Item Analysis Explanation

The item statistics below are examples of the type of information seen when reviewing examination items. This example represents a four-option, multiple choice item. The correct answer for this item is D.

The b-parameter is a way to evaluate item difficulty in item response theory (IRT). For a multiple choice item, the difficulty of an item is the point at which there is an equal chance of getting the item right or wrong. IRT b-parameters generally range from -3.0 (very easy) to +3.0 (very difficult), with a mean of 0.0.

The discrimination is a value that denotes whether or not the question can differentiate between candidates who possess the minimally acceptable level of knowledge to become certified and candidates who do not. Castle's recommendation for discrimination is a positive value at or above 0.15. The discrimination is stronger the closer it nears 1.

The classical item difficulty is the percentage of candidates who answered the question correctly. Castle's recommendation for classical difficulty is between 0.30 (difficult) and 0.92 (easy). However, there are some instances where ratings slightly below 0.30 or above 0.92 are acceptable.

B-parameter	**0.12**				
Discrimination	**0.3757**			**Item ID**	**0000**
N Candidates	501	**Difficulty (p-value)**	0.7285		
Option	**Key(s)**	**# Chose**	**% Chose**	**Mean Raw Score**	**Discrimination**
A	N	72	14.4289	77.8889	-0.2766
B	N	30	6.012	80.1	-0.13
C	N	32	6.4128	80.2188	-0.1323
D	Y	365	73.1463	90.3836	0.3757

The key (i.e., correct answer) for this item is D.

This column represents the percentage of candidates who selected the corresponding option. When the data in this column are lower than 3%, the accompanying distractor might not be plausible. If few candidates select a distractor, a key verification may be necessary to evaluate the distractor's plausibility.

The data in this column represent the average score that individuals who chose this response option received on the overall examination.

The data in this column represent the discrimination for each response option. The discrimination for the key should be positive. The discrimination for the distractors should be negative or near zero, and always lower than the discrimination of the key.

Glossary

Anchor Exam: An examination form that sets the standard of passing for a given series of examinations.

Certification: Authorized declaration validating that one has fulfilled the requirements of a given profession and may practice in the profession.

Classification System: A systematic arrangement of examination content in groups or categories according to specified criteria. Castle Worldwide, Inc. uses a six digit coding system to represent the domain, task, and knowledge or skill a specific question covers.

Content Domain: A body of knowledge, skills, and abilities defined so that items of knowledge or particular tasks can be clearly identified as included or excluded from the domain.

Cut Score: A specified point on a score scale at or above which candidates pass or are accepted and below which candidates fail or are rejected. This is also sometimes called the passing score or passing point.

Discrimination: The ability of a test or a test question to differentiate among qualified and unqualified individuals by measuring the extent to which the individual display the attribute that is being measured by the test or test question.

Distractor: The options that are not correct answers. Distractors must be plausible; hence, they distract the less qualified test-taker from the correct answer.

Equating: A process used to convert the score from one form of a test to the score of another form so that the scores are equivalent or parallel.

Equator: Questions that are on all forms of an examination, including the anchor form. These questions are used to equate test forms.

Internet-Based Testing: Computer-based testing where the examination is delivered via a secure, password protected Web site. The examination and the candidates' answers are uploaded to the test provider's secure server. Test security is assured through configuration management, controlled loading, and availability.

Item: A test question that consists of a stem, correct response, and distractors.

Item Analysis: The process of assessing certain characteristics of test questions, specifically the question difficulty, the discrimination, the candidates' mean scores, and the distractor discrimination.

Item Difficulty: The percentage of candidates answering a question correctly. This value can be computed to provide data about first-time candidates, retake candidates, ability level, etc.

Item Response Theory: A model-based measurement method that yields trait-level estimates.

Key: The correct answer in a list of options.

Knowledge Statement: An organized body of factual or procedural information.

Minimally Qualified Candidate: An individual's competence in a particular job role can be seen as a continuum ranging (theoretically) from the complete lack of ability to the highest level of mastery. The

term *minimum competence* suggests that the individual is capable of filling the role with sufficient mastery to not harm the public or the profession.

Options: The list of possible answers for a question including the correct answer.

Performance Domain: The major responsibilities or duties of a specific field of study. Each domain may be characterized as a major heading in an outline format and may include a brief behavioral description.

Psychometrics: The design, administration, and interpretation of quantitative tests that measure psychological variables such as aptitude, intelligence, skill, and learning.

Raw score: The unadjusted score on a test, usually determined by counting the number of correct answers.

Reliability: The reliability of a test refers to the consistency of the test result. We interpret the reliability of a test as a measure of the likelihood that if we gave the test again under the same conditions, we would then observe the same scores.

Role Delineation Study: Also known as job analysis study. The method of identifying the tasks performed for a specific job or the knowledge, skills, and abilities required to perform a specific job.

Scaled Score: A score to which raw scores are converted by numerical transformation (e.g., standardized scores).

Score: Any specific number resulting from the assessment of an individual. A number that expresses accomplishment either absolutely in points earned or by comparison to a standard.

Scoring Formula: The formula by which the raw score on a test is obtained. The simplest scoring formula is the raw score equals the number of questions answered correctly.

Skill Statement: The proficient physical, verbal, or mental manipulation of data, people, or objects. Skill embodies observable, quantifiable, and measurable performance parameters and may be psychomotor or cognitive in nature.

Standard Error of Measurement: The standard deviation of the hypothesized distribution of test score means if multiple samples from which to compute the mean were available. We interpret the standard error of mean as a measure of variability we would observe in multiple sample or test administrations.

Stem: The body of the question including any scenarios or qualifying information.

Subject Matter Expert: A person with expertise in a given field or profession. Subject matter experts are used to develop the content of examinations.

Task Statement: A comprehensive statement of work activity that elaborates upon the performance or content domain. Each task statement details a particular work activity in such a way that the series of task statements will offer a comprehensive and detailed description of each performance domain.

Test Specification: A content outline that specifies what proportion of the test questions will deal with each content area.

Validation: The process of rating each test question in order to determine how important, critical, and/or frequently the content tested by a specific question is used for a specific job.

Validity: Refers to the quality of the inferences made from a test score/result. If the purpose of a particular examination is to certify a minimally qualified candidate in a particular profession, then the question we ask is whether minimal qualification can be inferred from the examination. Alternatively, validity can be conceptualized as the accuracy of the test score.

Weighted Scoring: Scoring in which the number of points awarded for a correct response is not the same for all questions on a test.