*Shiken: JALT Testing & Evaluation SIG Newsletter. 13* (1). January 2009 (p. 35 – 41)

# Suggested Answers for Assessment Literacy Self-Study Quiz #6

by Tim Newfields

*Here are some suggested answers for the questions about testing, statistics, and assessment from the January 2009 issue of SHIKEN. If you feel an answer is unclear or conclusion is incorrect, please contact the editor.*

## *Part I: Open Questions*

**1 Q:** What key information is missing from the graph by the Japanese Ministry of Education, Culture, Sports, Science and Technology at http://jalt.org/test/Graphics/SSA6a.gif? What problems are likely to arise in interpreting this? How should this graph be revised?

**A:** First of all, the title of the graph is vague: we need to know precisely how many students from how many schools in what year offered this information – and also the response format and language of this survey. Sample size, sampling characteristics, response language, and response format all impact survey results. The title of any figure or table should be detailed enough to enable those skimming through an article to grasp the essential information.

If this survey is based on a rank-order forced-choice response format, the rationale for dividing the information into 17 categories should be explained. Some of the categories appear to overlap and if a factor analysis of the data were employed, it may be possible to truncate some categories. In fact, to make this information easier to comprehend it's probably best to compact the data into about seven items, since that is about all that most readers can remember. Minor factors could be truncated into a category such as "other reasons [*hoka no riyu*]".

Moreover, since this information is probably based on a Japanese language survey, it is best to list each category in a bilingual format since some of the Japanese concepts do not render smoothly into English – another option would be to have one version of the table in Japanese and another in English.

Finally, it's good to remember respondents will have a tendency to mark items at the top of the list more often than those are at the bottom. To compensate for this tendency, it's probably best to design at least two survey forms: one in which the items at the top in one form are shifted to the bottom in the other.

Here is a suggested revision of the graph:

> *Figure 2.3*. Reasons that [sample size] first year university students at [number] universities Japan indicated for selecting their universities in [year] based on a [mention format type] questionnaire in [language].
>
> Indicate the top 7 responses in this graph in a bilingual format.
> Be sure to include "no response [*kaitou nashi*]" and "other reasons [*hoka no riyu*]".
>
> To make this information richer, the x-axis should display not only the percentage of the total responses; it should also mention the actual numbers of responses in each category. For example, if the total sample size were 1,000 then beneath "25%" the number 250 would appear. This will give readers a rough idea of the statistical power of this survey.
>
> Finally, the precise source of the original study from which this data was obtained should also appear in small print below the figure. This is "orphaned" information and there is no way to corroborate it.

**2 Q:** In the context of multiple-choice testing, what does "proportionality" refer to and how is this concept important to test designers?

**A:** Ever since Tversky's (1964) pioneering paper into the optimal number of alternatives for multiple choice tests, researchers have employed various methodologies to ascertain how many distracters should appear in a MC exam. Obviously, as the number of distracters increases the chances of merely guessing the correct answer decreases. However, the efficiency of each item distracter also tends to diminish as choices increase. In short, test writers must consider the best way to trade-off distracter efficiency with reduced guessing. Osterlind (1989, p. 156) describes this scenario in terms of *proportionality*.

The body of research accumulated so far suggests that the optimal number of MC options depends on the test purpose as well as the target population. If the purpose of a test is to ensure that a minimum standard of instruction has been achieved and the majority of examinees are expected to know most of the material, then a 3-choice MC format is probably ideal. However, if a test is intended to identify a small group of high-performing candidates and the majority of examinees are likely to guess extensively on the exam, then a 4- or even 5-choice format is probably better.

Further reading:

Abad, F. J. , Olea, J., & Ponsoda, V. (2001). Analysis of the optimum number alternatives from the Item Response Theory. *Psicothema, 13* (1) 152-158. Retrieved December 2, 2008 from http://www.psicothema.com/imprimir.asp?id=427

Bruno, J. E. & Dirkzwager, A. (1995, December). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55* (6) 959-966. Retrieved December 2, 2008 from http:/Knowledgefactor.com/.../Bruno-Optimal Alternatives in Multi Choice.pdf

Budescu, D. V. & Nevo, B. (1985, Fall). Optimal number of options: An investigation of the assumption of proportionality: *Journal of Educational Measurement, 22* (3), 183-196.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology, 1*, 386-391.

**3 Q:** What ethical issues have been compromised in reporting the comparison of these three tests? What information needs to be included to enable teachers and school administrators to make an informed decision regarding the degree of correlation of the Pearson Education exam with the other two ETS exams?

**A:** A valid conversion of one test score into another can only be made if the tests in question measure the same construct. Since the TOEIC, TOEFL, and Longman English Assessment packages are designed for different purposes we should not take it as a given that these tests are tapping the same underlying construct. If they are not, attempts to convert the score of one exam into another is rather like seeking to express the value of lemons in terms of grapefruits or oranges. Since all three of these tests do tap into general English skills to some degree and many of their questions employ the same response format, we can expect that there is some degree of correlation between the scores of TOEIC, TOEFL, and Longman English Assessment (general) package scores. Unfortunately, Pearson Education has not mentioned what the correlation coefficient is. The correlation table would provide a more honest assessment if the Approximate Overall Scores were expressed as a range, such as 90% = TOEFL 440 – 480, or if standardized errors for each approximation were entered in a footnote. In mentioning the correlation coefficients between tests, it is also essential to describe the essential characteristics of the sampling group. The size, nationality, age, gender, and educational background of the sample must be specified when reporting correlation coefficients (APA, 1999, C-4). Two tests that correlate highly for a small group of Japanese undergraduates might not correlate so highly with a different type of group.

Moreover, when reporting test results it is important to indicate the significant limitations of a given test and indicate what it doesn't measure (APA, 1999, C-1). Unfortunately, most publishers are so intent on emphasizing what a test can do that they are remiss in reporting what it can't do.

Finally, test publishers have a responsibility to indicate what constitutes an inappropriate use of the products they develop. For example, using the Longman English Assessment packages to screen applicants for overseas study at a North American university would amount to test abuse since the LEA was designed to ascertain general English ability - not proficiency in academic English at the university level.

Further reading:

American Psychological Association Joint Committee on Testing Practices. (1999). *Code of Fair Testing Practices in Education*. Retrieved December 28, 2008 http://www.apa.org/science/fairtestcode.html

Pearson Education, Inc. (2005). *Longman English Assessment*. Retrieved December 28, 2008 from http://www.pearsonlongman.com/ae/multimedia/assess.htm

**4 Q:** What is one concrete situation for which each of the following tests would be appropriate?

(A) A mastery test
(B) A power test
(C) A speed test
(D) A -ze test

**A: Mastery tests** are straightforward pass or fail exams that ascertain whether or not a minimum standard of performance has been achieved. Most licensing exams can be regarded as mastery tests: anyone reaching a minimum cut-off point automatically passes and gets the desired license. The Japanese National Tourist Association's Interpreter-Guide Test (日本政府観光局の通訳案内士試験) is one example of such an exam.

According to Osterlind (1989) a **power test** is designed to identify only the highest performers in a given field: it would contain many items that the majority of applicants would not be able to correctly answer. Such a test might be appropriate for admission to a highly desired post or as one criterion in making decisions when awarding limited scholarship funds.

A **speed test** is an exam whose time limit is so short that few, if any, of the examinees will be able to finish all of the tasks. Score differences in such a test would be due primarily to the number of items completed. Such a test might be appropriate when it is important to measure how well job candidates can perform relatively simple tasks with rapidity. For example, in some types of translation work, applicants need to translate long passages rapidly. In a speed test designed for this scenario, none of the material would be difficult to translate, but the amount of material would be daunting.

According to Mousavi (2002, p. 837), a **–ze test** is a modified type of C-test in which the left half (plus one letter) of every other word is deleted after the first sentence, as in this example:

> Christopher Cleary is generally credited with creating the –ze test sometime around 1986.
> ____ead of __ng a ___ht hand ____ion procedure (as _n the –cl test), _he –ze __st uses a left __d deletion ___edure. In Clearly's ___w, this ___e of ___t works ___l for __me types _f students. _____ver, it __es not __em well ___ted to _____tary level _____ents.

Since this type of test is controversial, opinions about its appropriacy are divided. My suggestion would be experiment with its use in a low-stakes context in your own class and see how responses for this test compare with other test formats.

Further reading:

Cleary, C. (1988). The C-Test in English: left-hand deletions. *RELC Journal, 19* (2) 26-35. DOI: 10.1177/003368828801900203

Japan National Tourist Association. (n.d.). *Tsuuyaku Annai-shi Shiken Gaiyou*. [Outline of the Interpreter-Guide Test]. http://www.jnto.go.jp/jpn/interpreter_guide_exams/

Osterlind, S. J. (1989). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats . . .* Boston, Dordrecht, London: Kluwer Academic Publishers.

Rouhani, M. (2008). Another Look at the C-Test: A Validation Study with Iranian EFL Learners. *Asian TEFL Journal, 10* (1) Article 8. http://www.asian-efl-journal.com/March_08_mri.php

-ze test. (2002). In S. A. Mousavi *An Encyclopedic Dictionary of Language Testing*. (3rd Ed.). (p. 837). Taipei: Tung Hua Book Company.

## *Part II: Multiple Choice Questions*

**1 Q:** Which of these statements below is generally true of ANOVA tests? (Hint: Think a bit and don't expect a single, clear answer.)

(A) They examine only how groups and items interact.
(B) They point out pronounced items which differ markedly from the norm.
(C) They examine proportional score differences across total score categories.
(D) They are arbitrary as far as ability levels are concerned.
(E) They can help show whether or not there is a differential pull in terms of specific distracters between groups.

**A:** The wording of this question is rather tricky because it is better to think of ANOVAs as a family of tests rather than a single test. Choice (A) may be considered correct for one-way ANOVAs that describe how three or more groups respond to one instrument or variable. If dealing with several variables or instruments, then this statement could also be true for mixed-design ANOVAs or factorial ANOVAs.

Choice (B) is the focus of the transformed item difficulties (TID) index (also known as the Delta method) developed by Angoff in 1982. According to Roever (2005, pp. 3, 4), this method is no longer favored as it fails to distinguish between item difficulty and discrimination, nor does it indicate examinee ability. Like most ANOVAS, it identifies the interaction of groups and items. According to Osterlind (1983, p. 28) it goes a step further than ANOVAS do by also indentifying items that stand out far from the norm.

Choice (C) is true of some chi-square approaches, but not of ANOVAs.

Choice (D) is generally true of ANOVAs, Delta methods, and chi-square approaches. If one wants to systematically investigate ability levels, then some type of ICC or Rasch measurement would be appropriate. These two approaches could also be used to explore differential pulls between groups, as specified in Choice (E).

Further reading:

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting item bias*. Baltimore and London: The John Hopkins University Press.

Osterlind, S. J. (1983). *Test Item Bias*. (Chapter 2, pp. 20-28). Beverly Hills, London, New Delhi: Sage Publications.

Roever, C. (2005). "That's not fair!" Fairness, bias, and differential item functioning in language testing. *SLS Brownbag Lecture September 15, 2005*. Retrieved on December 28, 2008 from http://www2.hawaii.edu/~roever/brownbag.pdf

**2 Q:** Which statement about the test item in Figure 2 (at www.jalt.test/Graphics/SSA6c.gif) is true?

(A) It discriminates fairly well across all ranges of language ability.
(B) It discriminates fairly well only for mid- and upper-ability students.
(C) It discriminates fairly well only for mid- and lower-ability students.
(D) It does not discriminate well across various language ability ranges.

**A:** This graph suggests that the item discriminates well at mid-range ability. It also discriminates to some degree for those of lower ability, but not so well among higher ability learners. If the test in question has a sufficient number of items that discriminate among upper-ability students, then this test item can be considered adequate for the given sample.

Further reading:

Kelley T., Ebel R., Linacre, J.M. (2002). Item Discrimination Indices. Rasch Measurement Transactions, 16 (3) 883-4. Retrieved on December 28, 2008 from http://www.rasch.org/rmt/rmt163a.htm

**3 Q:** What type of scale is the survey on http://www.jalt.org/test/SSQ6.htm#MC3 based on?

(a) a Thurstone ratio scale
(b) a Guttman interval scale
(c) There is no way to tell by merely looking
(d) an ordinal rank-order scale
(e) a categorical dummy-variable scale

Also, how should this scale be revised to enhance its validity?

**A:** This appears to be an ordinal rank-order scale.

The issue of how to enhance the validity of such a scale is complex. There are three implicit assumptions latent in the Niigata University scale: (1) Most under-graduate students are capable of candidly expressing why they are studying a foreign language, (2) All respondents have at least three important reasons for doing so, and (3) All of those three main reasons are in the list of nine choices available.

Each of these assumptions should be considered critically. Are young university students reflective enough to be able to verbalize why they are studying a foreign language – and will they be candid about their reasons in a survey by their teacher?
Also, how should students with only one or two reasons respond to this survey? (The instructions demand that three reasons be indicated and some students will be tempted to arbitrarily circle a third item just to complete the task.) Also, how should students respond if their reason for studying is listed here?

The current survey design is problematic in a number of ways. Three ways to enhance this survey would be:

(1) to give respondents the option of writing out an additional reason not mentioned among the listed choices available.

(2) to repeat some of the questions in different response formats. For example, respondents could be asked in an open (constructed) response format a question such as:

What is the main reason that you are studying English now?
今英語を勉強している主な理由は何ですか。

Also, some of the survey items should appear as Likert scale opinion-response questions, as in this example:

It is important for me to understand music with English lyrics.
Circle One: Strongly agree somewhat agree mildly disagree Strongly disagree

英語の歌詞を備えた音楽を理解することは私にとって重要です。
1つを選んでください: 強く賛成、やや賛成 やや反対 強く反対

(3) At least two forms of this survey should be created in which items are rotated and opinion-response questions are stated in an antithetical way on the alternative form. For example, the previous opinion statement should appear this way on an alternative form:

It is not important for me to understand music with English lyrics.
Circle One: Strongly agree somewhat agree mildly disagree Strongly disagree

英語の歌詞を備えた音楽を理解することは私にとって重要ではありません。
1つを選んでください: 強く賛成、やや賛成 やや反対 強く反対

This will accomplish two things: (1) compensate for the tendency of respondents to favor responses appearing at the top of a list of choices, and (2) counterbalance the tendency of respondents to agree with the opinions expressed in surveys coming from teachers with whom they must interact with long after the survey is finished.

Further reading:

Ball State University. (1999, November 22). Using surveys for assessment. Retrieved on December 24, 2008 from http://web.bsu.edu/IRAA/AA/WB/chapter3.htm

Brown, J. D. (2001). *Using Surveys in Language Programs* (Cambridge Language Teaching Library) Cambridge: Cambridge University Press.

**4 Q:** In most 3- parameter IRT descriptions of applied linguistic research, what does the third parameter usually refer to?

(a) item difficulty (c) rater severity
(b) item discrimination (d) guessing

**A:** 3-parameter IRT models can be applied to any field, but within the discipline of foreign language education, the first parameter usually refers to item difficulty and the second parameter generally refers to the item discrimination. The third parameter most often refers to guessing. Though this parameter has a theoretical range of 0 to 1, according to Baker (2001, p. 28) only values under .35 are considered acceptable for 4 (or more) choice MC format questions. In other words, if a person has more than a 35% chance of guessing the correct answer in a multiple-choice question something is wrong with the question and/or distracters.

It is good to remember that all the parameters in IRT are interdependent, and in the 3-parameter model item difficulty changes because of guessing.

Further reading:

Baker, F. B. (2001). *The Basics of Item Response Theory*. (Second Edition). ERIC Clearinghouse on Assessment and Evaluation. Retrieved on December 25, 2008 http://echo.edres.org:8080/irt/

**5 Q:** At what point can a Pearson product-moment correlation be interpreted as "strong"?

(a) If it is within a range of .1 (anywhere from ±.90 to ± 1).
(b) If it is within a range of .3 (anywhere from ±.70 to ± 1).
(c) If it is within a range of .5 (anywhere from ±.50 to ± 1).
(d) If it is within a range of .7 (anywhere from ±.30 to ± 1).

**A:** The simple answer would be (b). The Pearson product-moment correlation describes the extent that two variables correlate, with values ranging from +1 and -1. Values between 0 and .3 are said to represent little, if any, correlation. Those between.3 to .5 are thought to signify a low correlation. Values between .5 to .7 are thought to indicate a moderate correlation. Values above .7 are generally interpreted as indicating a "strong" correlation and any correlation above .9 would probably be regarded as "extremely strong". Remember, all values can be either positive or negative.

Simple answers, however, are not necessarily accurate ones. The Pearson product-moment correlation assumes that a normal distribution curve exists for the given data set: a highly skewed or highly kurtotic data would not yield reliable Pearson product-moment correlation values. Moreover, The Pearson product-moment correlation values need to be interpreted in the light of other measures of statistical significance: measures from a small sample could suggest a strong relationship when, in fact, none exists.

It is also good to remember that the Pearson product-moment correlation only works for parametric data, that is data which is assumed to have a reasonably even distribution curve. For non-parametric data, the Spearman Rank Correlation Coefficient is usually used.

Incidentally, if a Pearson product-moment correlation is drawn from data thought to represent an entire population it's generally denoted by the Greek letter rho ( $\rho$ ). However, if it's only based on data for a sample subset of that population, then the Latin letter *r* is used.

Further reading:

AcaStat Software. (n.d.) *Pearson's Product Moment Correlation Coefficient*. Retrieved December 24, 2008 from http://www.acastat.com/Handbook/30.html

Brannick, M. T. (n.d.). *Correlation*. Retrieved December 24, 2008 from http://luna.cas.usf.edu/~mbrannic/files/regression/corr1.html

Cengage Learning. (n.d.). The Pearson Product Moment Correlation Coefficient. Retrieved December 24, 2008 from http://www.wadsworth.com/psychology_d/special_features/ext/workshops/correlation.html

**HTML**: http://jalt.org/test/SSA6.htm / **PDF**: http://jalt.org/test/PDF/SSA6.pdf