THE UNITED REPUBLIC OF TANZANIA

NATIONAL EXAMINATION COUNCIL OF TANZANIA

DIPLOMA IN SECONDARY EDUCATION EXAMINATION

762 EDUCATIONAL RESEARCH, MEASUREMENT AND EVALUATION

Time: 3 Hours. **ANSWER** **Year: 2003 a.m.**

Instructions

1. This paper consists of sections A, B and C.
2. Answer **all** questions in sections A, **two (2)** questions from section B and **one (1)** question from section C.
3. Question **11** is **compulsory**.
4. Section A carries 36 marks, section B carries 40 marks and section C carries 24 marks
5. Cellular phones and unauthorized materials are **not allowed** in the examination room.
6. Write your **Examination Number** on every page of your answer booklet(s).

Prepared by Maria Marco for TETEA

1. Define educational evaluation.

Educational evaluation is the systematic process of collecting and analyzing information to determine the extent to which educational goals and objectives have been achieved.

It involves measuring learners' performance, interpreting the results, and making decisions about teaching, learning, and educational programs.

It is used to improve instruction, assess curriculum effectiveness, and make informed decisions on student progression or certification.

It provides feedback to teachers, learners, and policymakers for continuous improvement in education.

2. Mention four characteristics of a good research title.

A good research title is clear and concise, avoiding unnecessary words while clearly indicating the study's focus.

It is specific, pointing directly to the variables, population, and setting under study.

It is informative, giving the reader an immediate idea of the research problem and scope.

It is engaging and relevant, attracting interest while reflecting the true content of the research.

3. State four reasons for sampling in educational research.

Sampling reduces the cost of research because studying a subset of the population is cheaper than studying the entire population.

It saves time, allowing researchers to collect and analyze data more quickly.

It makes research feasible when populations are too large to study in full.

It allows in-depth study of a representative portion of the population, making it easier to control quality and accuracy.

4. Give four limitations of using essays as a test format.

Essays take a long time to mark, making them less practical for large groups of students.

They are subject to scorer bias, as different examiners may grade the same response differently.

They may encourage memorization of content rather than demonstrating higher-order thinking.

They limit content coverage because only a few questions can be asked within the available time.

5. List three roles of a literature review in a proposal.

A literature review identifies gaps in existing research, guiding the focus of the new study.

It provides a theoretical background that supports the research problem and objectives.

It helps refine research questions and methodology by learning from previous studies.

6. State four threats to internal validity in quasi-experiments.

History effects occur when external events influence participants during the study.

Maturation effects result from natural changes in participants over time, unrelated to the experiment.

Testing effects occur when repeated testing influences participants' performance.

Instrumentation effects arise when changes in measurement tools affect the results.

7. Mention four principles for constructing multiple-choice items.

Each item should have one clearly correct answer and plausible distractors.

The stem should be clear and free from unnecessary information.

Distractors should be similar in length and structure to avoid clues.

Items should be based on important learning objectives, not trivial facts.

8. Explain four purposes of a table of specifications in test construction.

It ensures that test items cover all relevant topics and skills proportionately.

It balances questions across different cognitive levels, such as knowledge, application, and analysis.

It guides item writers, making test construction systematic and consistent.

It improves the validity of the test by aligning items with instructional objectives.

---

9. Give two situations where median is preferred to mean in reporting results.

When the data contains extreme scores that would distort the mean, the median is preferred.

When dealing with ordinal data such as class rankings, the median is more meaningful than the mean.

10. (a) Define measurement error and distinguish between random error and systematic error.

Measurement error is the difference between the observed score and the true score of a test-taker.

Random error occurs unpredictably and affects scores inconsistently, often caused by temporary factors such as distractions.

Systematic error is consistent and predictable, often caused by flaws in the test design or administration.

(b) Explain four practical strategies a test developer can use to reduce measurement error in school-based examinations.

Use clear and unambiguous wording to avoid misinterpretation.

Train examiners to ensure consistent administration and scoring.

Pilot test items to identify and correct weaknesses before the final test.

Standardize testing conditions, such as time limits and instructions, for all examinees.

11. (a) The following are Biology scores for 14 students:
38, 42, 45, 50, 55, 57, 60, 62, 65, 67, 70, 72, 78, 85

(i) Calculate the median.

Ordered scores: 38, 42, 45, 50, 55, 57, 60, 62, 65, 67, 70, 72, 78, 85

Number of scores = 14 (even)

Median = (7th score + 8th score) ÷ 2 = (60 + 62) ÷ 2 = 122 ÷ 2 = 61

(ii) Calculate the mean (nearest whole number).

Sum of scores = 856

Mean = 856 ÷ 14 = 61.14 ≈ 61

(iii) Calculate the variance and standard deviation (nearest whole number).

Step 1: Deviations from mean (61), squared:

$(38-61)^2 = 529$

$(42-61)^2 = 361$

$(45-61)^2 = 256$

$(50-61)^2 = 121$

$(55-61)^2 = 36$

$(57-61)^2 = 16$

$(60-61)^2 = 1$

$(62-61)^2 = 1$

$(65-61)^2 = 16$

$(67-61)^2 = 36$

$(70-61)^2 = 81$

$(72-61)^2 = 121$

$(78-61)^2 = 289$

$(85-61)^2 = 576$

Sum of squares = 2540

Variance = 2540 ÷ 14 = 181.43 ≈ 181

Standard deviation = √181 ≈ 13.45 ≈ 13

(iv) Transform X = 85 and X = 38 to T-scores using mean 50 and standard deviation 10.

Formula: T = 50 + 10 × ((X − Mean) ÷ SD)

For X = 85: Z = (85 − 61) ÷ 13 ≈ 1.85

T = 50 + (1.85 × 10) = 50 + 18.5 = 68.5 ≈ 69

For $X = 38$: $Z = (38 - 61) \div 13 \approx -1.77$

$T = 50 + (-1.77 \times 10) = 50 - 17.7 = 32.3 \approx 32$

(b) Interpret the two T-scores in part (iv) for a head teacher.

The student with a T-score of 69 performed well above the group average, indicating high achievement compared to peers.

The student with a T-score of 32 performed well below the group average, indicating the need for remedial support.

12. A researcher intends to evaluate the effect of a formative assessment intervention on Form Three Mathematics achievement across 12 schools in two regions.

(a) Propose an appropriate research design and justify the choice.

The most suitable design would be a quasi-experimental pre-test and post-test control group design. This is because it allows comparison between groups receiving the intervention and those that do not, even when random assignment of schools is not possible.

It enables the researcher to measure changes in achievement attributable to the formative assessment intervention by comparing pre-test and post-test results.

It is practical in real school settings where randomization may disrupt normal teaching schedules.

(b) Specify population, sampling frame, sampling technique, and sample size with reasons.

The population would include all Form Three students in secondary schools within the two selected regions.

The sampling frame would be the complete list of these schools, obtained from the regional education offices.

The sampling technique would be stratified random sampling, ensuring representation from both urban and rural schools in each region.

A sample size of 12 schools, with equal numbers in the experimental and control groups, would be used to maintain balance and allow for meaningful statistical comparison.

(c) Describe four data quality assurance procedures you would apply from instrument development to fieldwork.

Pilot testing the instruments to ensure clarity, appropriateness, and reliability before the main study.

Training data collectors thoroughly on administration procedures to maintain consistency.

Using standardized instructions and timing for both pre-test and post-test across all schools.

Double-checking data entry and cleaning to remove errors before analysis.

(d) Outline a data analysis plan linking each research question to suitable statistics.

For the question on differences in pre-test scores, use independent samples t-test.

For the question on post-test differences, use ANCOVA, controlling for pre-test scores.

For measuring improvement within groups, use paired samples t-test.

For examining the relationship between intervention participation and achievement, use regression analysis.

13. A new reading comprehension test is being validated for Form Two students.

(a) Explain how you would establish content validity using expert judgment and a content validity index.

Select a panel of subject experts to review each test item for relevance, coverage, and alignment with the syllabus.

Ask experts to rate each item on a scale, for example, from 1 (not relevant) to 4 (highly relevant).

Calculate the Content Validity Index (CVI) for each item by dividing the number of experts rating it as relevant (3 or 4) by the total number of experts.

Items with low CVI values would be revised or discarded to improve overall test validity.

(b) Explain how you would establish construct validity using exploratory factor analysis, including assumptions to check.

Administer the test to a large sample of students representative of the population.

Check assumptions such as adequate sample size (at least 5–10 participants per item) and sampling adequacy using the Kaiser-Meyer-Olkin (KMO) test.

Conduct factor extraction to see if items group into factors that match the intended sub-skills, such as vocabulary, inference, and main idea recognition.

Remove items that load poorly on their intended factors or load on multiple factors.

(c) Explain how you would establish criterion-related validity using both concurrent and predictive approaches, specifying appropriate external criteria.

For concurrent validity, administer the new test alongside an established standardized reading test and compute the correlation between the two sets of scores.

For predictive validity, administer the new test early in the academic year and compare scores with students' end-of-year reading performance.

A strong positive correlation in both cases would indicate good criterion-related validity.

14. Tanzania plans to report school performance using standardized scores rather than raw scores.

(a) Explain the logic of standard scores (z, T, stanines) and how they enable fair comparisons across forms and years.

Standard scores express performance in terms of deviation from the mean, adjusted for the spread of scores, making them comparable across different tests.

Z-scores show how many standard deviations a score is from the mean.

T-scores rescale z-scores to avoid negative values and decimals, making them easier to interpret.

Stanines group scores into nine broad categories, simplifying interpretation while maintaining comparability.

These measures remove the effects of differences in test difficulty, allowing fair comparisons across years or different forms.

(b) Discuss four risks of misinterpretation or misuse of standardized scores at school, district, and national levels, and propose practical safeguards for each.

One risk is overemphasis on rankings, which can create unhealthy competition between schools. This can be mitigated by also reporting growth scores.

Another risk is misinterpreting small differences in scores as meaningful. This can be addressed by training educators in statistical interpretation.

A third risk is ignoring contextual factors like resources, leading to unfair comparisons. This can be reduced by pairing scores with contextual data.

A fourth risk is using scores for purposes they were not intended for, such as teacher evaluations. This can be avoided by setting clear policies on score usage.

15. You are tasked to overhaul a national high-stakes examination to improve reliability and fairness.

(a) Propose four structural changes to the test blueprint and item formats that would increase reliability without inflating test length excessively.

Increase the number of items that assess each skill area to improve score consistency.

Balance the use of objective items, such as multiple-choice, with structured response items for breadth and depth.

Ensure that item difficulty is distributed to match the ability range of test-takers.

Include anchor items across test versions to maintain comparability over time.

(b) Propose four policy or operational changes in administration, scoring, and reporting that would reduce bias and enhance equity, explaining the mechanism for each.

Provide clear, standardized administration manuals to all examiners to ensure uniform procedures.

Use multiple independent scorers for subjective items, applying moderation to reduce scorer bias.

Translate test instructions into all relevant languages to minimize disadvantage due to language barriers.

Report both raw and scaled scores to give a fuller picture of student performance.

16. An experimental study will randomize classrooms to a digital learning tool in English language teaching.

(a) Identify and discuss four ethical issues specific to this cluster-randomized trial in Tanzanian schools.

Consent must be obtained from both school authorities and individual participants to respect autonomy.

Equity concerns may arise if only some students receive the digital tool, creating unequal opportunities.

Data privacy must be protected, especially since digital tools may collect student information.

Potential disruption to normal teaching schedules must be minimized to avoid disadvantaging any group.

(b) For each issue, propose concrete mitigation measures that are realistic for public-school settings and aligned with local regulations.

Use written consent forms explained in simple language to all stakeholders.

Rotate the intervention so that all classes eventually benefit from the tool.

Ensure all data is anonymized and stored securely, following national data protection laws.

Schedule digital tool sessions during non-critical teaching times to avoid curriculum coverage loss.