Shiken: JALT Testing & Evaluation SIG Newsletter. 14(1) March 2010 (p. 24 – 35)

Suggested Answers for Assessment Literacy Self-Study Quiz #9

by Tim Newfields

Possible answers for the nine questions about testing/assessment which were in the March 2010 issue of this newsletter appear below.

Part I: Open Questions

1. **Q:** What does the term *effect size* mean and how is it measured? When can an effect size be justifiably considered "large"?

A: Graziano and Raulin (2000, par. 2) define effect size as "an index of the size of the statistical difference between groups, independent of the size of the groups, and expressed in standard deviation units." In fact, effect sizes can be expressed in standardized and non-standardized units. However, it is important that the interval involved be clear to readers (Wilkinson & APA Task Force on Statistical Inference, 2004, p. 606). Effect sizes are widely employed in statistical power analyses (useful for determining the chances of a false negative) and meta-analyses (comparing the results of several research studies). Actually, they should be a feature of any experimental or quasi-experimental design.

Effect size can be calculated a number of ways. In fact, it is better to think of effect size as a family of measurements rather than as a single measure. One simple effect size measurement is to subtract the difference in mean scores between an experimental and control group, then divide that sum by the standard deviations for both groups according to this formula:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

Readers might recognize this as Cohen's *d* – one common effect size measure. Another way to calculate effect size, known as Glass's *delta*, is to divide the mean scores between an experimental and control group by the standard deviation of the control group. The Pearson *r* is also a common effect size index. A good introduction to that is provided by Ferguson (2009). Additional effect size indices have been suggested by Hedge (1981), Hedge and Olkin (1985), and Rosnow, Rosenthal, and Rubin (2000) in order to reduce measurement bias.

Regarding effect size strength, it is good to remember that a statistical *large* effect size is not necessary an *important* one. For example, it is easy to obtain large pre-test/post-test effect sizes for short criterion-referenced tests if examinees have a good idea what will appear on that exam. However, this does not mean that a significant change in linguistic ability has taken place. To interpret whether an effect size is "important" or not, it's essential to look beyond the numerical values derived from specific formulas and consider broader issues such as possible sampling bias and what the instrument might be measuring.

General guidelines for effect size interpretation have been offered (Cohen, 1988, p. 13, cited in Valentine & Cooper, 2003 p. 5). In most cases a Cohen *d* greater than .8 or Pearson *r* above .5 could be considered "large". Conversely, a Cohen *d* under .2 or *r* under .1 is small.

Since effect size estimates seek to ascertain the amount of non-overlap between the two sets of data, an more intuitive way to express effect size differences for lay readers may be Cohen's U3 index, which indicates how much the experimental and control groups differ in

terms of percentile means. A closely related term is the *improvement index*, which measures the percentile rank of the mean experimental and control scores, (Institute of Education Science & What Works Clearinghouse, 2008, p. 1).

Further Reading:

Carson, C. (n.d.) The effective use of effect size indices in institutional research. Retrieved March 14, 2010 from http://www.keene.edu/ir/effect_size.pdf

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
Cortina, J. M. & Nouri, H. (2000). *Effect size for ANOVA designs*. Thousand Oaks, CA: Sage Publications.

Effect Size. (2010, March 10). Wikipedia: The Free Encyclopedia. Retrieved March 9, 2010 from http://en.wikipedia.org/wiki/Effect_size

Ferguson, C. J. (2009). An Effect Size Primer: A Guide for Clinicians and Researchers. *Professional Psychology: Research and Practice, 40* (5) 532 - 538. DOI: 10.1037/a0015808

Graziano, A. M. & Raulin, M. L. (2000). Online Glossary to *Research Methods: A Process of Inquiry* (4th Edition). Retrieved March 11, 2010 from http://web.squ.edu.om/med-Lib/MED_CD/E_CDs/SPSS/glossary/glosse.htm

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational and Behavioral Statistics, 6* (2) 107-128. DOI: 10.3102/10769986006002107

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Levine, T. R. & Hullett, C. R. (2002). Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research. *Human Communication Research, 28* (4) 612-625. Retrieved March 14, 2010 from www.informaworld.com/index/912219870.pdf

Morris, S. B. (2008). Estimating Effect Sizes From Pretest-Posttest-Control Group Designs. *Organizational Research Methods, 11* (2) 364-386. DOI: 10.1177/1094428106291059

Rosnow, R. L., Rosenthal R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science, 11* (6) 446-453. DOIC:10.1111/1467-9280.00287

U.S. Department of Education Institute of Education Science & What Works Clearinghouse. (2008). WWC Standards (Version 1): Improvement Index. Retrieved March 9, 2010 from http://ies.ed.gov/ncee/wwc/references/iDocViewer/Doc.aspx?docId=20&tocId=4

Valentine, J. C. & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes.* Washington, DC: What Works Clearinghouse. Retrieved March 14, 2010 from http://ies.ed.gov/ncee/wwc/pdf/ essig.pdf

Wilkinson, L. & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54* (8) 594 - 604. Retrieved March 14, 2010 from http://www.loyola.edu/library/ref/articles/Wilkinson.pdf

2. **Q:** What is the difference between the *basket* and *Angoff* rating methods? What are the pros and cons of each procedure? When should they be employed?

A: Both these procedures are used in standard settings in attempts to establish empirically justified cut-off points for a test. A wide range of standard setting procedures exist and Cizek and Bunch (2006, pp. 65 - 215) provide a comprehensive overview. Two widely used item-based standard setting procedures are the basket method (sometimes known as the "in basket

method") and the Angoff procedure. Administered under standardized conditions with 5-15 raters or so, both are valuable data mining techniques offering insights as to how a group of individuals respond to a given task.

The basket method dates from at least the 1970s and was used by AT&T's Assessment Center (Byham, 1970). In 2003 a variant of it was used to help establish CEFR guidelines. One common form of the basket method could be described as a modified Angoff method in which raters make yes/no decisions as to whether a given performance fulfills a specified criterion.

In a traditional Angoff method, a panel of expert judges estimate the probability that a "minimally competent" individual with a set of defined skills would complete a given task successfully. Most often, ratings are done individually and the combined ratings are averaged and then subject to a broad range statistical procedures to ascertain inter-rater reliability.

The main advantage of the basket method is that it is easy to administer and does not require IRT scoring. However, Kaftandjieva (2009, p. 26-27) contends it has significant bias issues. In her opinion, it is prone to distortion judgments, particularly when applied to tests with narrow cut off ranges.

The main advantage of the Angoff method is its widespread credibility. However, the rating process is resource intensive and raters must think in terms of probabilities as well as conceptualize many characteristics of minimally borderline candidates simultaneously to make accurate assessments. Both tasks are often difficult.

In both the basket and Angoff methods, the importance of any given item with respect to a holistic rating standard is not considered. That raises concerns about representation issues for the test as a whole and whether all examination items should be weighted equally.

Further Reading:

Angoff, W. H. (1971, 1984). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measuremen*t (2nd ed.) (pp. 508–600). Washington, DC: American Council on Education. Retrieved on March 14, 2010 from http://www.ets.org/portal/site/ets/menuitem.c988ba0e5dd572bada20bc47c39215 09/ ?vgnextoid=78c5c2f348b46010VgnVCM10000022f95190RCRD&vgnextchannel=dcb3be3a864f4010VgnV CM10000022f95190RCRD

Byham. W. C. (1970, July/August). Assessment centers for spotting future managers. *Harvard Business Review, 59,* 150-167

Cizek, G. J. & Bunch, M. B. (Eds.) (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests* (New Edition). Thousand Oaks, CA: Sage Publications.

Cross, L. H., Impara, J. C., & Frary, R. B. (1984). A comparison of three methods for establishing the minimum standards on the national teacher examinations. *Journal of Educational Measurement, 21* (2) 113-129.

George, S. George, Haque, M. S. & Oyebode, F. (2006) Standard setting: Comparison of two methods. BMC Medical Education 6 (46). DOI: 10.1186/1472-6920-6-46

Kaftandjieva, F. (2009). Basket Procedure: The breadbasket or the basket case of standard setting methods? In N. Figueras & J. Noijons (Eds.) Linking to the CEFR levels: Research perspectives. (pp. 21-34). Arnheim: CITO/ EALTA. Retrieved March 11, 2010 from http://www.coe.int/t/dg4/linguistic/EALTA_PublicatieColloquium2009.pdf

Rock, D. A., Davies, E. L., & Werts, C. (1980). An empirical comparison of judgmental approaches to standard setting procedures (Research report #0-7). Princeton, NJ: Educational Testing Service.

3. **Q:** What is the university entrance exam item below probably attempting to measure? How could this item be improved?

I 次の 1 ～10 のうち，誤った英語表現を含んだ部分がある場合には a ～ d から誤りを 1 つ選び，誤りがない場合には e を選んでマーク解答用紙にマークせよ。

3. It (a) was customary (b) in that country, (c) the couple decided to (d) get married only after having received the permission of (e) both of their entire families. NO ERROR

A: This item was probably designed to measure the ability to *detect* sentence-level syntactic errors; that is an editing skill which appears to be distinct from the ability to *create* such sentences (Gray, 2004; Christensen, 2005). However, this particular item appears to be so muddled that it would be difficult to say it measures *anything*. Option A appears to be the most correct answer: by changing the first word to "As it" or merely "As", a problem with the syntax is resolved.

It could be argued more problematic issues exist with this test item. For example, although "the couple" might refer to one specific couple, if the sentence appears in isolation many readers will be tempted to think this refers to couples in general. The test item would have had more authenticity if several sentences had been embedded in a cohesive paragraph. Moreover, the phrase "both of their entire families" sounds archaic and patriarchal.

How should an item like this be rewritten? First, test designers should reflect on whether sentence-level grammar questions such as the one cited are sending the right message to examinees. The obsession with grammatical correctness can be seen as a detriment when we reflect on its likely educational backwash. The consequential validity of even well written sentence-level grammar correction items of this type should be questioned.

Also, it may be good to reflect on why such archaic English is being used. Most Asian EFL students still have difficulty mastering contemporary English. Should we also be teaching university applicants who are non-English majors the quaint writing conventions from earlier centuries?

Finally, from a statistical standpoint there is a question as to whether all of the distractors are functioning efficiently. As the number of distracters increases, the efficiency of each distracter tends to decrease, but – if the distractor is well-designed - so does the likelihood of guessing. For most types of multiple choice tests, having 3-4 choices per stem is considered optimum. According to Kehoe (1995, par. 21) and Bothell (2001, p. 4) the use of the "no error" option for multiple choice test items should be avoided. However, as Haladyna, Downing, and Rodriguez (2002, p. 319) point out, opinions about "no error" (or closely related "none-of-the-above") options are mixed.

Further Reading:

Bothell, T. W. (2001) 14 rules for writing multiple-choice questions. Retrieved on March 20, 2010 from http://testing.byu.edu/.../14%20Rules%20for%20Writing%20Multiple-Choice%20Questions.pdf

Christensen, C. A. (2005). The role of orthographic-motor integration in the production of creative and well-structured written text for students in secondary school. *Educational Psychology, 25* (5) 441 - 453 DOI: 10.1080/01443410500042076

Gray, R. (2004). Grammar correction in ESL/EFL writing classes may not be effective. *The Internet TESL Journal, 10* (11). Retrieved on March 15, 2010 from http://iteslj.org/Techniques/Gray-WritingCorrection.html

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15* (3), 309 - 334.

Kehoe, J. (1995). Writing multiple-choice test items. *Practical Assessment, Research & Evaluation, 4* (9). Retrieved March 20, 2010 from http://PAREonline.net/getvn.asp?v=4&n=9 . This paper has been viewed 80,293 times since 11/13/1999.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning 46* (2) 327-369. Retrieved on March 15, 2010 from http://hss.nthu.edu.tw/~fl/faculty/John/Grammar_Correction_in_L2_Writing_Class.pdf

4. **Q:** What is the *John Henry effect*? How does it differ from the *Hawthorne effect*? How can researchers minimize both of these effects?

A: Most readers are probably familiar with the Hawthorne effect, which is reputed to occur when subjects respond differently as a consequence of being studied. The John Henry effect has been described by Zdep and Irvine (1970) as a "reverse Hawthorne effect" because the control group rather than the experimental group was found to perform better in a test setting, likely because they felt themselves to be in competition with the experimental group. When both the control group and experimental group are competing for the same limited funds, for instance, it is quite likely each group will try to outperform the rival group.

Both the Hawthorne effect and John Henry are expectancy effects that can easily confound research studies. To reduce such confounding, expected results should be masked and multiple test items for each question under investigation should be used. Let's consider both of these points in more detail.

(1) *Mask expected results*

To the extent that it is ethically possible researchers can – and should – attempt to mask the results they expect to obtain. In many small scale research projects this does not appear to be done. Consider this following survey question that I designed in 2008:

Ex. 1 "I'm confident of my ability to understand most daily conversations in English." (circle <u>one</u> response)

Strongly Agree	Agree	Unsure	Disagree	Strongly Disagree
5	4	3	2	1

Setting aside the issue of whether or not there is actually a difference between the "[dis]agree" and "strongly [dis]agree" response options, if this were the <u>only</u> item seeking to measure student confidence in the survey, respondents would probably tend to give inflated responses since it is easier to agree with (or to express ambivalence about) survey items than it is to disagree with them. A well-designed survey should either use items that are neutrally worded or else counterbalance the previous item with one which has a different nuance, as in this example:

Ex. 2 "When it comes to most everyday spoken English, I'm not confident I can understand it." (circle <u>one</u> response)

Very true of me	True of me	Somewhat true of me	Not true of me	Never true of me
5	4	3	2	1

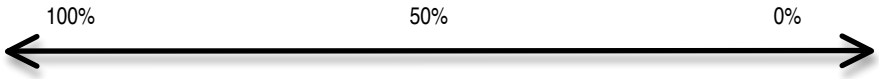
Notice that these two items differ not only in terms of nuance, but also response descriptors. Example 1 has a agreement-based descriptor scale, but Example 2 has a veracity-

based descriptor scale. When using similar questions to measure the same construct, altering the nuance and descriptors and nuances might enhance the robustness and range of the overall instrument.

(2) *Use multiple items for each question under investigation*

Nearly every survey item is problematic in some way and the overall robustness of a survey is enhanced if multiple survey items cover each research question. One of the advantages of this is that it helps alert us to random marking, a common problem if respondents feel little investment in completing a given survey. Another way that expectancy can be masked more effectively is with multiple questions that vary in nuance. Consider how Example 3 below explores the same theme as the previous two examples with a different nuance as well as a different response format:

Ex. 3 "What percentage of the time do you feel you can understand everyday spoken English?"
(Draw an "X" anywhere along the continuum below)



At this point we will ignore the ambiguity that is associated with the word "understand" – a design flaw of all three examples cited here. The visual analog scale in Example 3 has an advantage of being free of subjective descriptors such as "very" or "pretty much". However, such scales are difficult to score. Since no scale is perfect and each survey question has some kind of bias, asking several questions about each item being measured is a wise policy. Unfortunately, the majority of the surveys designed by language teachers that I have examined consist of only one item measuring each property under investigation. As a result, expectancy issues and other types of bias tend to compromise the research.

Further Reading:

Draper, S. W. (2009, December 23). The Hawthorne, Pygmalion, Placebo and other effects of expectation: Some notes. Retrieved on March 15, 2010 from http://www.psy.gla.ac.uk/~steve/hawth.html#Preface

Jones, R. A. (1981). *Self-fulfilling Prophecies: Social, Psychological, and Physiological Effects of Expectancies.* Hillsdale, NJ: Psychology Press.

Mizumoto, A., & Takeuchi, O. (2009). Comparing frequency and trueness scale descriptors in a Likert scale questionnaire on language learning strategies. *JLTA Journal, 12,* 116 - 130.

Van Bennekom, F. (2007). How Question Format Affects Survey Analysis. Retrieved on March 16, 2010 from http://www.greatbrook.com/survey_question.htm

Zdep, S. M. & Irvine, S. H. (1970). A reverse Hawthorne effect in educational evaluation. *Journal of School Psychology 8,* 85 - 95.

5. **Q:** What steps could be taken to improve the *differential validity* of a school entrance exam? How often are such steps taken at institutions you are familiar with?

A: Differential validity seeks to ascertain whether a test is fair to all examinees and measuring only what it claims to. For example, a test claiming to measure English reading skills should not advantage some test takers over others on the basis of non-construct relevant variables such as gender or ethnic background.

Let us consider a Japanese university entrance exam as an example. If over 99% of the exam applicants are ethnic Japanese, there is probably little value in seeing how Japanese and non-Japanese perform differently on the test. Likewise, if over 99% of the applicants are between ages 17 and 20, there may be little rationale for exploring how age differences impact performance. One non-construct variable that probably should be explored, however, is gender. If an EFL exam claims to measure only "language proficiency" yet males and females perform very differently on that exam, we are left with some questions that merit exploration. Within the given population, do men and women actually differ in terms of language proficiency? Or is there some sort of test bias that might disadvantage one gender? To answer those questions and ascertain the different validity of an exam, many different kinds of evidence would need to be examined.

How often are differential validity checks done on entrance exams in Japan? Although there are fortuitous exceptions, it seems that staff at most universities in Japan are rotated from to new departments every 2-6 years and as a consequence, few have any professional background in testing.

Part II: Multiple Choice Questions

1. **Q:** Which of the following procedures are best suited for comparing data from two 5-point Likert-like scales from the same sample in a pre-test/post-test research design?

(A) A t-test	(E) A Kruskal-Wallis H-test
(B) A Mann-Whitney U test	(F) Multiple linear correlations
(C) A Spearman test	(G) Pairwise multiple ANOVAs
(D) Somer's D coefficient	(H) Other: _____

A: First it is necessary to consider what sort of data can be obtained from a Likert-like scale. Since the range between most Likert scale responses is unknown, most Likert-like scale data should probably be regarded as *ordinal*. However, as Mizumoto and Takeuchi (2009, p. 119) point out, the practice of treating data from 4-choice (or more) Likert-like scales as *interval* data is prevalent. If you believe the Likert-like scale data you are using is merely ordinal at best, then options (C) – (F) are viable, as well as chi-square procedures. Of course, Rash analysis offers an elegant way to make ordinal sale data ostensibly interval scale data. If you believe that your Likert-like scale data can justifiably be considered interval data, then all options except (G) could perhaps be justified. According to Wilkinson and the APA Task Force on Statistical Inference (2004, p. 607) option (G) would "straightjacket" the research and lead to the rejection of many potentially fruitful hypotheses.

Moreover, although t-tests have been used with Likert scales (Sisson & Stocker, 1989, as cited in Clason & Dormody, 1994 p. 34), there is disagreement as to whether these are appropriate choices when dealing with Likert-like scale data.

For an in-depth discussion of the statistical methods mentioned above, refer to Rosenthal and Rosnow (2008), Ryan (2000) or Sheskin (2007).

Further Reading:

Clason, D. L. & Dormody, T. J. (1994) Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education, 35* (4) 31-35. Retrieved on March 16, 2010 from http://pubs.aged.tamu.edu/jae/pdf/ Vol35/35-04-31.pdf

Mizumoto, A., & Takeuchi, O. (2009). Comparing frequency and trueness scale descriptors in a Likert scale questionnaire on language learning strategies. *JLTA Journal, 12,* 116 - 130.

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). New York: McGraw-Hill.

Ryan, T. P. (2000). *Statistical Methods for Quality Improvement* (Wiley Series in Probability and Statistics). New York: John Wiley & Sons, Inc.

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL: Chapman & Hall/CRC.

Wilkinson, L. & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54* (8) 594 - 604. Retrieved March 14, 2010 from http://www.loyola.edu/library/ref/articles/Wilkinson.pdf

2. **Q:** In the field of statistics, which of the following terms correspond most closely with an "observed variable"? (Hint: More than one of the choices below fit.)

(A) a dependent variable
(B) an independent variable
(C) a criterion variable
(D) the predictor variable
(E) an extraneous variable
(F) outcome variable

A: There is a regrettable lack of uniformity concerning many statistical terms. In experimental and quasi-experimental research, a variable that is manipulated to ascertain how a specific outcome is influenced is variously known by these terms: causal variable, explanatory variable, exposure variable, independent variable, input variable, manipulated variable, moderated variable, and sometimes as a regressor.

A variable that changes as a consequence of shifting the independent variable is variously known by these terms: dependent variable, measured variable, observed variable, outcome variable, output variable, responding variable, or response variable.

A variable which is kept the same in an experiment is called a controlled or fixed variable.

Finally, a variable that is thought to influence the experimental outcome, but is not the focus of a given study is called as a extraneous variable, construct-irrelevant variable, lurking variable, or uncontrolled variable.

This is a simplification. Many of the terms above are used in slightly different ways in different branches of statistics, although the basic concept behind each of these four variable types is markedly similar.

Further Reading:

Marczyk, G., DeMatteo, D., & Festinger, D. (2005). *Essentials of research design and methodology*. New York: John Wiley & Sons.

Marion, R. (2004). The Whole Art of Deduction: Defining Variables and Formulating Hypotheses Retrieved March 14, 2010 from http://sahs.utmb.edu/pellinore/intro_to_research/wad/vars_hyp.htm

3. **Q:** Arranging the blocks of a test on the basis of an estimate of what should allow examinees to gain the maximum number of points in the least amount of time is an example of _____.

(A) efficiency ordering
(B) facility ordering
(C) difficulty ordering.
(D) reactive ordering.

A: There are a number of different ways to order the sections of a test. This issue is particularly relevant in *speeded tests*, in which examinees race against the clock. However, it is also relevant to long tests, since *test fatigue* can affect performance towards the end of a test.

In efficiency ordering - the correct answer to this question - the section of a test which should allow most examinees to garner the most points in the least time appears first. Those sections of a test which take more time to complete - but do not have as much payoff value for the time invested - appear later. In a typical Japanese university entrance exam that follows this scheme, multiple choice items would appear first and long reading passages would appear at the end of the exam if each test item carried the same weight. However, if the reading passage questions were worth more points than the multiple choice questions, they might appear first. The idea behind efficiency ordering is that students who are slow in completing the exam will not be penalized so heavily since they would have finished the highest dividend-yielding parts of the exam.

Placing the easiest blocks of an exam first and the most difficult parts last an example of facility ordering. An opposite strategy would be Option (C), in which the most challenging sections of a test appear first.

The most conceptually complex way to organize the blocks of a test is to consider how successful completion of one block might facilitate the completion of another block. If a test had perfect item-independence, this would be a non-issue. However, real life tests in general (and small scale tests in particular) often have some cross-item contamination. For example, having examinees complete some multiple-choice comprehension questions in a test might help explicate some vocabulary questions that appear later in that test. In reactive ordering, an attempt is made to reduce cross-item contamination. However, in cases when examinees can skip from section to section of a test or revise answers before handing in the test, the efficacy of reactive ordering may be negligible.

An alternative option to the four methods of ordering the sections of an exam listed above is *random ordering*, in which the placement of each set of test items is a matter of chance.

Further Reading:

Genesee, F. & Upshur, J. A. (1996). Classroom-Based Evaluation in Second Language Education (Cambridge Language Education). New York: Cambridge University Press.

Test Rubric: Problems Associated with Rubrics. (2002). In S. A. Mousavi. *An Encyclopedic Dictionary of Language Testing.* (3rd Ed.). (pp. 755-757). Taipei: Tung Hua Book Company.

4. **Q:** What does "truncation" generally refer to in test equating?

(A) Using just the sections of the respective tests being compared.
(B) Assigning scaled scores in a way that ignores the very highest or lowest raw scores.
(C) The shifting of data from an equivalent-groups design into a single-groups design.
(D) A form of statistical censoring that occurs when a given value is outside the range of the measuring instrument.

A: The correct answer is (B). In truncation, the extreme top and/or bottom scores of a test are removed from consideration. Truncation may also occur if the observation period is shorter than the events under investigation, such as in a mortality study.

Option (D) describes a different condition known as *censored data*, which occurs when responses go beyond the measurement range of a test. With truncation, data is cut *a posteriori* from analysis; with censored data, however, the instrument was simply not able to measure the data in the first place. As a case in point, data censoring could occur if a child who is highly proficient in a foreign language as a result of having lived overseas for many years took a test of proficiency in that language along with other children who have just begun to study it. Essentially the child belongs to a different group of examinees and consequently his or her level of proficiency would not be measured well by a test designed for a population with a relatively low proficiency level.

Further Reading:

Mandel, M. (2007). Censoring and truncation - Highlighting the differences. *The American Statistician, 61* (4) 321 - 324. DOI: 10.1198/000313007X247049.

Acknowledgement

Many thanks to Lars Molloy, Ed Schaeffer, and Chris Weaver for feedback on this article. The responsibility for any errors herein rests with the author.

HTML: http://jalt.org/test/SSQ9.htm / **PDF**: http://jalt.org/test/PDF/SSQ9.pdf