MAKING ACCURATE
&
FAIR TESTS

AS&K
Applied Skills & Knowledge, LLC
"A Leader in Measuring Success"

June, 2001

Introduction

Making tests that fairly and accurately measure what you want to measure involves a series of simple steps and some quality control checks. While these steps and checks can be simple for many types of tests, at an advanced level, they include the use of special mathematical procedures from the field of psychometrics. Tests are used for a wide variety of purposes. The decisions made from test results must be based test results that are as fair and accurate as possible. This is the challenge for any test maker – to make a test whose results lead to fair and accurate decisions. These decisions include:

- Hiring a salesperson who meets or exceeds sales targets
- Assessing mastery at the end of training
- Hiring a police officer who makes sound decisions and protects the public
- Promoting a manager to lead a team of software engineers
- Selecting managers for a special leadership training program
- Certifying that customer service representative knows the company products and services
- Waiving training for a programmer who may have already mastered visual basic

These are important decisions. Well-developed tests help you make decisions such as these and others more fairly and accurately. The accuracy of the test results and the confidence you can have in using the results to make better decisions depends upon (1) properly following the test making steps and (2) conducting the quality control checks. If you follow the steps you are likely to have good quality control results – but there are no guarantees.

The quality control checks are needed to assess the confidence you can have in your test results. If the quality control results are not good then you cannot be sure that using the results to make decisions is a good idea. However, if the quality control checks are good, the test results will help you make more accurate decisions. There are three types of quality control checks. They are:

- Reliability
- Validity
- Fairness

The steps for making a fair and accurate test are listed below. Each step contributes to the accuracy and fairness of the test – none can be skipped. And the steps must be followed in the order listed.

- Define what you want to measure
- Make a test budget
- Write the test questions
- Set a passing score
- Perform quality control checks
- Publish the test

This document describes how to complete each of these steps and describes the quality control checks. Some simple quality control check techniques are provided.

Good Housekeeping Criteria for Tests

There are three types of quality control checks for testing – reliability checks, fairness checks, and validity checks. These checks serve as the "Good Housekeeping" criteria for the quality of tests.

Reliability - an examinee obtains the same score on a second testing, assuming the person's skill level has not changes and the two tests are equivalent and not the same tests.

Validity – "….. is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment." (Messick, 1995, p. 741)

Fairness – examinees with the same skill level but different subgroup membership receive the same score. Important subgroups include males and females, majority and minority, handicapped and non-handicapped.

Write the Test Questions - Multiple Choice Items

In a multiple choice item, the examinee selects the one correct or best answer from a list or three or more possible answers. The incorrect answers represent common errors. The question part of the item is called the **stem** and the choices are called **alternatives**. The incorrect choices are called **distractors**.

There are some **advantages** of multiple-choice items, including:

- They are easy to administer and can be answered quickly by the examinee.
- They are easy to score and can be scored electronically.
- They are easy to handle statistically.
- They have good reliability (i.e., an examinee chooses the same answer again if the same item is presented on two separate occasions).
- They permit the measurement of several levels of learning because the incorrect answers can be selected to attract examinees of varying degrees of skill.

There are some **disadvantages** of multiple-choice items, including:

- Though better than true-false items, guesses will be correct a certain percentage of time. For example, if there are four alternatives, one-fourth of the guesses will be correct.
- The development of good multiple-choice items with "attractive" incorrect answers is difficult.

Multiple-choice items can measure a wide range of skill complexity including simple recall of facts to application, analysis and synthesis of information.

While there is no absolute rule about how many alternatives there should be, four or five are good choices. With more than five alternatives, it may be difficult for the examinee to keep all of the alternatives in mind simultaneously. Also, the preparation of more than three or four distractors is difficult.

Rules for Writing Multiple-Choice Items

Writing the Stem

1. Start the MC items with a full statement of the question (stem) or an open-ended statement. Either form is acceptable. The guiding principle is economy of words. Whichever approach results in the least amount of reading time and least difficulty for the examinee is preferred.

Which of the following is the best lubricant to use on Part A of Equipment XY?

OR

The best lubricant to use on Part A of Equipment XY is _____ ?

2. The examinee should be able to grasp the central point of the question from reading the stem. The stem should not be cut off too soon but focus the examinee on the essential concept being measured. "A psychological test …….. " followed by the alternatives provides too little information about the central point of the item.

Poor: If an automobile ignition will not turn the engine over _____

a. You should charge the battery
b. You should first try the horn.
c. The ignition system definitely has a short.
d. You should first check the level of the fluid in the battery

Better: If an automobile ignition will not turn over, which of the following actions should you normally take first?

a. Make sure the battery cables are not worn out.
b. Try the horn to make sure it is working.
c. Check that there is enough fluid in the battery.
d. Check the ignition system for shorts.

3. The language used in the stem should be precise but not technical unless the item is intended to measure the examinee's skill in handling highly technical material.

4. One of the most important features of a good item is economy of words. Stems that provide unneeded information such as an author's name or noting the controversial nature of the concept being tested, are not economical. Avoid unnecessary information in the stem.

Poor: Without explanation, a regular employee of the company fails to report for work following a period of 10 days annual leave on a trip to London, Paris, and Rome while on a honeymoon. After how many calendar days of unauthorized absence is the employee removed for abandonment of position?

Better: Without explanation, a regular employee of the company fails to report to work. After how many calendar days of unauthorized absence is the employee removed for abandonment of position?

5. Do not begin the stem with a pronoun such as "it." This causes the sentence structure to be awkward.

Poor: It is the most essential characteristic of a psychological test ……

Better: The most essential characteristic of a psychological test is ……..

6. Include in the stem as many words as required to avoid redundancy in the alternatives.

Poor: How many days are there in February?

a. 27, with the exception of leap year.
b. 28, with the exception of leap year.
c. 29, with the exception of leap year.
d. 30, with the exception of leap year.
e. 31, with the exception of leap year.

Better: Except for leap year, how many days are there in February?

a. 27
b. 28
c. 29
d. 30
e. 31

7. Generally, questions should be stated in positive form. Avoid the use of negatives such as not, least, none, etc. Particularly avoid double negatives.

8. When a negative word is used on the stem it should be emphasized by capitalizing (NOT), using italics (*not*) or underlining (<u>not</u>).

9. When a common alternative term is used, it is mentioned (usually in parenthesis) after the original term. This avoids testing vocabulary and focuses the question on the concept intended for measurement.

Poor: The Rorschach test is intended to measure which of the following?

Better: The Rorschach test (inkblot test) is intended to measure which of the following?

10. When acronyms are used, the complete name is spelled out the first time it is used.

Poor: Which of the following best expresses the advantages of WYSIWYG?

Better: Which of the following best expresses the advantages of What You See Is What You Get (WYSIWYG)?

11. Questions should be independent. A question is independent if it, (1) is not given away by information in another question and (2) does not depend on getting the correct answer to another question.

In the examples below, the clue is obvious. In the first question the examinee is told that the Work Order is a copy of the Production Order. There should be no trouble answering the second question because the answer to the second question is contained in the first question's stem.

What color is the Work Order copy of a Production Order?

a. Salmon
b. Green
c. Yellow
d. Blue

Which of the following is a copy of the Production Order?

a. BSN
b. Work Order
c. Request for Contact Investigation
d. Parts Replacement Request

12. In preparing a multiple-choice item, do not simply repeat problems or examples that were used in training materials provided to the examinees. Use fresh examples to test the examinee's understanding and problem solving skills.

13. Write problem-solving questions. Multiple choice questions have been criticized because it has been claimed that they measure only surface level knowledge.

Poor: What number can you call to obtain information about Omega Accounts?

a. (800) 854-7154
b. (800) 262-4636
c. (800) 535-5549
d. (800) 345-8265

Better: What is the proper procedure to follow when a client needs a new account number because the account was transferred to Investor Services?

a. Give the client the new account number after security questions are answered properly.
b. Tell the client to make the request for the new account number in writing.
c. Send through US Mail the new account information to the client's address of record.
d. Fax, email or send through US Mail the new account information to the client's contact information of record.

Rules for Writing Multiple-Choice Items

Writing the Distractors

1. Each of the distractors and the correct answer should follow from the stem conceptually and grammatically. In the first example, a cue is provided for the correct answer because it is the only alternative that logically follows from the stem. In the second example, the correct answer is cued by virtue of the fact that it does not follow the stem.

   What is the factor that has gone far ahead of the possibilities of treatment in any system of correctional institution?

   a. Custody
   b. Diagnosis (Correct Answer)
   c. Maintenance
   d. Discipline

   What sort of problems should the department head take up formally with the Board of Supervisors without reference to the County Administrator?

   a. Any departmental problem
   b. Any policy problem, if the matter is concerned exclusively with intradepartmental matter and does not have "administrative" implications.
   c. Any non-policy problem of primary concern to his department.
   d. Any problems that are so urgent that they cannot wait to be cleared with anyone.
   e. All Department Head – Board of Supervisor matters should normally be taken up with the County Administrator first. (Correct Answer)

2. Each distractor should be a common misconception or an error of the novice or poor performer. There should be no "throw-away" distractors such as distractor D in the example below.

   Poor: The Dow Jones Industrial average is best described by which of the following?

   a. The current average daily market values of the common stock for selected U.S. private industries.
   b. The current average daily market value of the common and preferred stock for selected U.S. industries.
   c. The current average daily market value of the common and preferred stock for selected U.S. transportation, utility and other industries.
   d. The bond yield in a company's portfolio.

Better: The Dow Jones Industrial average is best described by which of the following?

a. The current average daily market values of the common stock for selected U.S. private industries.
b. The current average daily market value of the common and preferred stock for selected U.S. industries.
c. The current average daily market value of the common and preferred stock for selected U.S. transportation, utility and other industries.
d. The current five day average market value of the common stock for selected U.S. transportation, utility, and other industries.

3. The choice of words for the correct answer compared to the alternatives can be a cue to the examinee. If common jargon or technical terms are used in the correct answer and layman's language in the distractors or vice versa, the examinee mayh be led to the correct answer by the difference in language usage.

4. Because a sufficient supply of good distractors is difficult to generate, a common flaw is to write short alternatives, compared to the correct answer. It is not necessary that all alternatives be the same length, but no pattern should be apparent.

5. The correct answer can be cued by repeating words from the stem in the correct answer alone. If words must be repeated, all alternatives should include the repeated words.

Poor: Many children who are so strongly blocked that they cannot establish a minimal relation in individual therapy are helped by which of the following?

a. Occupational treatment
b. Group therapy (Correct Answer)
c. Physical activities
d. Activity training
e. Music

Better: Many children who are so strongly blocked that they cannot establish a minimal relation in individual therapy are helped by which of the following?

a. Occupational therapy
b. Group therapy (Correct Answer)
c. Physical therapy
d. Activity therapy
e. Music therapy

6. When the item writer prepares the correct answer first and then the distractors, the writer may use pronouns in the distractors that have referents in the correct answer. Avoid the use of pronouns. In so doing, item writing is commonly found to be more redundant than ordinary prose.

7. An easy distractor to prepare is the opposite of the correct answer. This condition, a correct-incorrect (see A and C below) pairing usually signals that one of the pair is the correct answer. The choice of answers is reduced to two rather than one choice from among four or five. If the use of an opposite is necessary, include another pair of opposites.

One classification of information necessary for budget preparation should be made on the basis of the services and commodities to be purchased and of the obligations to be met by the governmental agency. Which statement is true about such a classification? It would …..

a. Include expenditures for salaries, heat, rent, and supplies.
b. Exclude considerations of new buildings and equipment.
c. Exclude expenditures for salaries, heat, rent, and supplies.
d. Include income that had been provided for but that has not yet been raised. (Correct answer)

One classification of information necessary for budget preparation should be made on the basis of the services and commodities to be purchased and of the obligations to be met by the governmental agency. Which statement is true about such a classification? It would …..

a. Include expenditures for salaries, heat, rent, and supplies.
b. Exclude considerations of new buildings and equipment.
c. Include services that have been committed but not yet received.
d. Include income that had been provided for but not yet been raised. (Correct answer)

8. Overlapping alternatives arise when one option includes concepts or portions of a aresponse found in one or more alternatives. This flaw occurs most frequently when the alternatives are numerical. For example, one alternative may be 50% and another may be less than 60%. These flaws are easily corrected once identified.

9. A flaw that is difficult to detect and occurs commonly is the generation of distractors what possess the same concept but in different words. Re-expressing a distractor permits the examinee to eliminate the two common alternatives leaving only two or three alternatives from which to choose.

Poor: Which one of the following procedures should a supervisor follow to get the best work possible from a new employee?

a. Let the employee figure the job out for him/herself.
b. Avoid bothering the employee if she/he has made the right start on the job.
c. Follow-up to see how the employee is doing after his/her initial training. (Correct answer)

d. Let his/her fellow employees tell him/her what to do and how to do it.

Better: Which one of the following procedures should a supervisor follow to get the best work possible from a new employee?

e. Give him/her only the general principles to follow and let him/her figure out the details.
a. Avoid bothering the employee if she/he has made the right start on the job.
b. Follow-up to see how the employee is doing after his/her initial training. (Correct answer)
c. Emphasize the mistakes others have frequently made in doing the job.

10. The use of "none of the above" or "all of the above" is generally not recommended. The examinee only needs to establish that any two alternatives are correct to determine that the "all of the above" alternative is the correct choice. When the question tests computational skills, "none of the above" may be used. Examinees often try out each alternative, working backwards to discover the correct numerical response. Use of "none of the above" is advisable here.

11. While there is no absolute rule as to how many alternatives responses there should be, four or five are good choices. With more than five alternatives, it may be difficult for the examinee to keep all the choices in mind simultaneously. Also, preparing more than three or four good distractors is very difficult.