Examination Development Guidelines

Compiled by:
Mary E. Lunz, Ph.D.

Measurement Research Associates, a division of Measurement Incorporated
180 N Michigan Ave
Chicago, IL 60601

Phone: 312-263-9411
Fax: 312-263-9413
Website: MeasurementResearch.com

1. INTRODUCTION

This handbook describes the procedures for writing, evaluating and reworking objective multiple choice items for examinations. The purpose of the examination process is to assess the competence of individuals to practice selected functions effectively and efficiently. Each examination is constructed through the coordinated efforts of the appropriately qualified item writers and examination developers who are subject matter experts, and psychometric experts who provide information on the measurement characteristics of the items and/or tests. In order to maintain the currency and relevancy of the examination, item development is an ongoing process.

2. PHILOSOPHY OF CRITERION-REFERENCED TESTING

Criterion-referenced testing has emerged over the past several decades as a multi-faceted concept (Berk, 1980). Generally, a criterion-referenced examination is designed to ascertain an individual's competencies. Content guidelines for each examination are often derived from the results of processes and procedures inventories, competency verification studies, job analysis studies, professional levels definitions and expert opinion of professional practicing in the field. These content guidelines link the skills and knowledge (theoretical and practical) expected of a competent practitioner. Because a test score from a criterion-referenced test is interpreted as a measure of how well a candidate performs in relation to the range of tasks and content domains represented by the test items, rather than the performance of other candidates, content guidelines must be carefully delineated prior to item development or test construction.

In a norm-referenced test, a table or schema is generally used to define the content areas to be measured (Ebel & Frisbie, 1991). However, decisions concerning whether the test functioned as intended are entirely determined by the candidate performance statistics on that test. A norm group of candidates is used to set the standard, and decisions to pass or fail are made by comparing the performance of the candidates to that of the norm group.

In criterion-referenced testing, analysis of the examination occurs both before and after the administration of the test. Statements of competence with clearly delineated content provide a basis for writing items. The standard is established on a benchmark scale and indicates the
expectations for competent performance by candidates. Post-examination item statistics are calculated and provide data concerning the interaction of the test population with the items. The items are content valid, because they are written to be representative of the content domain. This assumes that item writers work from established content guidelines that are valid for the field of practice.

The principles of criterion-referenced testing apply to an examination process which has specific characteristics and limitations. (1) The processes are comprehensive and include educational and experience requirements. (2) All candidates, regardless, of the route through which they qualify, must pass the examination. (3) The processes are designed to identify those persons who are capable of meeting acceptable standards of practice. Under the guidelines of the criterion-referenced test philosophy, it is incumbent upon the subject matter experts, to determine the task, content and cognitive skill areas to be tested in each item.

Criterion-referenced examinations are usually composed of examination items that

are representative of the field of practice and written to measure the knowledge and skills of qualified candidates.   In addition, the items are statistically evaluated to ascertain that they measure what they purport to measure, are appropriate for the test population, minimize the amount of test error, and are coherent in style and format.

The following are more complete definitions of the components of the philosophy of criterion-referenced testing.

Conceptualization of the field of Practice requires evaluation of the knowledge, skills, and ethical standards an individual must possess to meet the standard. This is accomplished through consensus of expert opinion, job analysis surveys, and field validation of processes and procedures. General statements serve as a model of professional performance are developed from this analysis.

Knowledge may be defined as the ability to identify, differentiate, conceptualize, classify from rules, principles, processes, operations, and strategies. It represents the information or content base upon which the field of practice is built.

Task may be defined as a purposeful activity. The implication is that the tasks can be defined and that one's ability to perform them can be measured.

Value underlies task and knowledge. Value indicates the importance or worth of the task or knowledge to the individual within the field of practice. Values are considered the basis for ethics and attitudes related to quality performance.

Cognitive skill is the level of mental process used by the candidate to determine the correct response to a item (See the section on Taxonomy Levels for more complete information.)

Item writing is an arduous task requiring not only mastery of the subject matter, but also an understanding of the examination population and mastery of verbal communication skills. The review process insures that the item adheres to appropriate technical and/or scientific principles (*Standards for Educational and Psychological Testing*, 2014). Items are selected by a group of experts for inclusion in the examination data bases.

3. RESPONSIBILITIES OF ITEM WRITERS

Responsibilities of item writer include, but are not limited to the following.

1. Developing new items on a continuing basis, as assigned.

2. Reviewing and selecting items for inclusion in the written examination.

3. Monitoring the content, task, and cognitive skill distributions of items.

4. Monitoring the content quality and difficulty of each item and avoiding duplicate items on the same knowledge/skill.

5. Providing expert input into the criterion standard against which candidates are measured.

6. Reviewing the performance of each item to ascertain the quality of the content and structure of the item.

The goal is to maintain a pool of examination items which are appropriate to measure the knowledge and skills necessary for safe and effective performance in the field of practice.

4. HOW TO CONSTRUCT MULTIPLE CHOICE ITEMS

Definition of a Multiple Choice Item:

A multiple choice item is designed for objective measurement and contains a STEM and four RESPONSES, one of which is the best answer (Airasian, 1991). The multiple choice item is unique in that the standard by which the best answer is selected is contained in the stem. Also, the best answer does NOT have to be the one and only indisputably correct response to the item, as long as the subject matter expert agree it is the best answer of those presented. The form is flexible so that items may be based on items, situations, laboratory results, etc. The following sections outline techniques for writing and evaluating multiple choice items by considering first the stem and then the responses.

Stem

The stem of a multiple choice item may:

    1. ask a question

    2. give an incomplete statement

    3. state an issue

    4. describe a situation

or any combination of the above.

The content of the stem focuses on a central theme or problem, using clear and precise language, without excessive length which can confuse or distract candidates (Gronlund, Linn, & Miller, 2012; see also Downing & Yudkowsky, 2009). The stem may ask a straight forward question, present a scenario or describe data or laboratory results. The question or issue presented in the stem should be relevant to the knowledge and skill level of

the population being evaluated.

Sentence structure in the stem should be grammatically accurate and logically related to the responses (Gronlund et al., 2012). It should present all relevant information to insure clarity and understanding.

Although the multiple choice item format is brief, sufficient information to make an interpretation, answer the question, or solve a problem must be included (Nitko, 1983). Avoid superfluous information, but be certain that all necessary details are included (Downing & Yudkowsky, 2009). Also avoid the use of personal pronouns such as "you" which are inappropriate and perhaps confusing.

As a general principle, the stem should be stated in a POSITIVE form (Downing & Yudkowsky, 2009). Negative statements are not characteristic of normal thought processes, and consequently may place the candidate who is attempting to decipher the item at a disadvantage.

Responses:

1. The "BEST" answer is the response the author and other experts consider the most appropriate answer.

2. The "DISTRACTORS" are logical misconceptions of the best answer.

Each multiple choice item should have mutually exclusive responses (Ebel & Frisbie, 1991).

The plausibility of the responses is the first consideration (Gronlund et al., 2012). The best answer should be the one agreed upon by the experts; however, the other three distractors should also seem plausible to the candidates who have partial, incomplete or inappropriate knowledge. The distractors may therefore be considered logical

misconceptions of the best answer. The responses should be parallel in content length, and category of information.

The grammatical structure of all the responses should be a logical conclusion to the situation, question, or statement presented in the stem (Gronlund et al., 2012).

When writing distractors, it is wise to avoid the use of superlatives such as "always" and "never" (Downing & Yudkowsky, 2009). Such words lead candidates from the response as they tend to be associated with suspect or exaggerated statements.

Repetitive language within the responses should be avoided (Wood, 1960). Words which are repeated in every response may be placed in the stem. Thus, the candidate has less to read and is less likely to be confused by the structure.

The length of each response should be approximately the same (Gronlund et al., 2012; see also Downing & Yudkowsky, 2009). There is a tendency among item writers to make the best answer the longest answer. Testwise candidates may key to this fact and answer correctly because of the format of the response.

Each distractor should be mutually exclusive and not overlapping (Wood, 1960). For instance, if a series of percentages is to be used for the responses, each range must be unique to the response. The following example illustrates this:

a. 10 - 20
b. 30 - 50
c. 55 - 60
d. 65 - 75
e. 76 - 100

If responses are overlapping, the candidate may not be able to determine the best answer not because they do not know the answer, but because the answer is incorporated into more than

one response. In addition, the candidate may be able to argue that more than one response is correct due to the overlap.

Avoid using "none of the above" as a response (Downing & Yudkowsky, 2009). This response does not test what the candidate knows, but only that he/she can recognize that the correct answer is **not** present.

For example:

What is the capital of Texas?

    1. Kansas City

    2. Pasadena

    3. New York

    4. None of the above

(the candidate confidently selects none of the above= because he/she *thinks* he/she knows that the capital of Texas is Lubbock)

Avoid using "all of the above" as a response (Downing & Yudkowsky, 2009). Essentially, this is an overlapping response, because it requires the candidate to consider the responses in combination. Knowing that two are correct leads a clever candidate to "all of the above" without knowing the importance or correctness of the remaining responses.

5. CRITERIA FOR EVALUATING MULTIPLE CHOICE ITEMS

The following may be used as criteria for reviewing an item.

Stem:

1. Is the statement in the stem positive?

2. Are the clinical data or laboratory information clearly described?

3. Is the fact to be recalled, the data to be interpreted or the problem to be solved clearly requested?

4. Does the stem contain excessive description which may be confusing?

5. Does the stem contain judgmental words such as "useful", "best", "treatment of choice" to provide some guidance for the candidate.

6. Do personal pronouns such as "you" appear in the stem?

7. Does the stem contain extraneous cues to the best answer?

8. Is the wording ambiguous?

9. Does the format of the item meet the criteria for the appropriate cognitive skill level?

Responses:

1. Do the responses appear in a logical order, if a logical order exists (example: time sequence)?

2. Do unnecessary words or phrases appear in each response that could be included in the stem?

3. Are the responses independent and mutually exclusive (not overlapping)?

4. Is the length of the response appropriate?
        Is the response too short to be clear?
        Is the best answer the longest or shortest response?
        Are all of the responses approximately the same length?

5. Are all responses plausible alternatives?

6. Are all the responses parallel in structure (e.g. all treatment or diagnoses)?

7. Are "none of the above" or "all of the above" used as responses?

6. ITEM CLASSIFICATION

   The classification of items assists the examination developer in monitoring the distribution of items across content and task domains as well as cognitive skill levels. Each item in the examination data base is assigned a classification code which is also used in the computerized item selection algorithm. There are four components to this classification; these components are as follows:

1. Content: Content reflects the major subject category of the item. It is the content classification that is used in selecting items to insure that the entire content domain is covered.

2. Task: Task is the skill performed (e.g. diagnosis or treatment).
   By structuring items to reflect different tasks, a greater variety of items are generated.

3. Taxonomy: Taxonomy refers to the cognitive processes required to answer the item (Bloom, et al, 1956). The construction of the stem and responses, utilization of visual materials as well as the process and content of the item all contribute to the classification of an item by taxonomy level. The following three taxonomy levels are common:

Taxonomy 1 - **Recall**: ability to recall or recognize previously learned (memorized) knowledge ranging from specific facts to complete theories.

Taxonomy 2 - **Interpretive skills**: ability to utilize recalled knowledge to interpret or apply verbal, numeric or visual data.

Taxonomy 3 - **Problem Solving Clinical Judgment**: ability to utilize recalled knowledge and the interpretation/application of distinct criteria to resolve a problem or situation and/or make an appropriate decision.

7. ITEM EVALUATION

The purpose of the item evaluation is to identify items that are **not** measuring as expected. Items that fail to perform properly increase the error of the exam and therefore do not contribute to the precision of the pass/fail decision made about candidates. After items are presented on a test, they are subjected to statistical, as well as, content analysis. The statistical analysis provides clues for the subject matter experts with regard to how well the content of the item yielded useful information about candidate ability.

The purpose of deleting items from an examination is always to create more precise and fair examinations. Any item that performs poorly is flagged for possible deletion. Items may perform poorly for many reasons. Many of these reasons are related to the initial construction of the item stem and responses. Proper development of item stems and responses leads to a higher probability that the item will perform successfully.

8. A QUICK SUMMARY OF ITEM ANALYSIS

Traditional Item Analysis

In the process of item review, the item statistics represent the performance of the item and provide guidance to the examination reviewers when revising items. Traditional item analysis consists of a p-value and point biserial correlation (RPBI). The p-value is the percent of candidates who selected each response. Hopefully, more candidates selected the keyed correct response than any distractor.   The point biserial correlation is the correlation between the performance of the candidates who answered the item correctly and the candidates who did well on the total test. The point biserial correlation should be positive and higher for the keyed correct response and negative for the distractors. This pattern suggests that candidates who did well on the test tended to select the correct answer on the item. More examples are provided in Appendix A.

The ideal ranges for the item statistics are as follows:

> (1) p-value: generally in the range of .30 to .80

> (2)   RPBI: around .20 for the correct answer and negative for all distractors

Item Response Theory

The Rasch Model Item Statistics

Rasch model analysis is based on the probability that a person with a given ability level will answer correctly a question representing a given difficulty or item calibration (Wright & Stone, 1979). If a person's ability is greater that the item's difficulty, the probability is <u>more</u> than 50% that the person will answer the item correctly. Conversely, if the person's ability is less than the difficulty of the item, the probability is less than 50% that the person will answer the question correctly. If the person's ability is equal to the difficulty of the item, the probability is 50% that the person will answer the question correctly.

The general equation that corresponds to this expectation is $\log(P_{ni}/1-P_{ni})=(B-D)$, where P_{ni} is the probability person (n) answering item (i) correctly and $1-P_{ni}$ is the probability of person (n) answering item (i) incorrectly. The probability of correct response depends upon the ability of the person (B) and the difficulty of the item (D). The difference between B and D is the basis for expecting a correct response.

Another way of looking at the ratio $\log(P/1-P)$ or log (right/wrong) is that it represents the "odds" of getting the question correct, or the log-odds since all measures are transformed to a log-linear scale. Thus the unit of measure is referred to as a "logit" (log-odds-unit). See **Appendix B** for sample output.

The Rasch item difficulties are represented in log-odds units (logits) and will generally fall within the range of -3.0 to +3.0 logits on the log-linear scale (mean of the scale is zero). An item with a negative logit value tends to be easy, while an item with a positive

logit value tends to be hard. The item calibration, or position of the item on the scale,

represents the relative difficulty of the item with respect to the other items on the scale.

Items retain their _relative_ difficulty on the equal interval scale, regardless of the ability of the

population that challenges the item. Thus, a difficult item will be difficult, relative to the

other items, an easy item will be easy, relative to the other items, regardless of the ability of

the population.

For certification examinations, items are usually targeted to the pass point, so that the

most information is gained about the candidate is gained about the candidate for whom

decisions are most difficult to make. Other forms of test administration, such as adaptive

testing, may target item difficulty to the current estimated ability of the candidate,

so the error of measurement is reduced more quickly for all candidates. Information which

can be derived from the Rasch item calibration statistics is as follows.

1. Items that are very hard (high positive values: +2 to +3 logits) suggest that the content
matter may be esoteric, or present new information that hasn't received wide dissemination
to the field.

2. Hard items (+2 to +3 logits) may be ambiguously worded such that there is no correct
answer or there are two or more acceptable answers.

3. Hard items (+2 to +3 logits) may deal with procedures that are rapidly being replaced by
new advances in the field and are thus becoming obsolete.

4. Easy items (high negative values: -2 to -3 logits) may have poorly written distractors.

5. Easy items (-2 to -3 logits) may have been keyed by other items on the test. For example,

the correct answer to the question may be in the stem of another question.

Every measurement system has some measurement error associated with it. This is

true for multiple choice items. The calibrated difficulty of any item is the calibration plus or

minus the associated error of measurement. Thus an item with a difficulty calibration of

1.00 and an measurement error of .20 actually has a difficulty ranging from .80 to 1.20 ($\Box$ 1

S.E.). Efforts are made to control measurement error by constructing and reviewing items

carefully; however, error cannot be completely eliminated and therefore must be considered

in the evaluation of item performance as well as candidate ability estimates.

The Rasch item fit statistic estimates how closely an item corresponds to the

expectation of the Rasch model (candidates who are more able have a high probability

answering the item correctly, while candidates who are less able have a lower probability of

answering the item correctly). The **fit statistic** compares the observed response distribution

of an item with the expected response distribution, similar to a Chi-square analysis. The fit

statistic is a standardized residual difference between expected and observed. The

difference between the observed and expected should be zero. However, not all items

measure as expected, so items two standard deviation units away from zero are generally

considered not to fit the expectations of the Rasch model.

9. REWORKING ITEMS

One of the responsibilities of subject matter experts is to revise or rework items. Both previously tested and new items may require revision. New items are reviewed to verify item clarity, accuracy, content and structure. If items require revision, the subject matter experts often determine, by consensus, the final revision (two heads are sometimes better than one). Once tested, items with poor statistical performance must be rewritten before they can be used again. The p-values are useful for assessing the contribution of each response. In addition, lower taxonomy items may be reworked into items representing higher taxonomy levels. While it is a vital part of the testing process to examine basic knowledge using recall items, data interpretation, diagnosis, and problem solving items come closer to testing critical abilities of a candidate.

The following sample items are laboratory-oriented. However, they provide examples of how item analysis can assist with item revision, and how lower taxonomy items can be rewritten as higher taxonomy items.

Poorly Constructed Distractors

   If a high percentage of the more able candidates *incorrectly* chose a distractor instead of the best answer, the item probably needs reworking. The p-values for the responses are acceptable; however, there is a low point biserial correlation (RPBI) for the best answer, positive discrimination for distractor b and low negative discrimination on distractor d.

Item Stem:

Results found 50% of the cells on a WBC differential were interpreted as blasts. Cytochemical stains were performed with the following results:

Peroxidase	Chloroacetate esterase	PAS
++	++	+/-

The blasts are probably:

 *a. myeloblasts
 b. monoblasts
 c. lymphoblasts
 d. undifferentiated blasts

	-A-	-B-	-C-	-D-
P-VALUE	0.61	0.18	0.09	0.12
RPBI	0.12	0.05	-.20	-.05

When reworking this item, the item writer must attempt to replace b with a response less likely to be confused with the correct answer, perhaps "erythroblasts" for **b** and "unclassified" blasts" for **d.**

Item Too Easy

Item Stem:

Which of the following tests is consistently prolonged by SMALL doses of aspirin?

  a. prothrombin time
  b. partial thromboplastin time
  c. thrombin time
*d. bleeding time

	-A-	-B-	-C-	-D-
P-VALUE	0.03	0.30	0.30	0.91
RPBI	-.17	-.12	-.11	0.25

The item is too easy for the candidate population (91% answered correctly) and, therefore, may not be the best item to measure the ability of the candidates. By attempting to make distractors with a low p-value more plausible to the minimally competent candidate, this item should become more difficult. For example, change **a** to clot retraction and change **c** to glass bead retention (adhesion) test. The more difficult responses may make the items more appropriate for the ability of the candidates.

Recall To Interpretation

Items which require the candidate to recall facts (taxonomy 1) can be upgraded to interpretation (taxonomy level 2) by referring to additional data.

Example Recall Item:

The best negative control for cetrimide agar is Pseudomonas:

    a. aeruginosa
    b. fluorescens
    c. putida
*d. maltophilia

REWORK

Item Stem:

A technologist observes the following quality control data on cetrimide agar:

Positive control: Pseudomonas maltophilia inhibited

    These results indicate:

    a. inappropriate quality control organisms
    b. appropriate quality control results
    c. inappropriate color of colonies
*d. mixed positive quality control organism

Using the content of a taxonomy level one item, but applying that concept to a specific situation, many taxonomy level two items can be written.

Recall To Problem Solving

This same recall item can be rewritten as a problem solving item by creating an item that poses a problem in need of a solution:

REWORK:

  Item Stem:

  A technologist observes the following quality control data on cetrimide agar:

                                      PIGMENT

  Positive control: Pseudomonas aeruginosa   blue-green
  Negative control: Pseudomonas fluorescence   clear

      The technologist should:

      a. repeat test with a Pseudomonas putida negative control and record odor.
      b. re-incubate plate for 24-hours at 25 C.
    *c. repeat test with a Pseudomonas maltophilia negative control and record growth.
      d. repeat test with a fresh subculture of control organisms.

By using the content from the taxonomy level one item, and a more complex task, then

interjecting aberrant results or sources of error (e.g. an incorrect negative control),the item

can be rewritten as a problem-solving item.

10. PRETESTING ITEMS

There are two methods of pre-testing items to insure that they will measure effectively. The first is **field testing,** and the second is **key validation**.

 Field testing allows the test developer to include new untested items in a test for the purpose of obtaining item difficulty data and verifying the validity of the item. Field test items are not scored as part of the test.

 The following characteristics should be met by field test items.

1.   Field test items are developed and selected from well defined explicit content and process domains.

2.   Field test items are carefully reviewed and modified by content experts for clarity, style, content and quality.

3.   Field test items are appropriately and evenly distributed over the content domains of the examination.

4.   Content experts scrutinize the new items carefully after the initial scoring of the examination. Items that did not perform as expected can be revised before they are used in the scored part of the examination. Items that cannot be reworked should be deleted.

The purpose of field testing is to ascertain that the item is measuring the content or concept, as the item writer and reviewers anticipated. Therefore, the item is pre-tested to collect data so that the item difficulty can be calculated. Assuming the item performs well, it will be added to the item bank for use on future examinations.

Key validation is somewhat different than field testing. This method can be used on fixed length written examinations only. After new items are developed and reviewed, they are placed on a form of the examination without prior testing. If they meet the criteria

for satisfactory performance, they are scored with the other items on the test. If they do not meet the criteria, they are deleted from scoring, and revised. When key validation is used, it is necessary to include more items than are needed to meet the test specifications. Usually, there is only a 50% probability that the untested items that are being key validated items will perform successfully and be scored. Key validation is a useful method of testing large numbers of new items. The items to be key validated should be carefully reviewed for accuracy of content and structure before they are included on the examination. They must also fit the test specifications.

References

Airasian, P.W. (1991). *Classroom assessment*. NY: McGraw-Hill.

Berk, R.A. (1980). *Criterion-referenced measurement*. Baltimore, MD: John Hopkins Press.

Bloom, B.S., Englehart, M.D., Furst, E.J., Hill, W.H. and Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1; Cognitive domain.* NY: David McKay Co.

Downing, S.M. & Yudkowsky, R. (Eds.). (2009). *Assessment in health professions education*. NY: Routledge.

Ebel, R.L. and Frisbie, D.A. (1991) *Essentials of educational measurement*. (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Gronlund, N.E., Linn, R.L., & Miller, M.D. (2012). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Hambleton, R.K., Swaminathan, H. and Rogers, H.J. (1991).   *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Nitko, A.J. (1983). *Educational tests and measurement*. NY: HBJ.

Standards for educational and psychological testing. (2014). Joint Committee of the AERA, APA, & NCME.

Wood, D.A. (1960). *Test construction*. Columbus, OH: Merrill Books.

Wright, B.D. and Stone, M.K. (1979). *Best test design*. Chicago: MESA Press.

Appendix A

```
Sample Item Analysis

Item Number  1     Correct answer =   D     Accession Number = 3763
        DISTRACTOR    A      B      C      D      E    NONRESPONSE
        N OF PEOPLE   39     17     67     48      6      0
        PVALUE       0.22   0.10   0.38   0.27   0.03   0.00
        PT BISERIAL +0.17  -0.09  -0.14  +0.06  -0.00  +0.00

Item Number  2     Correct answer =   B     Accession Number =  449
        DISTRACTOR    A      B      C      D      E    NONRESPONSE
        N OF PEOPLE   49     78     43      3      4      0
        PVALUE       0.28   0.44   0.24   0.02   0.02   0.00
        PT BISERIAL -0.18  +0.21  -0.06  +0.05  -0.02  +0.00

Item Number  3     Correct answer =   D     Accession Number =  952
        DISTRACTOR    A      B      C      D      E    NONRESPONSE
        N OF PEOPLE    2     17     38    109     11      0
        PVALUE       0.01   0.10   0.21   0.62   0.06   0.00
        PT BISERIAL -0.09  -0.02  -0.05  +0.14  -0.13  +0.00

Item Number  4     Correct answer =   C     Accession Number = 2057
        DISTRACTOR    A      B      C      D      E    NONRESPONSE
        N OF PEOPLE    8     27    100     42      0      0
        PVALUE       0.05   0.15   0.56   0.24   0.00   0.00
        PT BISERIAL -0.05  -0.09  +0.09  -0.01  +0.00  +0.00

Item Number  5     Correct answer =   D     Accession Number = 1268
        DISTRACTOR    A      B      C      D      E    NONRESPONSE
        N OF PEOPLE   14     24     51     86      2      0
        PVALUE       0.08   0.14   0.29   0.49   0.01   0.00
        PT BISERIAL -0.16  -0.10  -0.04  +0.21  -0.06  +0.00
```

Item Number - sequence number of item on the test
Correct answer - keyed correct response
Accession Number - pool number
Distractor - names of responses, usually a - e or 1 - 5
N of People - number of people who selected each response
P-Value - percent of candidates who selected each response
Point Biserial Correlation - discrimination index

Appendix B

```
Sample Rasch Item Output
+----------------------------------------------------------------------+
|ENTRY    RAW                         |   INFIT  |  OUTFIT  |SCORE||      |
|NUMBER  SCORE   COUNT   MEASURE  ERROR|MNSQ  ZSTD|MNSQ  ZSTD|CORR.|| ITEM|
|--------------------------------------+----------+----------+-----++-----|
|    1    262    304    -.79     .17| .98   -.1| .96   -.2| .20|| 9323|
|    2    269    304   -1.01     .18| .97   -.2| .85  -1.0| .25|| 2004|
|    3    255    304    -.78A    .17|1.11  1.0|1.04   .3| .23|| 9705|
|    4    170    304     .88     .12|1.04  1.4|1.05  1.4| .17|| 2002|
|    5    108    304    1.78     .12|1.08  1.8|1.13  2.5| .09|| 9306|
|    6    206    304    -.12A    .14|1.33  4.2|1.47  4.6| .09|| 9311|
|    7    208    304     .58A    .12|1.00   -.1|1.00   .1| .13|| 9313|
|    8     72    304    2.39     .14| .99   -.1| .99   -.1| .24|| 9315|
|    9    290    304   -2.02     .28| .99    .0| .90   -.3| .14|| 9317|
|   10    128    304    1.23A    .12|1.07  2.4|1.07  2.0| .12|| 9327|
|   11    168    304     .83A    .12| .98   -.5| .98   -.6| .30|| 9316|
|   12    244    304    -.18A    .14| .91  -1.2| .92   -.9| .20|| 9401|
|   13    195    304     .29A    .13| .99   -.2| .96   -.6| .41|| 9320|
|   14    272    304   -1.11     .19| .97   -.2| .93   -.4| .21|| 9321|
|   15    172    304     .86     .12| .96  -1.3| .95  -1.5| .33|| 9330|
|   16    269    304    -.98A    .18|1.00    .0| .99   -.1| .13|| 9403|
|   17    231    304    -.33A    .15|1.12  1.3|1.12  1.1| .26|| 9406|
|   18    231    304     .43A    .12| .87  -3.0| .87  -2.3| .20|| 9407|
|   19    160    304    1.03     .12|1.01    .4|1.00   .1| .24|| 9409|
|   20    193    304     .54A    .12|1.02    .5|1.02   .5| .21|| 9325|
|   21     79    304    2.26     .13| .96   -.6| .94   -.8| .31|| 9408|
|   22    167    304    1.05A    .12|1.00    .0|1.01   .2| .24|| 9604|
|   23    256    304    -.63     .16| .97   -.3| .95   -.4| .24|| 9508|
|   24    265    304    -.88     .17|1.01    .1| .97   -.2| .16|| 2001|
|   25     73    304    2.37     .14|1.01    .2| .99   -.2| .22|| 9513|
|   26    220    304     .56A    .12| .87  -3.5| .85  -3.2| .30|| 9601|
|   27    183    304     .70     .12| .99   -.3| .99   -.2| .26|| 9606|
|   28    187    304     .64     .12|1.04  1.1|1.04   .8| .17|| 9702|
|   29    163    304     .90A    .12| .99   -.3| .97   -.8| .29|| 9703|
|   30    150    304    1.17     .12| .96  -1.5| .97  -1.0| .33|| 2003|
|--------------------------------------+----------+----------+-----++-----|
| MEAN   195.   304.     .39     .14|1.01    .0|1.00    .0|     ||      |
| S.D.    61.     0.    1.08     .03| .08  1.4| .11  1.4|     ||      |
+----------------------------------------------------------------------+
Entry Number = sequence number
Raw score = number of candidates who answered correctly
```

Count = number of candidates who answered the item

Measure = calibrated difficulty of the item

Error = measurement error associated with the item

Infit statistics – consistency

Outfit statistics – consistency with regard to guessing

Score correlation = similar to point biserial correlation

Item = item pool or accession number