in Educational Research and Evaluation from Kasetsart University, Thailand. Email: sarhistthep.suk@stou.ac.th.

[2,*] Dr. Supamas Chumkaew, is currently a lecturer in the Office of Registration, Records and Evaluation, Sukhothai Thammathirat Open University, Thailand. She obtained a Ph.D. in Educational Research Methodology from Chulalongkorn University, Thailand. Corresponding Email: supamas.chu@stou.ac.th

garding the optimal number of choices for these tests. While some studies suggest that a 5-option format is superior in terms of reliability, others find little difference in quality between 4-option and 5-option multiple-choice tests (Esmaeeli, Esmaeili Shandiz, Norooziasl, Shojaei, Pasandideh, Khoshkholgh & Barkhordari Ahmadi, 2021; Haladyna & Downing, 1993; Lord, 1977). This raises questions on whether the time and effort invested in generating an additional option actually contributes to the test's effectiveness.

This study aims to address this gap in the literature by analyzing real-world data from Sukhothai Thammathirat Open University, a prominent distance-learning institution with a large student body. The study aims to evaluate whether reducing the number of options from 5 to 4 significantly impacts test quality in terms of reliability, discriminatory power, and difficulty. The results are intended to guide educational institutions in making informed decisions regarding test formats, thereby enhancing the evaluation process for both educators and students alike.

2. RESEARCH ASSUMPTION

The research relies on two types of data: Part 1 consists of secondary data, drawn from the five-choice test results of 1,966 undergraduate students at Sukhothai Thammathirat Open University during the second semester of 2019, covering a total of 120 exam questions. Part 2 involves simulation data, transforming the original five-choice test into a four-choice format based on three foundational assumptions.

Assumption 1 posits that high-ability examinees will correctly answer multiple-choice questions on their first attempt.

Assumption 2 asserts that all examinees, regardless of ability, have an equal probability of choosing any given answer option.

Assumption 3 stipulates that high-ability examinees will consistently answer questions correctly, regardless of the number of choices available, whereas low-ability examinees will not benefit from a reduction in choices from five to four, as they will continue to answer randomly.

3. LITERATURE REVIEW

3.1 Kernel Smoothing in Item Response Theory

McGuire (2012) examined the application of Item Response Theory (IRT) with Kernel smoothing in understanding the relationship between latent variables and observed variables in educational and psychological assessments. Observed variables typically stem from multiple choice exams, where one option is considered correct, and questions based on a scale in which each option carries different weightage.

A common challenge within Polytomous IRT is the limitations of mathematical models in accurately describing the probabilities of choosing among various options. This is encapsulated in the Option Characteristic Curve (OCC) (Effatpanah, & Baghaei, 2023). Traditional analysis of multiple-choice items often relies on basic statistical metrics such as p-values and Point Biserial correlations, which may not be sufficient. OCCs are considered pivotal starting points for advanced IRT analyses (Rajlic, 2020).

Two primary methods are employed for estimating OCCs. The first is Parametric IRT, where a predefined structure of parameters for OCCs is assumed. This method focuses on reducing the estimate to a parameter vector, which summarizes key statistical descriptors such as the difficulty level and discrimination power of each item. The second method, Non-Parametric IRT, directly estimates OCCs without a predetermined mathematical model. It offers the flexibility of being computationally convenient and is considered to produce OCC estimates closer to their true values. However, Non-Parametric IRT methods are less commonly used in comparison to Parametric IRT (McGuire, 2012; Rajlic, 2020).

In summary, this study employs equating—a rigorous statistical framework—to standardize scores from varying tests measuring the same construct, thereby ensuring

reliable and valid inter-test comparisons. Concurrently, Non-Parametric IRT is applied to analyze a 5-option multiple-choice exam, generating OCCs as benchmarks for discarding inefficient options. Principal Component Analysis (PCA) further complements this analysis by assessing the quality of both 5-choice and 4-choice multiple-choice tests.

3.2 Equating

Equating is a statistical technique used to convert scores from two different tests measuring the same attribute into comparable units (Kanchanawasi, 1998; Lord, 1980). This approach involves administering two distinct tests to a single group of test-takers. Scores from these tests are then standardized, making them directly comparable. However, a prerequisite for this process is that the scores from both tests must follow a normal distribution equate: An R package for observed-score linking and equating (Albano, 2016; Angoff, 1984; Petersen, Marco, & Stewart, 1982). In essence, equating enables the translation of scores from different tests into a common unit of measurement, thereby allowing for a valid and reliable comparison of performance across different assessments.

3.3 Conditions for Equating

Equating mandates certain conditions for the process of score equating (Albano, 2016; A; Petersen, Marco, & Stewart, 1982).

1. Both tests must assess the same construct, be it a skill, ability, or other characteristic.

2. After transformation, scores should exhibit a distribution similar to the reference test.

3. Invariance across groups necessitates that transformed scores remain constant irrespective of the test-taking population.

4. Symmetry dictates that score equating should be consistent, whether comparing from test version X to Y or vice versa.

3.4 Types of Equating

Equating techniques can be broadly categorized into two main forms, depending on the context in which they are applied (Effatpanah & Baghaei, 2023; Kanchanavasi, 1998):

Horizontal Equating: This type of equating is used to compare scores between different tests that focus on the same attributes and have similar levels of difficulty. Horizontal equating is particularly useful when multiple versions of a test are created on the same content, often to maintain test security or for administering the test at different times. In such cases, the test-takers come from the same population and possess similar abilities. This form of equating ensures that the scores from these different tests are comparable, thereby establishing fairness. For example, horizontal equating would be appropriate for calibrating scores between two different mathematics tests intended for Grade 6 students. Both tests would be designed to measure the same skill set at the same level of difficulty.

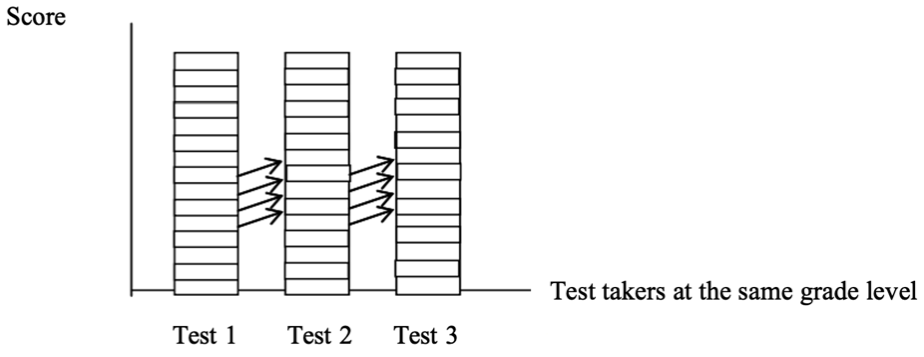

Figure 1 Horizontal Equating

Achieving perfect score equivalence in horizontal equating is complex. Tests should serve as parallel "alternate forms," closely matched in difficulty and content. Moreover, the competency distribution among test-takers must be similar across forms. These challenges warrant careful attention to maintain the validity and fairness of the equating process.

Vertical Equating: This type of equating is complex in its application. Vertical equating is designed to compare test scores across different grade levels, focusing on the same subject area but varying in difficulty. For instance, this method is frequently used to link math test scores from Grades 4, 5, and 6. The objective is to determine how a Grade 4 score correlates with those in Grades 5 and 6, enabling longitudinal analysis of student progress. The technique is intricate due to the varying test complexity and the distinct ability distributions in each grade level. Despite these challenges, vertical equating remains invaluable for educational assessments spanning multiple years.

In vertical equating, tests that evaluate the same subject but target different grade levels or skill sets are aligned for comparison. Each test version presents varying degrees of difficulty and is administered to groups with different ability distributions. These variations introduce complexities, making vertical equating theoretically and practically more challenging than its horizontal counterpart.

3.5 Data Collection Methods in Test Equating for Single Group Design

Uncounterbalanced Design

In this straightforward approach, a single sample group of test takers is given both tests in sequence. Since the same individuals take both tests, variations in ability levels are minimized, making score comparisons straightforward. However, factors like learning, practice effects, and fatigue from the first test may influence scores on the subsequent test (Effatpanah, & Baghaei, 2023; Kanchanavasi, 1998; Panidvadtana, 2019). Thus, the following hypotheses are proposed accordingly:

H1: According to Assumption 1, student scores from a 4-choice test, with one option randomly eliminated (excluding the correct answer), will be higher than scores obtained from a 5-choice test.

H2: According to Assumption 2, student scores from a 4-choice test, where one option is randomly eliminated (including the correct answer), will exceed scores from a 5-choice test.

H3: According to Assumption 3, scores from a 4-choice test, with option elimination based on the test-taker's ability level, will surpass scores from a 5-choice test.

Counterbalanced Design

To address the order effects found in the uncounterbalanced design, this model randomizes examinees into two subgroups. One

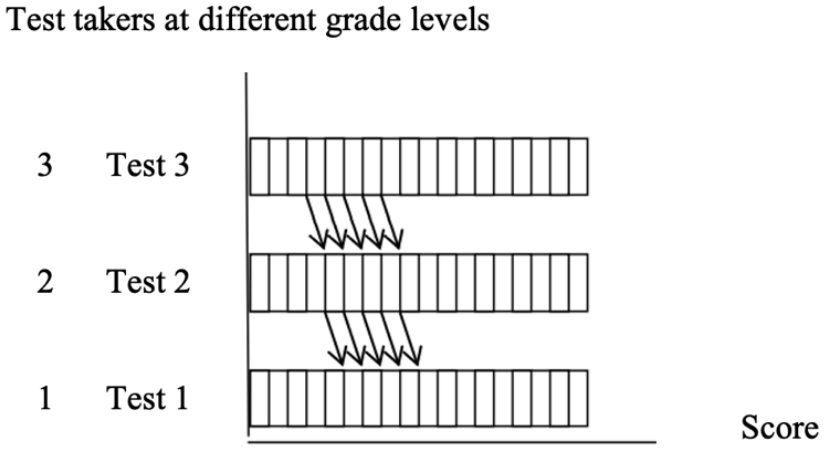

Figure 2 Vertical Equating

Table 1 Single Group Design

Data collection	Sample	Test	
		No. 1	**No. 2**
Single group of test takers	P_1	X	Y
1. Unbalanced single test-taker format	P_1	X	Y
2. Balanced single test-taker format	P_2	Y	X

subgroup takes the first test followed by the second, while the other subgroup does the opposite. This design aims to balance out influences such as test order, learning, practice, and fatigue across both subgroups (Effatpanah, & Baghaei, 2023; Kanchanavasi, 1998; Panidvadtana, 2019).

3.6 Equating Methodologies

Various methodologies exist for score equating, which can generally be classified into two primary categories according to the foundational principles of equating (McGuire, 2012; Rajlic, 2020; Kanchanawasi, 1998). These categories are Classical Equating Models and Item Response Theory-Based Equating.

In the context of this research, focus is directed solely on the Linear Equating method, a specific approach within classical equating models. This method serves as the chosen equating technique for score comparison, grounded in traditional theoretical frameworks.

3.7 Linear Score Equating

Linear score equating is predicated on the idea that scores from two different tests for a given group of test-takers are deemed equivalent if they align with the same standard scores (Angoff, 1984; Petersen, Marco, & Stewart, 1982). Recognized for its simplicity and convenience, linear score equating serves as a straightforward method for score comparison (Angoff, 1984). Several paradigms for the systematic collection and organization of data for linear score equating have been proposed. Although these paradigms vary in their approaches to estimating test means and standard deviations, they all determine equivalent scores based on the same standard score metrics.

$$\frac{Y - M_Y}{S_Y} = \frac{X - M_X}{S_X} \quad \text{----- (1)}$$

A reconfiguration of Equation (1) yields the following:

$$(Y - M_Y)/S_Y = (X - M_X)/S_X$$

where X,Y: Scores from Test X and Test Y
M_X, M_Y: Mean scores from Tests X and Y
S_X, S_Y: Standard deviation of scores from Tests X and Y

In this framework, scores from Tests X and Y can be linearly equated as follows: $Y = A_{YX}X + B_{YX}$

for $A_{YX} = \frac{S_Y}{S_X}$ and $B_{YX} = M_Y - A_{YX} M_X$

then $Y = A_{YX}X + B_{YX}$ transform to equation 2

$$Y = \left(\frac{S_Y}{S_X}\right)X + \left(M_Y - \frac{S_Y}{S_X}M_x\right) \quad \text{----- (2)}$$

$$(Y - M_Y)/\frac{S_X}{S_Y} = X - M_X$$

$$X = M_X + (Y - M_Y)\frac{S_X}{S_Y}$$

$$= M_X + Y\frac{S_X}{S_Y} - M_Y\frac{S_X}{S_Y}$$

$$= \frac{S_X}{S_Y}Y + \left[M_X - M_Y\frac{S_X}{S_Y}\right] \quad (3)$$

This study employs a single-group unbalanced design for data collection, wherein a single cohort of examinees serves as co-testers, taking both versions of the test under investigation. This approach is commonly referred to as a single-group design. For the purpose of score equating, the research utilizes linear equating, a method falling under the umbrella of classical equating models.

4. RESEARCH METHODOLOGY

This study employs a quantitative approach, analyzing data from a 5-option multiple-choice test administered to undergraduate students at Sukhothai Thammathirat Open University during the second semester of 2019. The data were analyzed using IRT with Kernel Smoothing to assess test quality and discriminatory power.

4.1 Data Sources

Data were obtained from two sets:

Secondary Data: Results from a 5-option multiple-choice test taken by 1,966 undergraduate students, featuring 120 questions.

Simulated Data: Based on the initial test results, simulations were conducted to create three scenarios for a 4-option test, each aligned with specific research assumptions.

4.2 Methodological Approach

Linear equating was employed for score equating, a method within the classical equating models. The simulated 4-option test data were analyzed alongside the 5-option test data for a comprehensive assessment of test quality.

4.3 Research Procedures

1) Data Selection: Data from academic years 2019 to 2021 were analyzed, focusing on tests with high sample sizes (>1000 students). The 2019 second-semester dataset, consisting of 1,966 students and 120 questions, was selected for its comprehensiveness and reliability.

2) Data Preprocessing: Quality checks were conducted to identify missing or outlier data, which were then replaced using calculations facilitated by the psych package in R.

3) Option Characteristic Curve Analysis: The KernSmoothIRT package in R was used to construct OCC for each question.

4) Option Elimination: Ineffective options were identified and removed based on their discriminatory power values and OCC data.

4.4 Criteria for Option Elimination

The criteria for eliminating options were grounded in their discriminatory power values and OCC data as outlined by Guo, Zu, & Kyllonen (2018) as follows:

Positive Discriminatory Values: Eliminate the option with the highest positive discriminatory value.

Negative Discriminatory Values: Target the option with the smallest negative value for elimination.

Mixed Discriminatory Values: Consider options with positive values for elimination.

Criteria for option elimination rely on the OCC, which graphically represents the likelihood that a test-taker will select a particular option. An option is considered for elimination if its corresponding curve on the OCC is closest to the horizontal axis, signaling a low probability of selection by the test-taker. In cases where the curve for a distractor option is absent or extremely low on the OCC, this further confirms that the distractor is ineffective and merits removal, as it fails to attract even the low-ability test-takers.

An Illustrative Case: Transitioning from a 5-Choice to a 4-Choice Test Format

Exam 1 underwent comprehensive data analysis, yielding the following results:

The initial phase focused on quantifying discriminatory power values.

Subsequent to the discrimination analysis, the OCC were evaluated as shown in Figure 3.

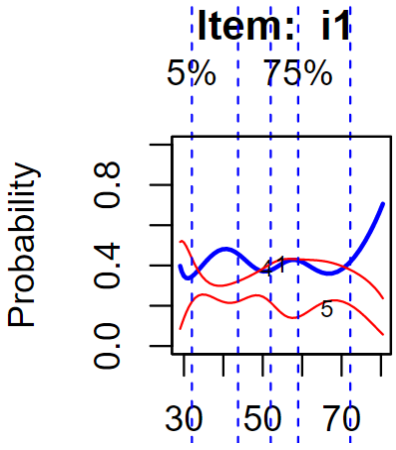

Figure 3 Illustrates the OCC for Question 1

Table 2 Discriminatory Power Values Derived from Transitioning from a 5-Choice to a 4-Choice Test Format

Item 1	i1. correct	i1. key	i1. n	i1. rspP	i1. pBis	i1. discrim	i1. lower	i1. mid50	i1. mid75	i1. upper
1		1	902	0.46	-0.04	0.01	0.44	0.47	0.49	0.45
2		2	34	0.02	-0.08	-0.02	0.02	0.02	0.02	0.00
3		3	23	0.01	-0.07	**-0.01**	0.02	0.01	0.00	0.01
4	*	4	716	0.36	0.01	0.05	0.36	0.33	0.35	0.42
5		5	291	0.15	-0.07	-0.03	0.16	0.17	0.13	0.13

An analysis of the discriminatory power values and OCC suggests the potential for eliminating Options 1, 2, 3, and 5. Based solely on discriminatory power, Option 1 would be the obvious choice for elimination. However, a deeper analysis reveals that Option 1 successfully misled up to 902 examinees. Moreover, the OCC indicates that a significant portion of test-takers still select Option 1 across a range of expected scores. As a result, Option 1 was retained for its effectiveness in misleading low-ability examinees.

Conversely, neither Option 2 nor Option 3 show evidence of being chosen across varying levels of ability in the OCC analysis. When combined with their negative discriminatory power values, the decision was made to eliminate Option 3 for Question 1.

4.5 Test Quality and Replacement Procedures

1) Replacement Options: replacement options will only affect candidates who initially chose the eliminated options. It's assumed that examinees who did not select the eliminated options will retain their original answers when transitioned to a 4-choice format. To guide the replacement of eliminated options, three basic research assumptions were followed, each corresponding to a specific case. The procedures for each case are summarized in Table 3.

Table 3 Method for Representing Answers According to Assumption in the Research

Assumption	Method
Assumption 1: High-ability test takers are likely to answer multiple-choice questions correctly on their first attempt.	Case 1: Represents randomly omitted options, excluding the correct answer.
Assumption 2: All test-takers have an equal opportunity to select any given answer.	Case 2: Represents randomly omitted options, including the correct answer.
Assumption 3: High-ability test takers are more likely to answer questions correctly, whereas low-ability test takers are less likely to do so.	Case 3: Represents options eliminated based on the test taker's ability level.

This section outlines the methodology for substituting values for omitted options, using the example of option elimination for question 1. Each case will be detailed individually as follows:

The example in Table 4 focused on analyzing the results from Question 1, which initially had five answer options. Option 3 was chosen for elimination, yielding a new 4-option version of the test. This option was found to be ineffective, serving as a distractor. Based on the data table, 23 test takers chose Option 3, necessitating a systematic method for its substitution. To achieve this, three different methods were employed for option replacement, as follows:

Case 1: Random Substitution, Excluding the Answer

In this case, any remaining option other than the correct answer (Option 4) can replace Option 3. Therefore, Options 1, 2, and 5 are candidates for substitution.

Case 2: Random Substitution, Including the Answer (Guo, Zu, & Kyllonen, 2018).

Here, any option including the correct answer (Option 4) can replace Option 3. Therefore, Options 1, 2, 4, and 5 could be used.

Case 3: Ability-Level-Based Substitution (Guo, Zu, & Kyllonen, 2018).

Test-takers are divided into high and low-ability groups. High-ability test-takers will have Option 3 replaced with the correct answer (Option 4), while low-ability test-takers will have it replaced with one of the remaining incorrect options: Options 1, 2, or 5.

2) Test quality analysis: the psych package in R was used for assessing test reliability, based on a minimum reliability score of 0.50.

Table 4 Approaches for Representing Answer Choices

Item 1	i1. correct	i1. key	i1. n	i1. rspP	i1. pBis	i1. discrim	i1. lower	i1. mid50	i1. mid75	i1. upper
1		1	902	0.46	-0.04	0.01	0.44	0.47	0.49	0.45
2		2	34	0.02	-0.08	-0.02	0.02	0.02	0.02	0.00
3		3	23	0.01	-0.07	**-0.01**	0.02	0.01	0.00	0.01
4	*	4	716	0.36	0.01	0.05	0.36	0.33	0.35	0.42
5		5	291	0.15	-0.07	-0.03	0.16	0.17	0.13	0.13

3) Difficulty and discriminatory power: the KernSmoothIRT package in R and PCA were used to evaluate the metrics.

4.6 Simulation Design

A Single-Group Uncounterbalanced Design was employed. Scores from the original and the new tests were compared using various methods, such as equating between the 5- choice and 4-choice tests and ability-level-based substitution.

4.7 Data Analysis and Hypothesis Testing

A comparative analysis of the 5-choice and 4-choice test scores was conducted using statistical tests such as the mean, standard deviation, and the Wilcoxon Signed Rank test.

5. RESULTS

5.1 Evaluation of Test Reliability and Validity Across 5-Choice and 4-Choice Question Formats: An Examination of Three Contextual Scenarios in Accordance with Initial Research Assumptions.

5.1.1 Evaluation of Test Quality for the Initial Case: The 5-Choice Test Format

The psychometric properties of the 5-choice test were assessed using the Psych package in R, focusing on the Cronbach's Alpha Coefficient as a measure of reliability. The test demonstrated a notably reliability score (raw_alpha) of 0.81, indicating high internal consistency for the selected assessment.

The assessment of difficulty and discriminatory power was carried out using PCA

```
Reliability analysis
Call: alpha(x = Score$scored, check.keys = FALSE)

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean    sd median_r
      0.81       0.8    0.82     0.033   4 0.006 0.43 0.092     0.03

    95% confidence boundaries
         lower alpha upper
Feldt      0.8  0.81  0.82
Duhachek   0.8  0.81  0.82
```

Figure 4 Reliability Analysis for the Initial Case: The 5-Choice Test Format

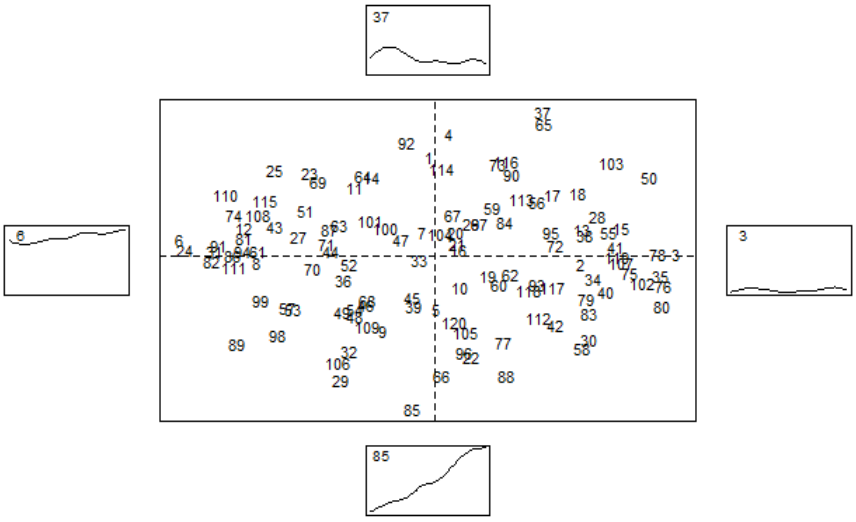

Figure 5 Difficulty and Discriminatory Power Analysis with PCA for the Initial Case: The 5-Choice Test Format

on ICC, obtained from the KernSmoothIRT package in the R programming environment. Kernel Smoothing in IRT served as a non-parametric method for estimating OCC, particularly when the model did not conform to parametric IRT assumptions.

Upon evaluation through PCA, it was discovered that out of 120 questions, 85 were well-classified. Specifically, the items 29, 66, 88, 106, 32, 22, and 96 were positioned close to the vertical axis dotted line and below the horizontal axis dotted line, indicating good classification. Conversely, the items 67, 114, and 1 were not as well-classified, with items 4, 92, 37, and 65 demonstrating poor classification.

When assessing the degree of question difficulty, items 103, 50, 3, 78, 35, 76, 80, 102, 75, 107, 110, 41, and 15 were identified as particularly challenging. Especially, items positioned at the end of the dotted line on the horizontal axis and to the right of the vertical axis (e.g., items 78, 3, 35, 76, and 80) were deemed highly difficult.

Questions positioned on the horizontal axis dotted line and to the left of the vertical axis dotted line (e.g., items 6, 24, 82, 31, 91, 88, 110, 74, and 89) were deemed comparatively easier. Among them, question 6 stood out as being particularly easy.

```
Reliability analysis
Call: alpha(x = Score$scored, check.keys = FALSE)

  raw_alpha std.alpha G6(smc) average_r S/N    ase mean    sd median_r
      0.79      0.78     0.8     0.029 3.6 0.0066 0.45 0.089    0.027

    95% confidence boundaries
         lower alpha upper
Feldt     0.78  0.79   0.8
Duhachek  0.78  0.79   0.8
```

Figure 6 Reliability Analysis for Case 1: The 4-Choice Test Format (Based on Preliminary Assumption 1)

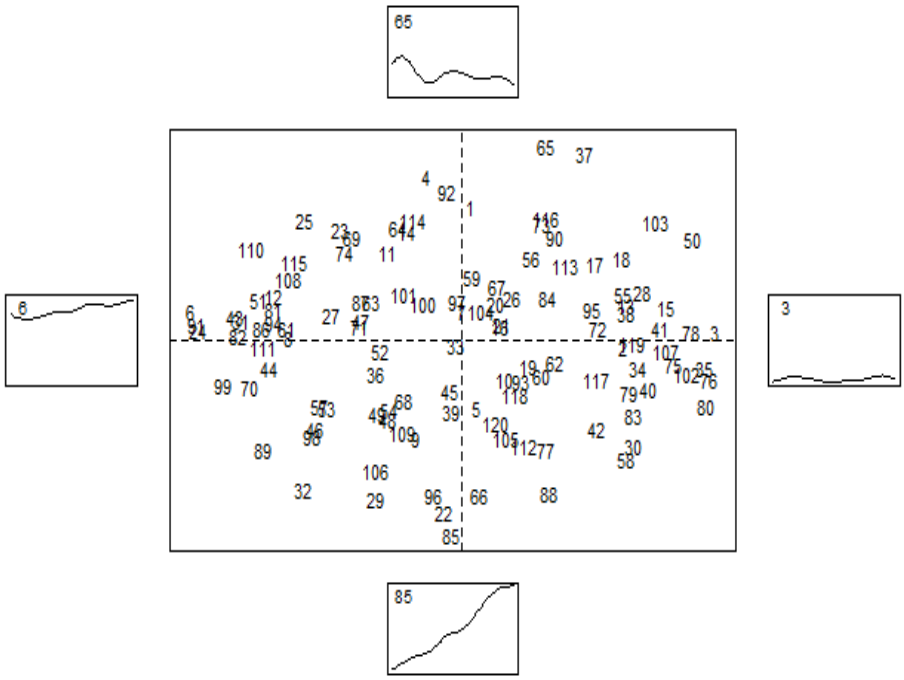

Figure 7 Analysis of Difficulty and Discriminatory Power for Case 1: The 4-Choice Test Format (Based on Preliminary Assumption 1)

For the remainder of the questions, the difficulty level was considered to fall within acceptable criteria.

5.1.2 Evaluation of Test Quality for Case 1: A 4-Option Multiple Choice Test (Randomly Eliminated Options Excluding Correct Answer, Based on Preliminary Assumption 1)

The reliability of the 4-option multiple-choice test was evaluated after excluding poorly classified distractors while retaining the correct answer options. Distractor elimination was guided by discrimination values (discrim values) obtained from the CTT package. Specifically, distractors with only positive discrim values were considered for elimination starting with the most positive. Conversely, for items with only negative discrim values, the distractors with the least negative values were considered for elimination. Additional criteria for distractor elimination were based on OCC graphs. Distractors that exhibited low probability of selection by test-takers, as represented by graph lines close to or adjacent to the horizontal axis, were also considered ineffective and were removed.

The raw_alpha reliability value for this modified 4-option test was 0.79, which is relatively high and closely aligned with the 5-option test. Notably, the reliability value experienced a minor decline after the elimination of only four options.

PCA was employed to assess difficulty and discriminatory power. Out of 120 questions, 85 were well-classified, including items 29, 66, 88, 106, 32, 22, and 96, among others. Items such as 65, 37, 4, 92, and 1 were poorly classified. Particularly challenging questions included 103, 50, 3, 78, 35, 76, 80, 102, 75, 107, 41, and 15. Questions positioned on the left of the vertical dotted line on the PCA plot, such as items 6, 91, 24, 99, 48, and 82, were identified as relatively easy, with item 6 being notably easy.

Considering the results, a parallelism in the difficulty and discriminatory power between the 5-option and the modified 4-option tests was observed. Specifically, the characteristics of both test formats appear to be aligned, demonstrating a consistency in the assessment regardless of the test format.

5.1.3 Evaluation of Test Quality for Case 2: A 4-Option Multiple Choice Test (Involving Random Omission of Options Including Correct Answers, Based on Preliminary Assumption 2)

The analysis examines the reliability and discriminative power of the 4-choice test format under the guidelines of the preliminary assumption 2. For this, the CTT package and OCC graphs were used to examine the probability of test-takers choosing particular distractors.

The calculated raw alpha reliability value was 0.79, which is comparable to that of the 4-choice format in other conditions and even to the 5-choice test. The reliability did not significantly differ after eliminating only four of the least effective distractors.

```
Reliability analysis
Call: alpha(x = Score$scored, check.keys = FALSE)

  raw_alpha std.alpha G6(smc) average_r S/N    ase mean    sd median_r
      0.79      0.78     0.8     0.029 3.6 0.0066 0.45 0.089    0.027

    95% confidence boundaries
         lower alpha upper
Feldt     0.78  0.79   0.8
Duhachek  0.78  0.79   0.8
```

Figure 8 Reliability analysis for Case 2: The 4-Choice Test Format (Based on Preliminary Assumption 2)

PCA was employed to assess the test questions for difficulty and discriminatory power. Out of 120 questions, 29 were classified as performing well. Notably, items such as 3, 78, 50, 76, 35, and 80 were identified as particularly challenging, while items such as 6, 94, 24, and 99 were easier for the test-takers.

Most of the test questions from the 5-choice test and the 4-choice tests (both with and without randomly eliminated options) showed similar levels of difficulty and discriminatory power. Therefore, preliminary assumption 2 was substantiated, affirming that each option, including the correct answers, had an equal probability of being selected by the test-taker regardless of their skill level.

This analysis supports the reliability and effectiveness of the 4-choice test format, especially when compared with the 5-choice format, both in terms of reliability and discriminatory power.

5.1.4 Results of Quality Analysis of the Multiple-Choice Test, Case 3, 4-Choice Test (Elimination of Choices According to Test-Taker Ability Level, Based on Preliminary Assumption 3)

Reliability was measured using the CTT package and was complemented by the OCC. Test options were eliminated based on their discrim values, and replacement choices were determined by the test takers' ability levels. Test takers were divided into two ability groups:

Group 1: High-performing test takers (scores >= 60 out of 120)

Group 2: Low-performing test takers (scores < 60 out of 120)

For Group 1, eliminated distractors were replaced with correct answers, while for Group 2, eliminated distractors were randomly replaced with the remaining options. This was done in accordance with observed OCC trends that indicate high-ability test takers are more likely to choose correct answers, and low-ability test takers are more likely to opt for incorrect choices.

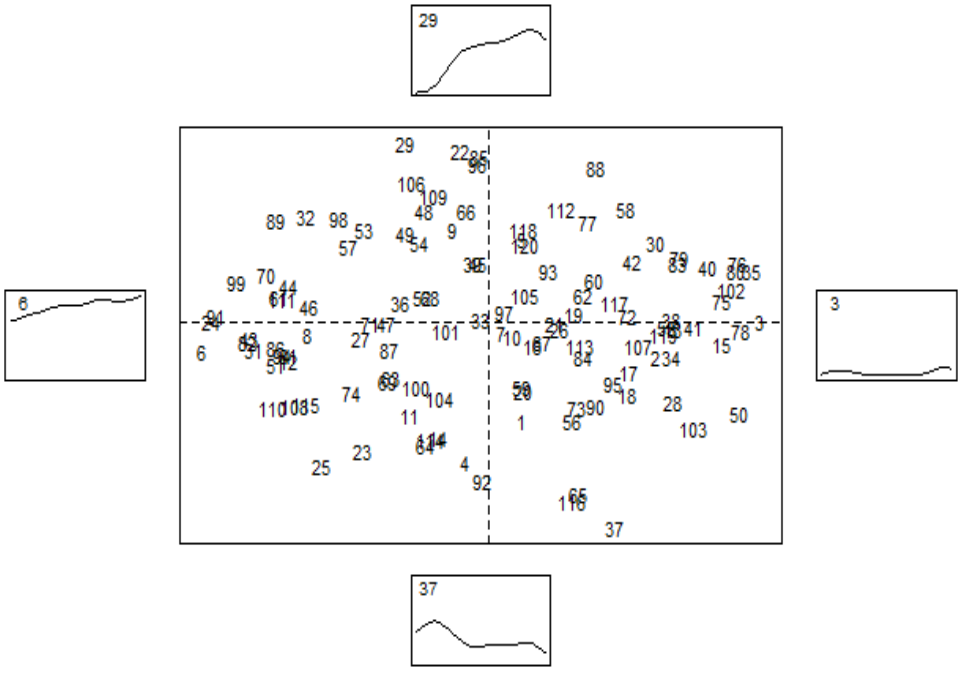

Figure 9 Analysis of Difficulty and Discriminatory Power for Case 2: The 4-Choice Test Format
(Based on Preliminary Assumption 2)

```
Reliability analysis
Call: alpha(x = Score$scored, check.keys = FALSE)

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean   sd median_r
      0.87      0.87    0.88     0.053 6.7 0.004 0.44 0.11     0.05

    95% confidence boundaries
         lower alpha upper
Feldt     0.86  0.87  0.88
Duhachek  0.86  0.87  0.88
```

Figure 10 Reliability Analysis for Case 3: The 4-Choice Test Format
(Based on Preliminary Assumption 3)

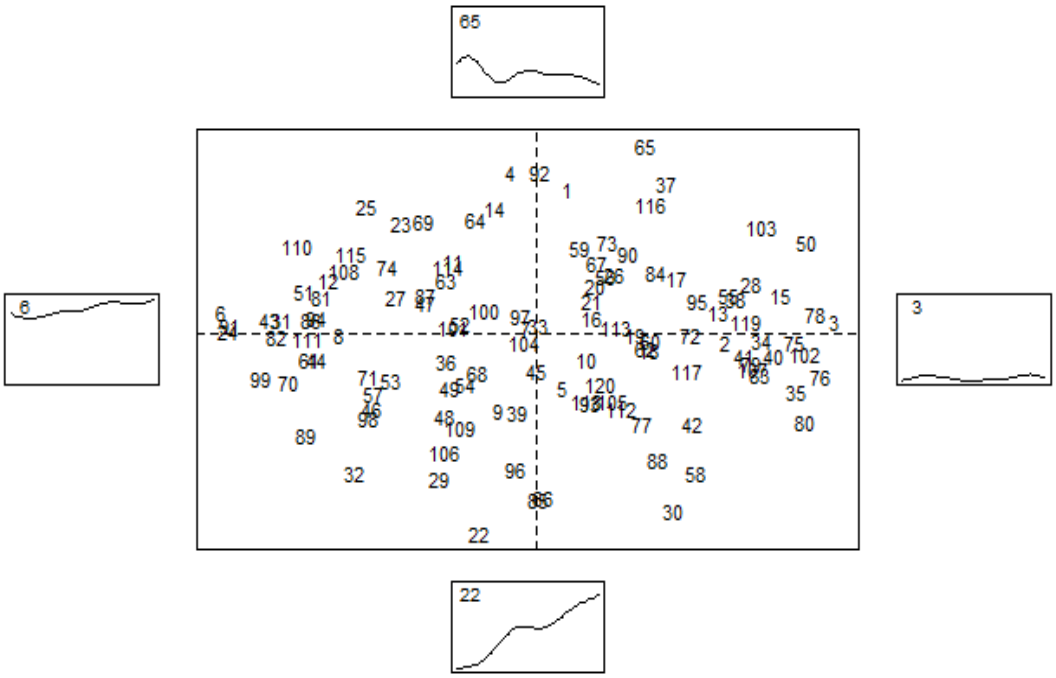

Figure 11 Analysis of Difficulty and Discriminatory Power for Case 3: The 4-Choice Test Format
(Based on Preliminary Assumption 3)

The raw alpha reliability value for this format was notably high at 0.87, exceeding both previous 4-choice formats (0.79) and the 5-choice test (0.81).

Using PCA, it was found that out of 120 questions, only 22 could be classified as well-performing. Questions such as 3, 78, 50, 76, 35, 80, 102, 75, and 15 were identified as notably difficult. On the flip side, questions 6, 82, 24, and 99 were found to be relatively easier.

The analysis suggests that there is little difference in difficulty and discriminatory power across the 5-choice test and the different 4-choice test formats (based on Preliminary Assumptions 1, 2, and 3). Therefore, each type of test—regardless of the preliminary assumptions guiding its construction—appears to offer consistent levels of challenge and discrimination among the test items.

This high reliability, along with comparable difficulty and discrimination metrics across different test formats, validates the efficacy of the 4-choice test model, especially when adapted to suit test takers of varying ability levels as per preliminary assumption 3.

5.1.5 Results of the Comparative Analysis of Test Quality for 5-Choice and 4-Choice Tests Across 3 Preliminary Assumption Scenarios

Upon analyzing the four different test cases—each based on different test structures and replacement methods—it was observed that the test cases varied significantly in their reliability coefficients:

Initial Case: 5-choice test format.

Case 1: 4-option multiple choice test (randomly eliminated options excluding correct answer, based on preliminary assumption 1)

Case 2: a 4-option multiple choice test (involving random omission of options including correct answers, based on preliminary assumption 2)

Case 3: 4-choice test (elimination of choices according to test-taker ability level, based on preliminary assumption 3)

Among these, Case 3 demonstrated the highest reliability, followed by the initial case, Cases 1 and 2 yielded similar reliability scores, which were the lowest among the four cases.

Table 5 Reliability Coefficients for Various Test Formats Based on Preliminary Assumptions

Test Case Descriptions	Reliability
Initial Case: The 5-Choice Test Format	0.81
Case 1: A 4-Option Multiple Choice Test (Randomly Eliminated Options Excluding Correct Answer, Based on Preliminary Assumption 1)	0.79
Case 2: A 4-Option Multiple Choice Test (Involving Random Omission of Options Including Correct Answers, Based on Preliminary Assumption 2)	0.79
Case 3: 4-choice test (Elimination of Choices According to Test-Taker Ability Level, Based on Preliminary Assumption 3)	0.87

Table 6 Difficulty and Discrimination Power for Various Test Formats Based on Preliminary Assumptions

Test Case Descriptions	Difficulty	Discrimination
Initial Case: The 5-Choice Test Format	**High**: 6, 24, 82, 31, 91, 88, 110, 74, 89	**High:** 85 29, 66, 88, 106, 32, 22 96
	Low: 103, 50, 3, 78, 35, 76, 80, 102, 75, 107, 110, 41, 15	**Low:** 65, 37, 4, 92 1, 4, 37
Case 1: A 4-Option Multiple Choice Test (Randomly Eliminated Options Excluding Correct Answer, Based on Preliminary Assumption 1)	**High:** 6, 91, 24, 99, 48, 82	**High:** 85, 29, 66, 88, 106, 32, 22, 96
	Low: 103, 50, 3, 78, 35, 76, 80, 102, 75, 107, 41, 15	**Low:** 65, 37, 4, 92, 1
Case 2: A 4-Option Multiple Choice Test (Involving Random Omission of Options Including Correct Answers, Based on Preliminary Assumption 2)	**High:** 6, 94, 24, 99	**High:** 29, 22, 85, 88, 96
	Low: 3, 78, 50, 76, 35, 80, 102, 75, 40, 15	**Low:** 65, 37, 116
Case 3: A 4-Option Multiple Choice Test (Elimination of Choices According to Test-Taker Ability Level, Based on Preliminary Assumption 3)	**High:** 6, 82, 24, 99	**High:** 22, 30, 88
	Low: 3, 78, 50, 76, 35, 80, 102, 75, 15	**Low:** 65, 4, 37, 92

PCA was employed to examine item difficulty and discriminatory power across the four test cases. The analysis revealed that the characteristics of difficulty and discriminatory power generally aligned across all four test cases. No significant differences were observed in item difficulty or discriminatory power among the four scenarios. while Case 4 showed the highest reliability, all four test cases demonstrated comparable levels of item difficulty and discriminatory power, as indicated by PCA.

5.2 Results of Test Score Equating of the 5-Choice and 4-Choice Tests (All 3 Cases Based on Preliminary Assumptions)

Test score equating for the 5-choice and 4-choice multiple-choice tests, conducted for preliminary assumption 1, revealed that the linear equation for score conversion was $y = 0 + 1x$, which features a slope of 1 and an intercept of 0.

Test score equating for the 5-choice and 4-choice multiple-choice tests, conducted for preliminary assumption 2, revealed that the linear equation for score conversion was $y = 3.999 + 0.969x$, which features a slope of 0.97 and an intercept of 4.00 with all values rounded to two decimal places.

Test score equating for the 5-choice and 4-choice multiple-choice tests, conducted for preliminary assumption 3, revealed that the linear equation for score conversion was $y = –9.722 + 1.222x$, which features a slope of 1.22 and an intercept of –9.72, with all values rounded to two decimal places.

The details of these equations, including the rounded values, are summarized in Table 7 below.

5.3 Hypothesis Testing Results

H1: According to Assumption 1, student scores from a 4-choice test, with one option randomly eliminated (excluding the correct answer), will be higher than scores obtained from a 5-choice test.

No significant difference was observed between the 4-choice (4cht1) and 5-choice tests (5ch), with both showing a Mean of 51.5 and Standard Deviation of 11.1. Statistical hypothesis testing was deemed unnecessary.

H2: According to Assumption 2, student scores from a 4-choice test, where one option

Table 7 Results of Equating Student Scores Between The 5-Choice and 4-Choice Tests

Assumption	Linear Equation	Intercept (constant)	Slope
Preliminary Assumption No.1	$y = 0 + 1x$	0.00	1.00
Preliminary Assumption No.2	$y = 3.99 + 0.97x$	4.00	0.97
Preliminary Assumption No.3	$y = –9.722 + 1.222x$	9.72	1.22

Table 8 Summary of Hypothesis Testing Results

Hypothesis	Test Case	Descriptive Statistic		Normality Test (Shapiro-Wilk p-value)	Statistical Test Used	Significance (p-value)	Conclusion
		Mean	SD				
H1	5ch	51.5	11.1	N/A	N/A	N/A	No Difference
	4cht1	51.5	11.1				
H2	5ch	51.5	11.1	< .001	Wilcoxon	< .001	4cht2 > 5ch
	4cht2	53.9	10.7				
H3	5ch	51.5	11.1	< .001	Wilcoxon	< .001	4cht3 > 5ch
	4cht3	53.2	13.5				

is randomly eliminated (including the correct answer), will exceed scores from a 5-choice test.

The 4-choice test (including the correct answer among the eliminated options) (4cht2) outperformed the 5-choice test (5ch), with Mean = 53.9 vs. Mean = 51.5, and a slightly lower standard deviation for the 4-choice test. The Wilcoxon Signed Rank test confirmed the difference as statistically significant. The 4-choice test produced significantly higher scores than the 5-choice test.

H3: According to Assumption 3, scores from a 4-choice test, with option elimination based on the test-taker's ability level, will surpass scores from a 5-choice test.

The 4-choice test (4cht3), with options eliminated based on test-taker ability, also surpassed the 5-choice test (5ch). This test yielded an average score of 53.2 and a higher standard deviation of 13.5. The difference was confirmed as statistically significant by the Wilcoxon Signed Rank test. The 4-choice test produced significantly higher scores than the 5-choice test.

In summary, the 4-choice tests (4cht2 and 4cht3) consistently yielded higher average scores than the 5-choice test (5ch). Among the 4-choice tests, the one tailored to the test-taker's ability (4cht3) had a broader score distribution, evidenced by a higher standard deviation. Hypothesis testing revealed that scores from the 4-choice tests were statistically higher than those from the 5-choice test in all three cases, achieving statistical significance at the 0.05 level. These results are further detailed in Table 8.

6. DISCUSSION

6.1 Objective and Scope of Multiple-Choice Test Design

The primary objective in developing a high-quality test hinge on clearly defining the measurement objectives and systematically structuring the exam to ensure content validity. A significant component of this design process is the number of answer choices included in each multiple-choice question. The selection of answer choices has a dual impact: it not only influences the overall quality of the test but also affects the likelihood of guessing (Esmaeeli et al., 2021; Haladyna & Downing, 1993; Lord, 1977; Nguyen et al., 2021). Beyond these considerations, the number of choices also bears psychological implications, affecting the cognitive load on the test-taker. Reducing the number of choices may lighten the cognitive burden, thereby facilitating more accurate responses. Established guidelines generally recommend four or five options as the optimal range for most academic and internationally standardized tests, such as the TOEFL (Haladyna, Downing & Rodriguez, 2002).

6.2 Criteria for Selecting Answer Choices

After clearly defining the measurement objectives, the next critical step in test design is to establish content validity. This foundation is instrumental in guiding the selection of the number of answer choices for each question. While it is a common practice to offer 4 or 5 choices, it's crucial to understand that this decision is influenced by several factors, such as the complexity of the content and the cognitive level of the target audience (Haladyna, Downing & Rodriguez, 2002).

Each multiple-choice question typically comprises one correct answer and one or more distractors. The quality of these distractors is integral to the test's reliability and validity (Adams, 1964; Kumar et al., 2023; McDaniel & Little, 2019; Rodriguez, 2005). Poorly designed distractors compromise both these aspects, underscoring the need for meticulously crafted options that provide an adequate challenge to the test-taker (Lekakul, 2016).

It's worth noting that the optimal number of answer choices may vary depending on the target audience. For example, tests aimed at younger children often feature 2–3 choices to reduce confusion and cognitive overload, allowing for a more accurate assessment of their knowledge.

6.3 The Art and Science of Option Crafting

The meticulous development of answer options is fundamental to both the reliability and validity of a test. Research supports that a high-quality test relies heavily on well-designed distractors (Thurstone, 1931). Examples of effective distractors include options that are plausible yet incorrect, often capitalizing on common misconceptions or frequent errors. For instance, in a history test asking for the first U.S. President, an effective distractor might be "Benjamin Franklin," as he is a prominent figure from the same era but was never President.

In contrast, poor distractors not only waste valuable test-taking time but also undermine the test's integrity by making the correct answer too obvious. For example, in a math question asking to solve for X in $X+2 = 5$, a poor distractor would be "10," as it is far removed from the plausible range of correct answers. Such distractors compromise both the test's reliability and its validity (Thurstone, 1931).

It is, therefore, imperative to craft distractors that closely resemble the correct answer in both form and content, challenging test-takers who have not mastered the subject matter (Greving et al., 2023; Lekakul, 2016). This nuanced approach ensures that the test effectively discriminates between varying levels of knowledge and understanding.

6.4 Optimal Number of Choices

Current research, such as studies by Chutinuntakul, Wongnam, & Panhoon (2018), underscores the benefit of eliminating ineffective options to enhance test quality. A well-designed four-choice test can be nearly as effective as a five-choice one, particularly when crafting a fifth plausible distractor proves to be challenging (Sirirungphan, 2016). Additionally, using more than five options does not necessarily improve the test's ability to differentiate between levels of knowledge.

6.5 Implications of Choice Reduction

Transitioning from a five-choice to a four-choice format can yield several advantages, including reduced test-taking stress and lower administrative costs (Chutinuntakul, Wongnam & Panhoon, 2018). Such a shift could positively influence educational policies aimed at reducing student attrition rates and improving performance (Aydin, Öztürk, Büyükköse, & Sönmez, 2019; Budiman, 2018). This aligns with psychological theories indicating that fewer choices may reduce cognitive load and facilitate quicker decision-making.

This discussion outlines the importance of various aspects of multiple-choice test design, from the initial stages of defining objectives to the final implementation of the test. Understanding the optimal number of choices and the art of crafting effective distractors is crucial for achieving high test reliability and validity.

ACKNOWLEDGEMENTS

The Institutional Research and Information Division Fund of Sukhothai Thammathirat Open University supported this research.

REFERENCES

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, *74*, 1-36. https://doi.org/10.18637/jss.v074.i08

Angoff, W. H. (1984). (1971). Scales, norms, and equivalent scores. In RL Thomdike (Ed.), Educational measurement (pp. 508-600). Washington DC: American Council on Education.

Aydin, S., Öztürk, A., Büyükköse, G. T., Er, F., & Sönmez, H. (2019). An Investigation of Drop-Out in Open and Distance Education. *Educational Sciences: Theory and Practice*, *19*(2), 40-57. http://dx.doi.org/10.12738/estp.2019.2.003

Chutinuntakul, Sasithorn., Wongnam, Pirat ., & Panhoon, Sompong. (2018). The comparison of test scores derived through Kernel Equating and IRT Equating Methods under varied conditions. *STOU Education Journal, 11(1),* 294-306.

Effatpanah, F., & Baghaei, P. (2023). Kernel Smoothing Item Response Theory in R: A Didactic. *Practical Assessment, Research, and Evaluation*, *28*(1), 7.

Esmaeeli, B., Esmaeili Shandiz, E., Noroozi-asl, S., Shojaei, H., Pasandideh, A., Khoshkholgh, R., ... & Barkhordari Ahmadi, F. (2021). The optimal number of choices in multiple-choice tests: a systematic review. *Medical Education Bulletin*, *2*(3), 253-260. doi: 10.22034/MEB.2021.311998.1031

Greaney, V., & Kellaghan, T. (Eds.). (2008). *Assessing national achievement levels in education* (Vol. 1). World Bank Publications.

Greving, S., Lenhard, W., & Richter, T. (2020). Adaptive retrieval practice with multiple-choice questions in the university classroom. *Journal of computer assisted Learning*, *36*(6), 799-809. https://doi.org/10.1111/jcal.12445

Greving, S., Lenhard, W., & Richter, T. (2023). The testing effect in university teaching: Using multiple-choice testing to promote retention of highly retrievable information. *Teaching of Psychology*, *50*(4), 332-341. https://doi.org/10.1177/00986283211061204

Guo, H., Zu, J., & Kyllonen, P. (2018). A Simulation-Based Method for Finding the Optimal Number of Options for Multiple-Choice Items on a Test. *ETS Research Report Series*, *2018*(1), 1-17. https://doi.org/10.1002/ets2.12209

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item?. *Educational and psychological measurement*, *53*(4), 999-1010. https://doi.org/10.1177/0013164493053004013

Kanchanawasi, Sirichai. (1998). Equalizing scores between tests (Test Equating). Bangkok: Academic Textbooks and Documents Center, Faculty of Education, Chulalongkorn University.

Kumar, A. P., Nayak, A., Shenoy, M., & Goyal, S. (2023). A novel approach to generate distractors for multiple choice questions. *Expert Systems with Applications*, *225*, 120022. https://doi.org/10.1016/j.eswa.2023.120022

Lekakul, Anupap. (2016). *Creating a multiple-choice exam.* Songkhla: Prince of Songkla University.

Lord, F. M. (1977). Optimal number of choices per item: A comparison of four approaches. *Journal of Educational Measurement*, 33-38.

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ, Lawrence Erlbaum Ass.

McDaniel, M. A., & Little, J. L. (2019). Multiple-choice and short-answer quizzing on equal footing in the classroom: Potential indirect effects of testing.

McGuire, B. (2012). *KernSmoothIRT: An R Package allowing for Kernel Smoothing in Item Response Theory* (Doctoral dissertation, Montana State University).

Nguyen, T., Bui, T., Fujita, H., Hong, T. P., Loc, H. D., Snasel, V., & Vo, B. (2021). Multiple-objective optimization applied in extracting multiple-choice tests. *Engineering Applications of Artificial Intelligence*, *105*, 104439. https://doi.org/10.1016/j.engappai.2021.104439

Panidvadtana, Panida. (2019). Concept and application of multidimensional irt equating procedure. *Yanasangworn Institute Research Journal, 10(20),* 318-329.

Petersen, N.S., Marco, C.L., & Stewart, E.E. (1982). *A test of the adequacy of linear Score Problems*. Hillssdale, N.J.: Erlbaun.

Rajlic, G. (2020). Visualizing items and measures: An overview and demonstra-

tion of the kernel smoothing item response theory technique. *The Quantitative Methods for Psychology, 16*(4), 363-375. http://doi.org/10.20982/tqmp.16.4.p363

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational measurement: issues and practice, 24*(2), 3-13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Sirirungphan, Chusak. (2016). Techniques for writing multiple-choice exams. In Narong Theepprachai (ed.), Handbook of measurement and evaluation in distance education systems. For undergraduate programs Sukhothai Thammathirat Open University (pp. 78-84). Sukhothai Thammathirat Open University Press.

Thurstone, L.L. (1931). *The reliability and validity of tests: Derivation and interpretation of fundamental formulae concerned with reliability and validity of tests and illustrative problems*. Ann Arbor, MI: Edwards Brothers.