



**Item Writing Guide**  
**January 2008**

1 N. Dearborn Street  
Suite 1600  
Chicago, Illinois 60602



## Table of Contents

---

1. INTRODUCTION .....	1
2. ITEM WRITING .....	1
3. CRITERION-REFERENCED TESTING .....	12
4. EXPLANATION OF ITEM STATISTICS .....	12
5. OVERALL REVIEW OF ITEM STATISTICS .....	15
6. ITEM WRITING REFERENCES.....	16

## 1. INTRODUCTION

---

Pearson VUE has prepared this guide to facilitate the process of developing items for multiple-choice examinations. The purpose of this guide is threefold: (1) to help writers create valid, accurate and reliable items; (2) to inform test developers of Pearson VUE's technical requirements; and (3) to improve item writers' understanding of item statistics and how to interpret them.

Section 1 presents how—and how not—to write items, with points to remember and pitfalls to avoid. This section also explores mechanical, grammatical and stylistic considerations. In addition the section delivers item import/export format specifications and other functional requirements. Section 2 discusses how to use statistical results to make informed decisions about retaining and rewriting items. Section 3 provides additional item writing references.

## 2. ITEM WRITING

---

### Test Blueprints & Item Specifications

Tests are designed to measure a domain or range of knowledge. Subsets of expertise exist within this given domain. For example, a test measuring knowledge of cuisine might have questions about Italian, French, Chinese and Mexican cooking. These categorizations represent content classification at the primary level. Furthermore, these subsets might have subsets within them. The subset “Italian cuisine” might contain antipasti, entrees, desserts, etc. These are content classifications at the secondary level.

The breakdown and categorization of the information within the domain of knowledge and the percentage of total items on the test that fall within each category is the test blueprint or content outline. For example a 200-item test on cuisine might contain:

I.	French cuisine	20 items	(10%)
II.	Italian cuisine	70 items	(35%)
III.	Chinese cuisine	60 items	(30%)
IV.	Mexican cuisine	50 items	(25%)
Total exam		200 items	(100%)

This particular example presents the selection criteria at the primary content level. The examination, consisting of 200 total items, contains 20 items related to French cuisine, 70 items related to Italian cuisine, 60 Chinese cuisine items and 50 Mexican cuisine items.

Some tests have more specific selection criteria. Here is an examination blueprint with selection criteria at the secondary content level:

I.	French cuisine	(20 total items)
	A.	Regional (7 items)
	B.	Appetizers (5 items)
	C.	Entrees (5 items)
	D.	Desserts (3 items)

II.	Italian cuisine	(70 total items)
	A. Regional	(30 items)
	B. Appetizers	(20 items)
	C. Entrees	(15 items)
	D. Desserts	( 5 items)
III.	Chinese cuisine	(60 total items)
	A. Regional	(25 items)
	B. Appetizers	(20 items)
	C. Entrees	(10 items)
	D. Desserts	( 5 items)
IV.	Mexican cuisine	(50 total items)
	A. Regional	(20 items)
	B. Appetizers	(15 items)
	C. Entrees	(10 items)
	D. Desserts	( 5 items)

Each primary content area is broken down into more specific categories and each of those categories is ascribed a percentage or raw number of items to appear on the examination. Test blueprints may be as specific or as general as required to adequately test the overall domains of knowledge as well as the subcategories within that domain.

For example, an item that tests knowledge of ingredients for a crepe recipe would have a classification code of “I.D.,” identifying its primary classification as “French cuisine” and its secondary classification as “Desserts.” Generally, a classification scheme will be provided for item writers that specifies the exact code to be included at each level of classification.

Content classifications are derived from task analyses (sometimes called job or practice analyses) or by consensus of content experts.

Providing correct classifications is essential when writing items. These item classifications or labels enable test developers to easily identify various types of questions and ensure that the entire domain of knowledge is represented on the examination.

## Writing Multiple-Choice Items

Item writing can be a difficult process. It requires not only thorough knowledge of a subject matter but also the ability to verbalize thoughts and ideas and the creativity to present those ideas in novel and challenging ways. It also requires the ability to write consistently across different concepts. Adhering to basic item writing standards is necessary to obtain a reliable and valid measure. Please use the following guidelines to assist you in your item writing.

### ***Choosing a Topic***

The first step in writing an item is choosing a topic. Item writers should choose topics that are familiar to them and that represent aspects of the domain of knowledge to be tested. In many cases a test development committee or chairperson will assign topics. These assignments usually reflect the item distribution across subtopics within the discipline as defined by the test blueprint or content classification scheme. Whether items are assigned or categories are chosen by the author, constructing items that are relevant to the purposes of the test and fit into the test blueprint is crucial to developing a valid, reliable test.

Several sources can help you generate ideas for item writing. One such source is professional practice or experience. Everyday job or practice situations can be an excellent source of ideas for sound test items. Common professional situations often have standard procedures to be followed, and those procedures are generally well referenced.

Other good sources are textbooks and journal articles. Concepts presented in textbooks are generally well researched and time tested. Journals provide an excellent resource for some of the more recent and cutting-edge concepts within professions.

Once ideas have been generated for item writing, the process of constructing the actual item begins.

### ***Constructing a Multiple-Choice Item***

First, an introduction to some of the terminology used in association with test development is in order. Tests are made up of **items**. An **item** refers to an individual scoring unit (commonly called a question). There are several types of items including multiple choice, true/false, fill in the blank and essay. This document will focus on how to write multiple-choice items.

The main goal is to develop items and cases that are CLEAR, SPECIFIC and ACCURATE.

Multiple-choice items are made up of two parts, the **stem** and the **responses**. The stem introduces the problem and is followed by the responses. The responses consist of the **answer**, which is the correct response, and the **distractors**, which are the incorrect responses.

### ***Constructing Stems***

If an item performs poorly on an examination, the stem may be badly worded. A good stem will consist of an introductory question, an incomplete statement, a description of a situation or a statement of an issue.

- The stem's content should focus on a single theme or problem. Focus is essential to writing clear test items. A lack of focus can contribute to confusion on the part of the examinee.
  - The stem should be economically worded.
  - The concept tested in the stem should be consistent with evaluation goals and germane to the discipline.
-

- The stem should be grammatically correct, both alone and in conjunction with the responses.
- The stem should not contain personal pronouns.
- Negatively worded stems should be avoided.

### ***Constructing Responses***

Responses should be plausible alternatives to the correct answer. They should be **homogenous**, i.e. conceptually and structurally parallel.

### ***Item Examples***

Chicago is

- A. adjacent to a Great Lake.\*
- B. culturally homogenous.
- C. the largest city in the USA.
- D. one of the original 13 colonies.

This item has three fundamental problems. The first is the unfocused stem. A stem should be focused enough so that all responses fall within a particular topic and structure. An examinee should be able to read the stem and construct the answer without reading the responses. This stem gives no indication of the problem to be solved.

The second problem is the **heterogeneity** of the responses. Heterogeneous responses are conceptually and grammatically divergent and are not limited to one topic. The responses in this example cover geographical, historical and demographic concepts.

A third problem is the plausibility of the distractors. Distractors should be attractive alternatives to the correct response without being deceitful or tricky. Distractors should not be so absurd as to be obviously false. The likelihood that anyone would confuse Chicago with one of the 13 original colonies is slim to none.

### **Better:**

Chicago is adjacent to which Great Lake?

- A. Ontario
- B. Huron
- C. Michigan\*
- D. Erie

The stem in this example is much more focused. Additionally, the responses in this example are homogenous and parallel in both content and structure. This parallelism adds to the plausibility of the distractors.

### ***Window Dressing***

Another common problem with item construction relates to items with extraneous or irrelevant information, also known as "window dressing." Excess information not essential to the question serves only to confuse examinees and slow them down. An item should only contain the information necessary to arrive at the correct response.

---

A train departs from Milwaukee at 12 noon on Friday headed due south, traveling at 60 miles per hour. Its destination is Union Station in Chicago, which is 123 miles away. Assuming that the train does not stop and maintains its speed, when will the train arrive at its destination?

- A. 1:23 p.m.
- B. 2:03 p.m.\*
- C. 2:23 p.m.
- D. 3:03 p.m.

A more efficient way to present this concept is:

A train departs at 12 noon traveling at 60 miles per hour. Assuming that the train maintains a constant speed, when will the train arrive at its destination, 123 miles away?

- A. 1:23 p.m.
- B. 2:03 p.m.\*
- C. 2:23 p.m.
- D. 3:03 p.m.

The second stem is more streamlined than the first. It eliminates unnecessary information. It is not important to include the day of the event. Nor is it important to precisely name the destination or the point of origin. Extraneous information is omitted and the abridged item is much more straightforward.

### ***Cueing***

Another typical problem with items is cueing. The term refers to an item inadvertently tipping off or “cueing” the examinee to a correct answer. Several types of cueing frequently occur.

One obvious type of cueing is an item that contains information that can be used on another item in the test. Consider one of our earlier examples:

Chicago is adjacent to which Great Lake?

- A. Ontario
- B. Huron
- C. Michigan\*
- D. Erie

Including the above item on a test with the following item would be unwise:

Which of the following is a Great Lake?

- A. Winnebago
- B. Louise
- C. Erie
- D. Placid

The former item cues the reader to the answer of the latter.

Another form of cueing relates to the grammar and structure of the item. Grammatically both the correct answer and the distractors should mimic the stem. Occasionally items are written in which the distractors are obvious because they do not conform grammatically to the stem. For example:

---

When baking bread it is important to:

- A. letting the dough rise for 10 to 15 minutes.
- B. preheat the oven prior to baking.
- C. sifted flour is generally more difficult to work with.
- D. do not use road salt.

The above example has only one response that appropriately completes the sentence. The distractors are obvious because of their grammatical incompatibility with the stem. This item, although it may test a worthy concept, would need reworking.

The words “always” and “never” are another common source of cueing. As a general rule, it is a good idea to avoid the words “always” and “never” in both stems and responses as those questions and answers tend to be obviously correct or incorrect.

Often, the length of a response can cue the reader. The length of an item frequently is an indication of its correctness because item writers unintentionally put the most information in the answer. Many examinees know this and other tricks and look for such structural cues. As a general rule it is good to keep responses parallel in length, content and structure.

### ***Bias***

Neutrality and even-handedness is central to effective item writing. When writing items avoid using positive or negative stereotypes, in particular those about race, gender, ethnicity, culture, disability, age or sexual preference.

With regards to examinations involving case assessment, do not mention a client/patient’s age, race or sex unless the information is necessary to answer the item.

### ***Cases***

In some situations the use of a **case** may be appropriate. A case is a picture (e.g. drawing, photograph, x-ray, schematic), a block of text (e.g. personal history, medical history, specifications) or a combination of both that pertains to one or more items and assists in presenting a complete scenario. Cases should be clearly drawn or represented and only used to illustrate or clarify a concept. Cases, particularly pictures, should never be used for interest or entertainment. Any labels on pictures should be clearly visible and easily read.

When using a case, writers should ask themselves, “Can this item be answered without the benefit of the case?” If the answer is yes, the case is unnecessary and will only clutter the test. Consider restructuring the item.

When submitted, cases should be clearly identified and labeled with the correct AN and text of the corresponding item. For further information see the section on technical specifications.

### ***References***

Test committees frequently request that item writers include at least one reference for each item. Test developers use references to check the accuracy and importance of concepts when reworking items. Acceptable references can include textbooks or journal articles. A good rule of thumb to follow when choosing the references is that they should be widely accepted, accessible and representative of the mainstream thought in the subject matter.

A reference should minimally contain the author, title of the book or journal, chapter of the book, edition number, page numbers and publisher. Requirements for references and their form may vary from

---



organization to organization. Consult your test development committee for specific reference requirements.

### ***Reworking Items***

On occasion, an item may need to be pulled from a test or a bank. At this time a choice must be made: Should the item be reworked or edited, removed from the bank or replaced with an entirely new item? Use the guidelines provided for creating new items when reworking or replacing items.

Items containing outdated information must be reworked. If the information can be updated, the item can be made appropriate for continued use.

### **Standard Item Writing Guidelines**

Please follow established standard procedures and formats.

### ***All Items***

Items should be written in 11-pt. Arial font and saved as a word-processing document. Responses should be identified with a capital letter and separated from the text with a period and a tab. When responses are numbers, they should be aligned by place value and listed in ascending order.

How many legs does a giraffe have?

- A. 0.75
- B. 4.00
- C. 15.00
- D. 100.00

### ***Direct Questions***

Items in which the stem asks a question and the responses provide complete answers.

Stems should begin with a capital letter and end with a question mark.

Responses, if they are **complete sentences**, should begin with a capital letter and end with a period. A complete sentence has a subject and a verb, and does not usually begin with a conjunction. Response that are not complete sentences should begin with a lowercase letter and end without punctuation. Proper names/nouns, however, **should** be capitalized in all circumstances.

In the nursery rhyme "Hickory Dickory Dock," what caused the mouse to run down the clock?

- A. He saw a cat.
- B. The master wound the clock.
- C. The clock struck one.
- D. He smelled cheese.

What color is the sky?

- A. red
  - B. yellow
  - C. green
  - D. blue
-

### ***Incomplete statements***

Items in which the responses finish a sentence begun in the stem.

Stems should begin with a capital letter and only grammatically necessary punctuation should end the sentence fragment. For example if the responses are lists of four or more items, then the stem should end with a colon.

Responses should begin with a **lowercase** letter and end with a period. Proper names/nouns, however, **should** be capitalized in all circumstances.

A person who is ambidextrous is able to write

- A. better with her right hand.
- B. better with her left hand.
- C. equally well with both hands.

New England consists of six states. They are:

- A. Illinois, Wisconsin, Michigan, Iowa, Minnesota, Indiana.
- B. Maine, New Hampshire, Vermont, Massachusetts, Connecticut, Rhode Island.
- C. New York, New Jersey, Pennsylvania, Delaware, Maryland, Virginia.
- D. North Carolina, South Carolina, Kentucky, Tennessee, Georgia, Florida.

### ***Consistency***

Consistency is the key to successful item writing. Following these suggestions will promote such consistency and consequently lessen the possibility that an item tests an examinee's logic skills or reading ability, as opposed to the intended domain of knowledge.

### **Rules of Grammar & Usage**

Items exhibiting poor grammar, mechanics and style may confuse examinees and hence are poor determinants of candidates' knowledge.

Frequently encountered problem areas are listed below in bold. A description of the correct style and usage follows, accompanied by examples of poorly and well-written items.

#### ***Parallelism***

Responses should be **parallel** in grammar, sentence structure and length. In other words, the verb tenses and type of nouns, subject-verb-object order, punctuation, number of words and sophistication of language should be similar in each. This format enables the examinee to focus his/her energies on ascertaining the correct answer based on its content rather than looking for clues in its structure. Often item writers unintentionally point the examinee toward the answer by writing it differently from the rest.

#### **Poor:**

Why did Little Miss Muffet run away?

- A. Along came a spider that sat down beside her.
  - B. She finished her curds and whey.
  - C. A bear frightened her.
-

- D. She stepped on a slug.

The correct answer is A. Notice that it is considerably longer than the distractors and the word order is inverted. In the answer the word order is verb-subject-object/preposition whereas in the distractors it is subject-verb-object/preposition. Smart examinees look for such grammatical clues, so they should be avoided.

**Better:**

Why did Little Miss Muffet run away?

- A. A spider sat down beside her.
- B. She finished her curds and whey.
- C. A bear frightened her.
- D. She stepped on a slug.

The word length ranges from 4-6 words, each sentence is organized in the subject-verb-object/preposition format, the voice is active and the verbs are all past tense. If the examinee did not know the nursery rhyme, he/she could not deduce the answer by looking for inconsistencies in the response style.

**Hyphenation<sup>1</sup>**

Hyphens join words or word parts closely associated with one another. Most frequently they connect the parts of compound modifiers or a prefix to a word. In the former instance, they tend to be underused, but in the latter they are severely overused.

**Compound Modifiers**

A compound modifier is an adjective or adverb made up of two or more words that **together** modify another noun. A hyphen should go between the words, unless one of the words is *very* or an *-ly* adverb.

**Poor:**

a pencil and paper test  
the gaily-dancing couple

The first example should be hyphenated because the words “pencil and paper” modify “test,” not each other. The second example should not be hyphenated because the adverb ends in *-ly*.

**Better:**

a cost-effective proposal  
a happily married couple

The modifier “cost-effective” is hyphenated because together these words describe “proposal.” “Happily married” is not hyphenated because of the adverb’s *-ly* ending.

When the compound modifier appears after the noun, the hyphen is dropped.

a three-year-old boy  
The boy is three years old.

---

<sup>1</sup> Kessler, L.; McDonald, D. (1992) *When Words Collide: A Media Writer's Guide to Grammar and Style*. 3rd. (pp.110-112). Belmont, CA: Wadsworth Publishing Company.

## Prefixes

As a general rule, do not place a hyphen after a prefix unless:

The prefix ends in a vowel and the root word begins with the same vowel

Example: *pre-eminent* (Exceptions: *coordinate*, *cooperate*)

The root word begins with a capital letter.

Example: *pro-Democrat*

The prefix and root word together make another word with a different meaning.

Example: *re-create* vs. *recreate*

Note that numerous exceptions abound. If unsure, check in the dictionary.

## Commas before Conjunctions

The most misunderstood, abused mark of punctuation is the comma. Myriad situations call for the use of commas, but in items commas most often appear before conjunctions.

### Before *and*

**Objects in a simple series:** When a series consists of three or more objects, a comma should appear between each object except before the conjunction. EXCEPTION: If the final object contains a conjunction, put a comma before the final *and* in the series.

The athlete won medals in diving, swimming and gymnastics.

She listened to classical, jazz, and rock'n'roll music.

**Objects in a complex series:** If the objects in a series are long or confusing, a comma should appear between each, including before the final conjunction.

In deciding to run for office, the politician had to consider whether she could raise enough money to fund a campaign, whether she could withstand public scrutiny, and whether voters would support her often-controversial views.

**Linking sentences:** Use a comma before *and* unless the clauses are closely related.<sup>2</sup>

Joe dressed meticulously, and off he went to pick up his date.

Joe dressed meticulously and he wore his new jacket.

### Before *but*

Use a comma before *but* if the second subject is stated.<sup>3</sup>

Elle went to the post office but forgot to buy stamps.

Elle went to the post office, but she forgot to buy stamps.

### Before *or*

---

<sup>2</sup> Ryan, L. (1995). *The Editor's Toolbox: A Reference Guide for Beginners and Professionals* (M. O'Donnell, Ed.) (pp. 23-25). Forest City, IA: Pug Publishing.

<sup>3</sup> Ibid.

Use a comma before *or* when the second subject is different from the first.<sup>4</sup>

Mark must swerve immediately or he will hit the pedestrian.

Mark must swerve immediately, or David will get hurt.

### **Numerals**

The rule for use of numerals in item writing is a deviation from that of standard written English. To improve readability and reduce confusion, particularly on computer-based tests, use numerals to represent numbers **in all circumstances**.

Even the numbers “zero” through “nine,” which traditionally are spelled out unless referring to measurements or percentages, should be written as numerals.

#### **Poor:**

Suzy has \_\_\_\_\_ fingers on her left hand.

- A. three
- B. five
- C. seven
- D. 11

The “11” is much easier to read than the spelled-out numbers, even though all technically are correct grammatically. An examinee might lose time just figuring out what numbers the words represent. Items should be as clear as possible, so examinees can quickly and fully understand what information they must provide.

#### **Better:**

How many fingers does Suzy have on her left hand?

- A. 3
- B. 5
- C. 7
- D. 11

Using numbers in all circumstances will improve the items’ readability, as demonstrated above. The responses are clear and their discreteness apparent.

### **Conclusion**

The rules of grammar and style listed are intended to improve the clarity and accuracy of written communication, and they usually do. They can be bent, however, if strict adherence to them would make an individual item stilted, awkward and/or overly complex. Determinations of whether to bend the rules should be made on a case-by-case basis.

Questions regarding item specifications should be directed to the Psychometric team at 1-847-866-2001 or 1-800-255-1312.

---

<sup>4</sup> Ibid.

### 3. Criterion-Referenced Testing

---

A score on a criterion-referenced test is a measure of how well the examinee performs on questions across the content domains represented on the test. This is in contrast to a norm-referenced test in which the examinee's score is a measure of how well he/she performs in relation to the performance of other examinees.<sup>5</sup>

A criterion-referenced test is grounded in a job/practice analysis, specifications of curriculum objectives and/or consensus of content matter experts that validates the content presented on the examination.

On a criterion-referenced test, clearly delineated content outlines or blueprints provide the framework for item writers. The item bank is balanced across content areas and candidates are administered tests that meet the blueprint specifications.

Item writing for a criterion-referenced test requires not only mastery of the subject matter, but also an understanding of the examination population and good communication skills.<sup>6</sup> Review of all items by content experts before they are included in the item bank ensures that they are relevant to practice, technically correct and of appropriate difficulty for the examinee population.

The passing score on a criterion-referenced test is established on a benchmark scale. Candidates who meet the passing standard have demonstrated minimal competency by correctly answering questions across the content domain and have demonstrated that they have the knowledge necessary to successfully perform in the field of practice. The domains of knowledge, defined by committees of content experts and supported by the job/practice analysis, are the basis for the criteria against which examinees are measured.

On a criterion-referenced test, it is theoretically possible for all examinees to pass the test, just as it is theoretically possible for all examinees to fail the test. Examinees' scores on the test reflect their degree of knowledge as measured by their performance on the criterion-referenced items, not their performance relative to other examinees.

### 4. Explanation of Item Statistics

---

Statistics commonly used in item analysis can be divided into two general categories: Classical statistics and Item Response Theory (IRT) statistics.

#### ***Classical Statistics***

Classical statistics are "population-dependent." This means that they apply only to the group of examinees or items that were used to generate them. They are very useful in describing the results of a particular test and for diagnosing the individual characteristics of an item.

The most commonly used Classical statistics are:

#### **P-Value**

Referred to as the "difficulty" of the item, this statistic is the percentage of examinees that selected a particular response. A P-value can be calculated for the correct answer and each of the distractors by dividing the number of individuals that selected the response by the total number of examinees. P-values provide the following general guidance:

---

<sup>5</sup> Lunz, M. E. (1993) *Examination Development Handbook*. Chicago, IL: American Society of Clinical Pathologists Board of Registry.

<sup>6</sup> Ibid.

---

- If the P-value is low (in the range of .00 to .20, the item is very hard and should be reviewed to determine that it has not been miskeyed or that there is not more than one correct answer.
- If the P-value is greater than .95, the item is very easy. The item should be deleted or revised to present a greater challenge to the examinees.
- If the P-value is approximately equal across all responses, examinees may be guessing at the answer. There may be insufficient information in the question to adequately determine the correct answer, or the content of the question may be very esoteric.
- If the P-value is zero for any response this is called a “Null Distractor.” Null distractors are indicative of obvious answers, nonparallel distractors or nonsensical distractors. Nonparallel distractors result from difficulty developing a sufficient number of distractors and from unfocused questions. In this case one or two distractors are radically different from the others in form and/or content such that they “stick out” as obvious wrong answers. Nonsensical distractors are those that are absurd or totally unrelated to the question.

### **Point Biserial Correlation**

Often referred to as “item discrimination,” the point biserial correlation is an association measure between an item response and the total test score. It can be understood as a correlation between the examinees who selected a particular response and the overall distribution of examinees across the test. It is calculated by standardizing the difference between the mean score of those selecting the response and the overall mean on the test and then restricting this value to a correlation range by multiplying by the root of the odds of a correct response. The formula for calculating the point biserial correlation is as follows:

$$PtBIS = \frac{(\bar{Y}_i - \bar{Y}_o)}{S_y} \sqrt{\frac{P_x}{1 - P_x}}$$

Where  $Y_i$  is the mean score of the individuals getting the item correct

$Y_o$  is the mean score of the total test group on the exam

$S_y$  is the standard deviation of the scores of the total group

$P_x$  is the percent of examinees answering the item correctly

The interpretation of the point biserial provides general guidance in evaluating item performance. A positive point biserial is desired for the correct answer and a negative point biserial is desired for each incorrect response. Some guidelines in the interpretation of the point biserial are as follows:

- The point biserial for the correct response should be .20 or higher.
- When the point biserial for the correct response is near 0.00, approximately equal numbers of more-able and less-able individuals selected that response.
- A negative point biserial for the correct response and a positive point biserial for an incorrect response indicates that more-able examinees are selecting the incorrect response instead of

the correct response. The item may have been miskeyed, worded in an unclear manner, obscure, or there may be more than one correct answer.

- The point biserial tends to be less informative at the score extremes. This means that items having very low p-values or very high p-values may have lower point biserials. This is also true with examinations having very small numbers of candidates. With a small test population, one or two more-able examinees selecting an incorrect response will cause the point biserials to change radically.

### **Item Response Theory Statistics**

Item Response Theory (IRT) provides a useful mathematical framework for the calibration and equating of items onto a common scale and the development of calibrated item banks.<sup>7</sup> The explanation that follows presents information on the Rasch model or one parameter IRT model. Rasch model statistics are considered to be "population independent."

With the Rasch model more able candidates have a greater probability of answering any question correctly than less able candidates and more difficult items have a lower probability of being answered correctly than easy items.

The equation that models these general assumptions is as follows:

$$\ln\left[\frac{P}{1-P}\right] = B_x - D_i$$

where P is the probability of answering the item correctly  
 B<sub>x</sub> is the ability of candidate x  
 D<sub>i</sub> is the difficulty of item i

The portion of the equation to the left of the equal sign is expressed in the natural logarithm of the probability odds. This transformation places the ability estimates and the difficulty calibrations calculated from the above equation onto a log-linear scale. The units of this scale are "logits" or log-odds-units.

Rasch model statistics are:

### **Item calibrations**

Item calibrations represent the degree of difficulty of the item on the logit scale. This scale generally runs from about -3.0 to +3.0 and is an equal-interval scale. Items with a negative difficulty are relatively easy, whereas items with positive difficulty are relatively hard. Because of the nature of the model, items retain their relative difficulty on the equal-interval scale regardless of the test population used for analysis. This means that once different test administrations have been equated, item difficulty remains relatively constant and directly comparable across administrations.

Following are some guidelines for interpretation of item calibrations:

- Estimates of item difficulty greater than +2 logits refer to hard items. These items should be reviewed to determine that they have not been miskeyed or have more than one correct answer.

---

<sup>7</sup> Wright, B. D., & Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement*, 21(4), 331-345.

- Estimates of item difficulty less than  $-2$  logits refer to easy items. They may have obvious answers, have poorly written distractors or be cued by information in other questions.

### **Rasch Fit Statistics**

The data-to-model fit statistics check how closely the performance of an item approximates the expectations of the Rasch model that more-able candidates will perform better on the item than less-able candidates and that harder items will be answered correctly less frequently than easier items.

The fit statistics are mean-squared residuals. The residuals are calculated as the ratio between the observed value and the value predicted by the model. The expected value of the mean-squared residual is 1. Values greater than 1.2 indicate more than 20% unexplained noise in the item calibration.

Fit statistics are also reported in a standardized format with an expected mean of zero and a standard deviation of 1. The interpretation of standardized fit statistics is similar to the interpretation of a z-score. Items with fit statistics greater than 2.0 or less than  $-2.0$  differ more than two standard deviations from expected performance.

Depending upon the test, either mean-squared residuals, standardized mean-squared residuals or both are reported.

Guidelines for the interpretation of the fit statistics are as follows:

- High mean-squared residuals (Fit  $> 1.2$ ) and high positive standardized residuals (Fit  $> 2.0$ ) indicate that the item has poor discrimination. The item may be worded ambiguously or have multiple correct answers
- Low mean-squared residuals (Fit  $< .80$ ) and high negative standardized fit statistics (Fit  $< -2.0$ ) indicate that the item is highly discriminating. These items should be reviewed, however, to make sure that they 1) do not contain especially seductive distractors, 2) require special knowledge to answer correctly 3) involve a common misconception, or 4) are cued by another item.

## **5. Overall Review of Item Statistics**

---

Item statistics are reviewed after each examination or in the case of an adaptive test, on a periodic basis. Any item that exhibits unusual statistical performance is flagged for further review by content experts. Content experts must make the final determination as to whether items will be retained, deleted from scoring, reworked or retired from the item bank.

---

## 6. Item Writing References

---

- Crehan, K., The validity of two item-writing rules. *Journal of Experimental Education*. 1991;59(2):183-192.
- Cross, L.H., Grading Students. *ERIC/AE Digest*. 1995;95(5) (October):2.
- D'Costa, A.G., Watson, J.E., A critical-incident technique for developing criterion-referenced tests. *Educational Technology*. 1983 July:13-16.
- Frary, R.B., Hints for Designing Effective Questionnaires. *ERIC/AE Digest*. 1996;96(8) (November):2.
- Frary, R.B., More Multiple-Choice Item Writing Dos and Don'ts. *ERIC/AE Digest*. 1995;95(4) (October):2.
- Guttman, L, Schlesinger I., Systematic construction of distractors for ability and achievement test items. *Educational and Psychological Measurement*. 1967;27:569-580.
- Haladyna, T.M., Downing SM. Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*. 1989;2(1):51-78.
- Kehoe, J., Writing Multiple-Choice Test Items. *ERIC/AE Digest*. 1995;95(3) (October):2.
- Lunz, M.E., *Examination Development Handbook*. Chicago, IL: American Society of Clinical Pathologists Board of Registry; 1993.
- Millman, J., Greene, J., The specification and development of tests of achievement and ability. In: Linn RL, ed. *Educational Measurement, Third Edition*; 1989:335-366.
- Osterlind, S.J., *Constructing Test Items*. Boston: Kluwer Academic Publishers; 1989.
- Stone, M. Steps in Item Construction. *Rasch Measurement*. 1997;11(2) (Autumn):559.
- Test Plan for the National Council Licensure Examination for Registered Nurses*. National Council of State Boards of Nursing, Inc.; 1995.
- Trevisan, M.S., Sax, G., Michael, W.B., Estimating the optimum number of options per item using an incremental option paradigm. *Educational and Psychological Measurement*. 1994;54(1):86-91.
- Williams, R.G., Haladyna, T.M., Logical operations for generating intended questions (LOGIQ): A typology for higher level test items. In: Roid G, ed. *A Technology for Test-Item Writing*. New York: Academic Press; 1982:161-186.
-