BSD Certification Group Inc.

Psychometrics and Exam Construction

December, 2006

Table of Contents

Introduction

The primary goal of any certification organization is to accurately assess a specified set of knowledge and skills. For example, the BSD Certification Group Inc.'s mission is to certify system administration skills on BSD operating systems.

But how is that set of knowledge and skills defined? And how can the accuracy of the assessment be determined? Through the science of psychometrics.

Psychometrics isn't something that just occurs once; it is an ongoing process that begins in the planning stages of the certification and continues after the exam is available. In this document, Sandra Dolan, a psychometric consultant for the BSD Certification Group Inc., describes the psychometric process with a focus on the question creation stage of a certification exam.

In ***Objective Measurement***, she provides a bird's eye view of the science of psychometrics by comparing it to other systems of measurement.

Properties of Objective Measurement explains three fundamental goals used in the creation of assessment exams.

In ***Planning the Test***, Sandra describes the steps used to create the Table of Specifications for the BSDA. For the BSDA, the Task Analysis Survey Report was used to determine the required skillset which in turn was used in the creation of the BSDA Exam Objectives.

The ***General Item Writing Principles*** section is used by SMEs (Subject Matter Experts) when they create exam questions (items). As you read through this section, you'll get a better understanding of the difference between well worded questions which fairly assess skill and poorly worded questions which try to confuse or frustrate the test taker.

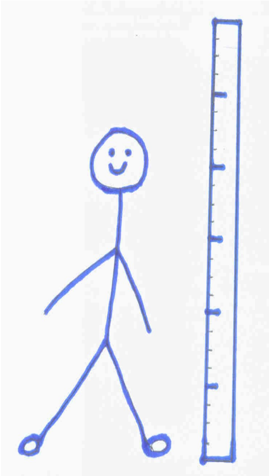
The ***Appendix*** provides a summary checklist of the principles used in the creation of good exam questions.

It is our hope that the examples provided within this document help to de-mystify the science of psychometrics and to explain why psychometrics is an important component in the creation of certification exams.

Objective Measurement

The test as a measurement tool.

A test is just a tool to measure the amount of Knowledge, Skills and Abilities (KSAs) a person has in some area. It is often difficult to comprehend a quantity of knowledge, since it seems to be so abstract. But in actuality, any quantity measurement is just an abstraction.



For instance, the measurement of height in inches/feet appears on the surface to be a real and concrete measurement. But if you think about it, the inch was simply created and defined by people. There is no naturally occurring inch and there really are no natural units of measurement at all. One cannot hold an inch, and it really is just an abstraction that is generally agreed upon. It is this general agreement that makes the inch a useful measurement tool. It is this common frame of reference that makes a unit of measurement functional and useful.

What is Psychometrics?

Psychometrics is the field of study concerned with the theory and technique of psychological measurement, which includes the measurement of knowledge, abilities, attitudes and personality traits. Psychometrics is applied widely in a variety of testing areas including educational, intelligence, psychomotor and personality assessments.

Within a common frame of reference, height begins to take on more meaning than just inches. When one marks a ruler with inches/feet, there is the intrinsic meaning that something that barely reaches the 4' mark is shorter than an object that exceeds the 6' mark. The common frame of reference becomes a measure of height in inches/feet, a continuum from shorter to taller. It is a continuum because, in order for someone to be 6' tall, they must also be at least 4' tall. In other words, one must pass ALL the marks up to 6' mark.

One can think of the inch/feet marks on the ruler as the test items, with each item asking if the top of the person's head reaches that mark. You can then move progressively up the ruler and when the answer becomes "no," you then know the person's height. If you were measuring the height of a group of adults, you might want to use a ruler that focuses in on the range of adult heights. One need not ask at every inch from 1" to 10,000" if the head reaches that mark. In order to save time, you might start at the 48" mark, end at the 84" mark, and ask if the head reaches each mark in between. The marks included are just a sampling from the universe of marks. There are many heights less than 48" and greater than 84", but since the general range of adult height is within this range, this is the area we want to concentrate our questions in. If we were measuring newborn height, we might wish to ask if the babies reach inch marks from 10" to 30". But in either case, the meaning of the inch and height does not change. The frame of reference is stable.

Similarly, a test must take on more meaning than the items it contains. A test's frame of reference is its items, and like the ruler, the items must be able to be placed in a continuum, from easier items to more difficult ones. It is logical that, if a “smarter” candidate gets the harder items correct, he should get most/all of the easier items right, just as when a taller person also reaches all the inch marks below his height. A candidate who ONLY gets a few easy items right would not be expected to get the hardest items correct, just as when a shorter person reaches the lower inch marks, but not the higher ones. As a candidate answers items on this difficulty continuum, one can estimate his knowledge to be at the point where he can no longer answer the items correctly. As in the height example, a certification exam contains questions that are only a sampling from the universe of possible items in that field. We may choose to target the items within a particular difficulty range for our purposes. If we were creating a junior sysadmin exam, we might want to have items on our test that are more basic. For a senior administrator exam, we’d want to test more complex and higher order knowledge, skills and abilities at a greater difficulty level.

We must keep in mind that tests are NOT given to determine if a candidate knows the answers to particular items, but rather we are interested if the candidate understands the concepts that the items tap. We are not interested in the items themselves, but rather in the underlying latent trait or concept that the item investigates. Again, the test should take on more meaning than the items it contains.

Properties of Objective Measurement

There are three major properties of objective measurement.

1) person measures are test-independent: this means that, if a person takes test A and test B on the same day, and if both tests measure the same thing, that person should receive the same ability measure for both tests, regardless of test difficulty

2) item difficulties are sample-independent: this means that the relative ordering of item difficulties (from easy to hard) should NOT significantly change from administration to administration; items should not be biased against certain groups

3) unidimensionality: this means that a test measures one, and only one, dimension or ability; you do NOT want to combine items in one test that measure different things -- it's like adding apples and oranges -- you have no idea what you have in the end; this is why combining written, oral and clinical scores can be unwise

Planning the Test: Creating the Table of Specifications

Ideally, when a test is being planned, the Table of Specifications, or exam blueprint, is defined first using training requirements, educational curricula, a job analysis and other data. This Table is a two-way table, with learning outcomes (or educational objectives) as one dimension and content areas as the second dimension.

Once the learning outcomes (or educational objectives) and content areas have been delineated, the items are then written that match those specifications.

STEP 1 -- IDENTIFYING THE LEARNING OUTCOMES

Specific learning outcomes should:

·be concise (1-2 sentences each)
·focus on one aspect of behavior
·describe the behavior as a desired end product
·focus on observable behaviors only
·use definite terms (write, define, list, identify, predict, select, etc.)
·avoid vague terms (learn, see, realize, develop, understand, apply, etc.)

Bad Examples...

1) Develop **accuracy** (undefined trait)
2) **Know** the rules for constructing essay tests (indefinite term)
3) Define and calculate the mean and explain its uses (multiple behaviors)

Good Examples...

Identifies the correct definition of terms.
Mount a USB key drive.
Schedule a cron job.

After the learning outcomes have been formulated, the outcomes are placed within some classification scheme for entry into the Table of Specifications. Often, Bloom’s Taxonomy of Educational Objectives is utilized for this purpose (Table 1). In the end, each item is then classified in two ways - according to the Taxonomy (Knowledge, Comprehension, Application) and according to its content area.

Table 1 -- Taxonomy of Educational Objectives: Cognitive Domain

Knowledge

1.00 **KNOWLEDGE** (remembering previously learned material)
1.10 Knowledge of specifics
1.11 Knowledge of terms
1.12 Knowledge of specific facts
1.20 Knowledge of ways and means of dealing with specifics
1.21 Knowledge of conventions
1.22 Knowledge of trends and sequences
1.23 Knowledge of classifications and categories
1.24 Knowledge of criteria
1.25 Knowledge of methodology
1.30 Knowledge of the universals and abstractions
1.31 Knowledge of principles and generalizations
1.32 Knowledge of theories and structures

Intellectual Abilities and Skills

2.00 **COMPREHENSION** (grasping the meaning of the material)
2.10 Translation (converting one form to another)
2.20 Interpretation (explaining or summarizing material)
2.30 Extrapolation (extending meaning beyond the data)

3.00 **APPLICATION** (using info in concrete situations)

4.00 **ANALYSIS** (breaking down material into its parts)
4.10 Analysis of elements (identifying the parts)
4.20 Analysis of relationships (identifying relationship)
4.30 Analysis of organizational principles (identifying the ways parts are organized)

5.00 **SYNTHESIS** (putting parts together into a whole)
5.10 Production of a unique communication
5.20 Production of a plan or set of operations
5.30 Derivation of a set of abstract relations

6.00 **EVALUATION** (judging value of a thing using definite criteria)
6.10 Judgments in terms of internal evidence
6.20 Judgments in terms of external criteria

For simplicity, the last four classifications are often combined under one heading -- APPLICATION.

STEP 2 -- IDENTIFYING THE CONTENT AREAS

“The content domain to be covered by a licensure or certification test should be defined clearly and explained in terms of the importance of the content for competent performance in an occupation. A rationale should be provided to support a claim that the knowledge or skills being assessed are required for competent performance in an occupation and are consistent with the purpose for which the licensing or certification program was instituted.” (Standards for Educational and Psychological Testing, standard 11.1, p.64).

The next step is to outline the subject-matter content (subscales) to be measured by the exam. Again, training/educational curriculum and job analyses can be extremely helpful. The subject-matter content defines the exam subscales. The content should be detailed as needed.

STEP 3 -- PREPARE A TABLE OF SPECIFICATIONS

After the learning outcomes and subject-matter content have been defined, a two-way Table of Test Specifications should be created. This Table relates the learning outcomes to the content and indicates the relative importance that will be applied, in terms of numbers of test items, to various areas. Below is the Table of Specifications for the BSDA exam.

Content Area (Domain)	Content %
Installing and Upgrading the OS and Software	13
Securing the Operating System	11
Files, Filesystems and Disks	15
Users and Accounts Management	16
Basic System Administration	12
Network Administration	15
Basic Unix Skills	17
Cognitive Classification (Bloom's)	Bloom's %
Knowledge	TBD
Comprehension	TBD
Application	TBD

As items are written, they need to be classified by both objective number and cognitive classification. The item writer's name should be included, as well as the reference source for the correct answer, including the full name and version of the source and its page number.

General Item Writing Principles

1. Use items that measure important learning objectives.

Although the evaluation of knowledge or specific facts is often a proper measurement goal, testing of trivial knowledge should be avoided. A common error is to focus on trivial facts that are irrelevant.

On what day in 1990 did Iraq invade Kuwait?

The following item stem focuses on a more important idea, specifically the reasons for the invasion.

What was the stated Iraqi reason for invading Kuwait in 1990?

2. Avoid items containing ambiguous language or phrasing.

If the word or phrase you use in a question is ambiguous, the examinee may misunderstand the question. There are several ways in which words can be ambiguous. Words can be ambiguous when they have two or more meanings. Examples are: poor; fire; open; fit

Another example of a potentially ambiguous word is the use of where in the following question.

Where would you go to find out the effects of a drug?

The examinee could interpret where to mean a place (pharmacy), a book, or even a person (pharmacologist).

A set of words which must be used precisely are should, could and might. An example of how the use of these words could affect questions follows.

Outline what should be done to improve U.S. balance of payments.
Outline what could be done to improve U.S. balance of payments.
Outline what might be done to improve U.S. balance of payments.

Each question has a different connotation that changes the way in which the examinee answers the question. Should connotes a moral issue of needs, could connotes possibility and might conveys the issues of probability. Other words that cause similar problems are can, may, and must.

3. Follow standard rules of punctuation and grammar.

A. If the item is a direct question, it should end in a question mark.

B. If the stem is an incomplete sentence, then the first word in each distractor should be lower-case and the options should end with a period.

> The most common bird in Washington, D.C. is the
> a. mynah.
> b. robin.
> c. Ladybird.
> d. mockingbird.

C. If the options for an incomplete stem are proper names, they should be capitalized and followed with a period.

> The dog that starred in the T.V. show My Three Sons, is
> a. Lassie.
> b. Rin Tin Tin.
> c. Tramp.
> d. Lady.

D. If both the stem and the options are complete sentences, each option should begin with a capital letter and end with a period or question mark.

> Which of the following supports the statement that glaciers covered the Great Lakes region in the past?
> a. The region is relatively cool in the summer.
> b. Unsorted deposits are found in the region.
> c. Tropical vegetation is absent in the region.
> d. Igneous rocks are found in the region.

4. Write items that have only one right or clearly best answer.

Especially for multiple choice exams, the alternatives should include only one correct answer. Multiple correct answers invalidate the question and often cause the item to be dropped from the exam. Keeping track of the source of each item provides a reference for the correct answer if the answer is ever questioned by an examinee or anyone else.

5. Keep the level of reading difficulty appropriate to the group being tested.

6. Avoid items based on statements taken verbatim from instructional materials.

The only thing this type of item evaluates is the examinee's ability to memorize. Also, when the passage is taken out of instructional materials it can lose its original context and can become trivial or meaningless.

The second principle for good education is _____________.

This question is only relevant within the context of the material.

7. If any items are based on opinion or authority, state whose opinion or what authority.

The people best suited to rule the ideal state are the
a. managers.
b. philosophers.
c. soldiers.
d. nurses.

This question could be (and has been) debated endlessly. A better stem is, "According to Plato, the people best suited to rule the ideal state are the..."

8. Avoid items containing irrelevant clues.

An irrelevant clue is anything that enables an examinee to identify the correct answer without knowing the correct answer.

The Soviet Air Defense Forces, although primarily defense oriented,
a. can also provide air transport.
b. may take offensive action.
c. will form the initial strike force in a conventional war.

The phrase beginning with "although..." asks for a contrast. The opposite of defense is offense which leads the examinee to the correct answer B.

9. Avoid items that help examinees answer other items in the test (enemies).

When either of the two items below stand alone, they are acceptable items. However, when the two items are combined, the testwise examinee has been given the correct answer for the second item. These are called enemy items.

> In the TV show, The Partridge Family, David Cassidy played what character?
> a. Danny.
> b. Keith.
> c. Rueben.
> d. Chris.
>
> What hit TV series featured David Cassidy and Shirley Jones?
> a. The Brady Bunch.
> b. The Monkees.
> c. The Partridge Family.

10. Avoid trick items.

Trick items tend to anger and frustrate an examinee, in addition to reducing your credibility to zero.

11. Check all items with other colegues to eliminate ambiguity, teknical errors and other erors in item writeing.

We all tend to overlook errors when editing our own material. Have a colleague check your items when you first write the items and after you have edited the items. Before the exam is assembled and administered, it should be reviewed by an entire committee.

12. Avoid items that test more than one idea.

The following asks about two concepts, “what” and “why.” If both parts are important to the exam and its Table of Specifications, each part should be addressed in separate items.

> In the Andy Griffith Show, what did Opie kill with a slingshot and why?
> a. A bird for target practice
> b. A mouse for snake food
> c. A rabbit for dinner
> d. A bird because he was afraid of it

13. Avoid items that have answers that are subject to change within a short period of time.

Who is the President of the United States?

A better version is...

Who was the President of the United States in 2005?

APPENDIX: Item Writing Checklist

GENERAL CHECKLIST

1. Use items that measure important learning objectives.
2. Avoid items containing ambiguous language or phrasing.
3. Follow standard rules of punctuation and grammar.
4. Write items that have only one right or clearly best answer.
5. Keep the level of reading difficulty appropriate to the group being tested.
6. Avoid items based on statements taken verbatim from instructional materials.
7. If any items are based on opinion or authority, state whose opinion or what authority.
8. Avoid items containing irrelevant clues.
9. Avoid items that help examinees to answer other items in the test.
10. Avoid trick items.
11. Check all items with other colleagues to eliminate ambiguity, teknical errors and other erors in item writeing.
12. Avoid items that test more than one idea.

13. Avoid items that have answers that are subject to change within a short period of time.

THE STEM CHECKLIST

Note: the ***stem*** is the exam question itself.

1. Clearly define the question in the item stem.

2. Include as much of the item in the stem as possible.

3. Omit irrelevant material in the item stem.

4. Avoid grammatical cues in the item stem.

5. Use the one best answer format.

6. Eliminate or reduce the number of negatively stated items.

THE OPTIONS CHECKLIST

Note: the ***options*** are the possible answers to choose from.

7. Avoid correct options that are noticeably longer or shorter than the incorrect options.

8. Avoid stating the correct option in more detail than the incorrect options.

9. Avoid generalizing the correct option so that it has wider application than the incorrect options.

10. Avoid correct options that are one of two options that state the idea of fact in diametrically opposite fashion.

11. Avoid correct options that contain familiar or stereotyped phraseology.

12. Make all options grammatically consistent with the stem.

13. Vary the position of the correct option.

14. If the options can be put in some natural order, avoid making the first and last options always the incorrect options.

15. Avoid incorrect options that contain language or technical terms that the examinee may not know.

16. Avoid incorrect options that contain emotive words like nonsense, foolhardy and harebrained.

17. Avoid incorrect options that are flippant remarks or unreasonable statements.

18. Make all options parallel in form.

19. Avoid the options all of the above and none of the above.

20. Avoid clang correct options.

21. If the item is testing the definition of a word, put the word to be defined in the stem and make the options alternative definitions or meanings.

22. Write at least three distractors for every question.

23. Make each distractor plausible and attractive to examinees who have NOT mastered the material.

24. When possible, arrange the options in a logical or sequential order.

25. Make all options independent of each other.

26. Avoid specific determiners like sometimes and never in the options.

27. Use important-sounding words in the distractors as well as in the correct option.

28. Avoid repeating key word(s) from the stem in the correct option.

29. List options on separate lines, beneath each other.