SHIKEN: JALT Testing & Evaluation SIG Newsletter. October 2011. *15*(2) 30 - 42

Assessment Literacy Self-Study Quiz #11

by Tim Newfields

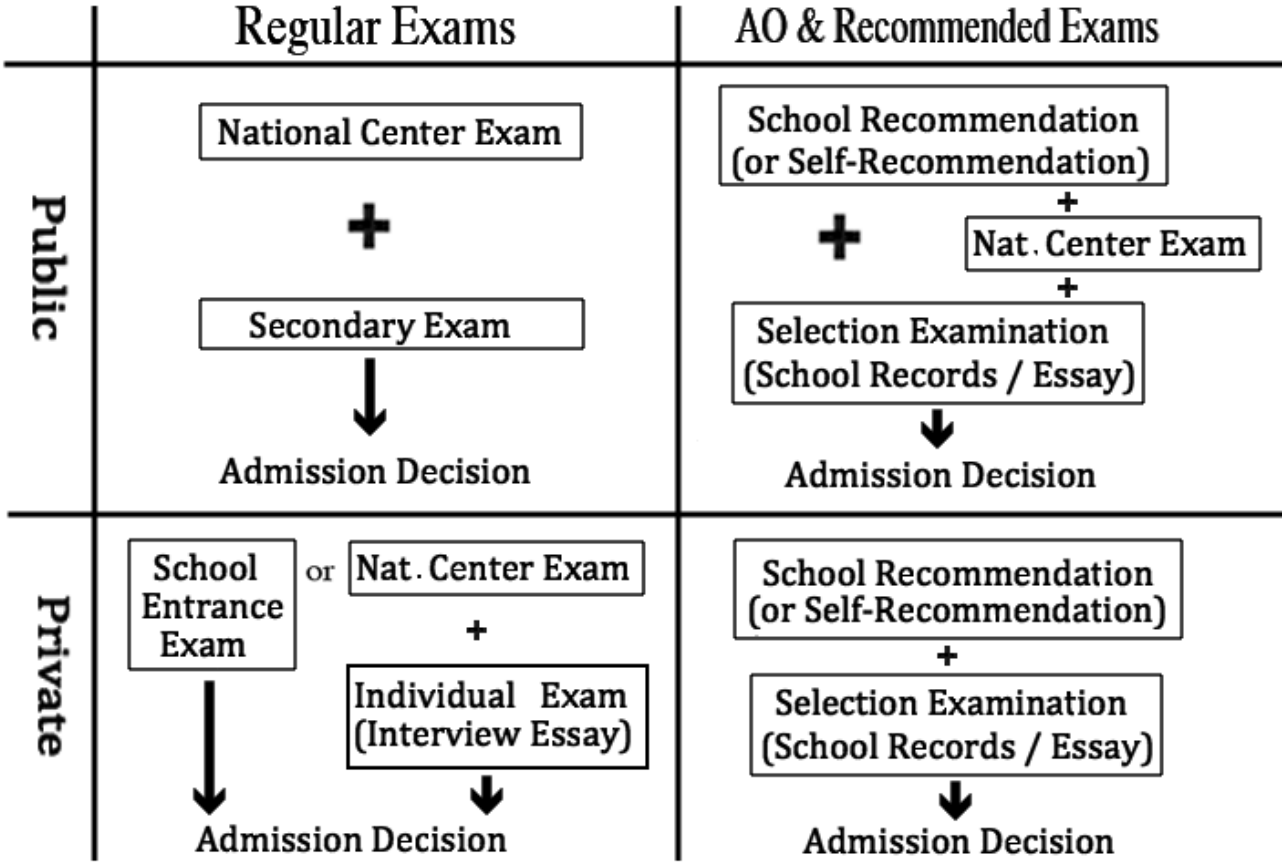
Possible answers for the ten questions about testing/assessment which were in the October 2011 issue of this newsletter appear below.

Part I: Open Questions

1. **Q:** At most universities in Japan, how do the AO entrance examinations tend to differ from regular entrance exams and the National Center Test?

A: Japanese universities have a wide variety of admission pathways, and these vary slightly from year to year. Most faculties have anywhere from four to nine different types of entrance examinations. Table 1 offers an overview of the main entrance exam pathways at tertiary educational institutions in Japan.

Table 1. *An Overview of the Main Admissions Procedures Used by Japanese Universities*



Source**:** *Daigaku juken annai 2011*, Shobunsha Inc. (2011, p. 34)

The AO exam could loosely be translated as a "self-recommendation" or "self-referral" exam. Its precise content varies from school to school. At Keio University's Faculty of Economics, for instance, this year it consists of an essay written during a 90-minute period, a 5-10 minute interview, and a portfolio of achievements through which candidates can score admission points. Items on the portfolio checklist include community volunteer work, awards in art, science, or sports, or record of holding high school student council office.

Most Japanese universities do not include essays in their AO exams. By contrast, high school grade point averages generally weigh heavily in admissions decisions. Particularly at large schools, admission decisions tend to be based on simple mathematical formulas. Usually AO examinees can bypass standard written exams, though some national universities require such examinees to take subjects on the National Center Exam deemed relevant to their major. In 2010 10.5% of all students at private universities and 2.5% of all students at public universities entered university via the AO gateway (Daigaku Shinbunsha Shinrou Jouhou Kenkyuu Sentaa, 2011, p. 63).

At most private schools, the so-called "A Exam" consists of their regular institutional exam. Although the content is somewhat similar to the National Center Test (*Sentaa Shiken*), institutional exams are generally not developed by testing experts and they are more apt to have problems which have already been pointed out by others such as Schoppa (1990), Brown (1998), and Zeng (1999). The number of exam subjects applicants must take varies from faculty to faculty. Many schools now have a system enabling applicants to choose their "best" three, two, or even one exam results. Because of shifting demographics, it is now far easier to enter Japanese universities than it was a generation ago. How do Japanese students navigate though the maze of university examinations? Many refer to reference works such as Shoubunsha's annual *Daigaku Juken Annai*.

Further Reading:

Brown, J. D. (with T. J. Leonard). (1998). Japanese university entrance examinations: An interview with Dr. J.D. Brown. *The Language Teacher, 22* (3) 303-318. Retrieved from http://www.jalt-publications.org/tlt/files/98/mar/leonard.html

Daigaku Shinbunsha Shinrou Jouhou Kenkyuu Sentaa. (2011). *2011 nen-ban shinrou adbaiza kentei koushiki tekisuto*. [2011 public examination advisory text] Tokyo: Author.

Schoppa, L. J. (1990). *Education reform in Japan: A case of immobilist politics*. London: Routledge.

Shoubunsha. (2011). *Daigaku juken annai*. [University entrance exam guide]. Tokyo: Author.

Zeng, K. (1999). *Dragon gate: Competitive examinations and their consequences*. London & New York: Cassell.

2. **Q:** What is a "false consensus effect" (FCE)? How might it skew some social science data and how should it be avoided?

A: This is a form of cognitive bias in which respondents overrate the extent that others agree with them. In other words, it is a tendency to conflate personal opinions about an issue with those of

a larger group. Fields and Schuman (1976) describe this tendency as "looking glass perceptions" in which individuals believe their views represent the views of a majority.

Recognizing this proclivity, when conducting surveys it is probably wise to avoid asking about the opinions of *others* – and focus on the opinions of respondents themselves. Survey questions such as "how common is study abroad among university students in Japan?" are vulnerable to false consensus artifacts. Rather than include questions like this on a survey, it is better to seek data from so-called "objective sources" such as government statistics, at least for research informed by standard empirical frameworks. However, if a study is informed by a post-modern narrative study framework, questions about the behavior of others might have value. However, the answers to such question should not be regarded as "factual" in any traditional, empirical sense. Instead, they should be interpreted as a constructed discourse that may or may not seem "true" to others. Most studies based on *attribution theory* (Jones, et al, 1972), for example, do ask informants to report to indicate why they think others have a specific proclivity. As long as their answers are interpreted as narratives rather than as "facts" this is not problematic.

Further Reading:

Fields, J. M. & Schuman, H. (1976). Public beliefs about the beliefs of the public. *Public Opinion Quarterly, 40* (4) 427-448. DOI: 10.1086/268330

Jones, E. E., D. E. Kannouse, H. H. Kelley, R. E. Nisbett, S. Valins, and B. Weiner, (Eds). (1972). *Attribution: Perceiving the causes of behavior*. Morristown, NJ: General Learning Press.

Wojcieszak, M. (2008). False consensus goes online: Impact of ideologically homogeneous groups on false consensus. *Public Opinion Quarterly, 72* (4) 781-791. DOI: 10.1093/poq/nfn056

3. **Q:** What unstated assumptions are inherent in the TOEIC "Can Do" chart at http://jalt.org/test/Graphics/SSQ11q3.gif? What issues need to be considered when interpreting this chart? How could this information be more ethically presented?

A: This chart appears to be based on naïve and rather misleading assumptions about how TOEIC scores correlate with English ability and employment outcomes. Even if we conceded that the TOEIC was a measure of formal English knowledge, there is certainly no guarantee that merely knowing the formal rules of English or having a substantial vocabulary will translate into actual English usage. Studies by Macintyre and Charos (1996), as well as Macintyre, Clément, Dörnyei, and Noels (1998) suggest that affective variables such as a "willingness to communicate" significantly impact target language usage. The TOEIC is not designed to measure such factors. Moreover, the job categories in this chart seem to ignore the fact that English ability is merely one of several key variables in determining job outcomes. Other variables such as social skills, job expertise, and age are not considered in this chart.

How should this information be ethically presented? The authors could begin by mentioning their sample size and the demographic characteristics upon which the information appears to be built. They also need to mention the degree of correlation amongst each of the variables. As it stands, this chart suggests a perfect correlation between all of the variables. Obviously, it seems to be based upon idealized scenarios rather than actual data because real data is seldom as clean cut as this chart suggests. This chart illustrates what can happen if advertising imperatives overrun concerns with accuracy.

Further Reading:

Macintyre, P. D., & Charos, C. (1996). Personality, attitudes, and affect as predictors of second language communication. *Journal of Language and Social Psychology, 15 (1)* 3-26. doi: 10.1177/0261927X960151001

Macintyre, P. D. Clément, R., Dörnyei, Z. & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence andaffiliation. *Modern Language Journal, 82,* 545-562.

4. **Q:** Illustrate how a study with a *regression discontinuity design* could be structured to ascertain the effects of a specific intervention on English language reading fluency among a group of EFL high school students.

A: A regression-discontinuity design is a type of quasi-experimental design in which there is a pretest-posttest comparison of two groups based upon a fixed pre-test cutoff criterion. To illustrate this, let's consider how the high school scenario might be implemented. Let us say all students entering a high school are given a test supposedly measuring English reading fluency just before classroom assignments are made. Students scoring *below* a pre-determined point on that test would be assigned to one group of classes, and receive a specific intervention. Those scoring *above* the cut off point would be placed in different classes and given "normal instruction". Towards the end of the academic year, a different version of the same test would be re-administered. If the score results in the treatment group were significantly higher than their initial regression line would predict, there is a good chance that intervention was responsible for the resultant score gains. The problem, of course, is that other variables besides the specific intervention might be responsible for the score gains. For example, if some of the teachers in the control group were considerably more skillful than those in the experimental group, the effects of the specific intervention might seem null. However, if the sample size is sufficiently large, such cases will even out. For this reason regression-discontinuity designs are regarded as a good way to ascertain causality, as true random designs are difficult if not impossible to implement in ordinary classroom settings.

Another aspect of this question concerns how EFL reading fluency should be measured. First we should recognize that the term "reading fluency" has a wide range of meanings. If we wish to

define it narrowly as "the ability to read accurately, quickly, effortlessly, and with appropriate expression" (Rasinkski, 2003), then perhaps one of the measurement procedures outlined by Wagner and Lawton (2011) would be appropriate. If we define reading fluency more broadly to include comprehension, then some of the tests described in the Testing Reading section of the *JLTA Language Testing Bibliography* (2009) might be useful. Extensive field testing would be needed to ascertain whether or not any specific test would be appropriate for a given high school context.

Further Reading:

Japan Language Testing Association. (2009). *The JLTA language testing bibliography, Category 3: Testing reading*. Retrieved from https://e-learning.ac/jlta.ac/mod/resource/view.php?id=36

Moss, S. (2009). *Regression discontinuity design*. Retrieved from http://www.psych-it.com.au/Psychlopedia/article.asp?id=256

Rasinski, T. V. (2003). *The fluent reader*. New York: Scholastic Professional Books.

Trochim, W. M. K. (2006). The regression-discontinuity design. *Research Methods Knowledge Base*. Retrieved from http://www.socialresearchmethods.net/kb/quasird.php

Wagner, R. &, Lawton, R. O. (2011). Assessment of word reading and reading fluency in English. *Encyclopedia of Language and Literacy Development* (pp. 1-8). London, ON: Canadian Language and Literacy Research Network. Retrieved from http://literacyencyclopedia.ca/pdfs/Assessment_of_Word_Reading_and_Reading_Fluency_in_English.pdf

5. **Q:** What additional information should have been mentioned in the chart at http://jalt.org/test/Graphics/SSQ11q5.gif to help readers to interpret the data more meaningfully? Also, how would you interpret this data distribution?

A: Although the authors did include one measure of central tendency (the mean), they neglected to indicate the degree of dispersion (usually expressed by the standard deviation) for their sample. To their credit, later in the article they did contrast the SD for males and females. In the chart or elsewhere in the paper, the reliability measure of the test should be indicated. Other descriptive statistics such as the standard error of measurement or kurtosis might be worth mentioning. Most importantly, it is not clear from the article whether SPOT is meant to function as a criterion-referenced test or a norm referenced test.

The SPOT test is a 60-item gap-filling listening test reputed to measure "integrated proficiency" in Japanese (Kobayashi, Ford-Niwa, & Yamamoto, 1996, par. 2). The fact that a sizable number of the respondents obtained very low scores should lead us to question the incentives they had for completing this test properly. If adequate incentives are not provided to complete a test, some respondents might leave portions of it uncompleted. Also, the distribution curve raises the question of whether the SPOT test was an appropriate instrument for this particular population. The discrimination index of this test for this population is worth analyzing.

Further Reading:

Kobayashi, N., Ford-Niwa, J., & Yamamoto, H. (1996). Nihongo nouryoku no atarashii sokutei-hoo: SPOT. [A new way of measuring integrative ability of Japanese: SPOT]. *Japanese Language Education Around the Globe, 4*, 201-218. Retrieved from http://www.jpf.go.jp/e/japanese/survey/globe/06/report.html#13

6. **Q:** In factor analysis, how does a *communality* differ from a *commonality*?

A: The term "communality" – not a misspelling - has a very specific meaning in factor analysis. It is "the sum of the squared factor loadings for all factors for a given variable (row) is the variance in that variable accounted for by all the factors" (UNESCO, 2005, par. 10). A communality of zero indicates that the factors under study have no direct relation to the variance observed. By contrast, a communality of 1 indicates all of the variance can jointly be explained by the factors. One way to think of communality is as a way of assessing the reliability of an indicator.

The term "commonality" has no specialized meaning in testing – it may refer to the shared characteristics of two or more variables. It could also be a loose way of explaining some types of correlation, a topic covered in depth by authors such as Brown (1988) and Spatz (2011).

Further Reading:

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. Cambridge University Press.

Spatz, C. (2011). *Basic statistics: Tales of distributions* (10th edition). Belmont, CA: Wadsworth Cengage.

StatSoft. (2011). *Principal components and factor analysis*. Retrieved from http://www.statsoft.com/textbook/principal-components-factor-analysis/

UNESCO. (2005). *Factor analysis*. Retrieved from http://www.unesco.org/webworld/idams/advguide/Chapt6_3.htm

Suggested answers online at http://jalt.org/test/SSA11.htm

Part II: Multiple Choice Questions

1. **Q:** A tendency of respondents to concur with the opinions implicitly expressed in a given rating scale (and avoid offering counter-opinions) is a known as _____.

(a) contamination bias (b) acquiescence bias (c) social desirability bias

(d) central tendency bias (e) nonresponse bias

NOTE*: One of these choices is most relevant; two are somewhat relevant, and two are irrelevant.*

A: Perhaps the most relevant answer is (b). *Acquiescence bias* (also known as agreement bias) occurs when respondents falsely concur with survey questions. It can be understood as a form of researcher expectancy bias. This should not be confused with *compliance bias*, which occurs when respondents vary in the extent that they comply with a given treatment regimen, skewing the results.

Of course, acquiescence bias can also be interpreted as a form of *social desirability bias* – a tendency to answer questions in a way that will be viewed favorably. Hence Option (c) is partly correct.

The relation of Option (d) with the described scenario is indirect. Rather than express clear disagreement with an issue, respondents might choose a neutral middle ground. *Central tendency bias* (also known as end-aversion bias) occurs when respondents hesitate to select responses at either end of a response scale, safely opting for mid-scale responses. Japanese informants, often raised in a culture with strong pressures towards conformity, might be especially susceptible to this type of bias. Option (d) can therefore be understood as partly correct.

Option (e) describes a type of sampling error some surveys are subject to. Persons with strongly negative feelings about an issue might choose to opt out of the survey altogether. The remaining respondents might be less negative about an issue than the population as a whole. *Nonresponse bias* is the main reason why it is so important to indicate what percentage of the initial pool of informants in a research project were discounted or opted out of the study. Option (e) is not directly related to the scenario described.

The scenario described in this stem is also unrelated to Option (a) c*ontamination bias*, a condition that occurs when control group members are inadvertently exposed to an intervention. For example, if an experimental group of students were taught the value of shadow echoing (Murphey, 2001) and some of them interacted with control group members outside of class, some of the control group might also start adopting that behavior. Whenever control and treatment groups interact with each other, contamination bias may occur.

Further Reading:

Moss, S. (2008). *Acquiescence bias*. Retrieved from http://www.psych-it.com.au/Psychlopedia/article.asp?id=154

Murphey, T. (2001). Exploring conversational shadowing. *Language Teacher Research 5* (2) 128-155. doi: 10.1177/136216880100500203

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15* (3) 263–280. DOI: 10.1002/ejsp.2420150303

Viswanathan, M. (2005). *Measurement error and research design*. Thousand Oaks, CA: Sage.

Weiksner, G. M. (2008). *Measurement error as a threat to causal inference: Acquiescence bias and deliberative polling*. Retrieved from http://polmeth.wustl.edu/media/Paper/Weiksner-Measurement%20Error%20as%20a%20Threat%20to%20Causal%20Inference.pdf

2. **Q:** What is one difference between the *standard error of measurement* (SEM) and *standard error of difference* (SED)?

(a) The former is a test of significance between two individual test scores.

(b) The latter indicates how widely an individual's given test score is likely to differ from his/her true score.

(c) The latter is the critical value of a ***t-test*** of the difference between two means.

(d) To calculate the latter, the former must be known.

NOTE*: Just one answer is considered fully correct.*

A: In this *SHIKEN* issue, JD Brown gave us some background information about the standard error of measurement. A less common measurement he alluded to is the *standard error of difference*, which is also known as the *standard error of the sample mean difference* as well as the *standard error of the difference between two means*. Mathematically, this can be expressed as:

$$SE_{diff} = \sqrt{SEM^2}_{\text{score1}} + SEM^2_{\text{score2}}$$

As you can see, the standard error of measurement is needed to calculate this measure, so Option (d) is correct.

The significance between two individual test scores can be ascertained by the *p-value*. Option (b) describes the standard error of measurement - not the standard error of difference. Option (c) describes a concept known as the *least significant difference* (a.k.a. Fischer's LSD). The least significant difference is most often used when one is attempting to compare two groups after an ANOVA hypothesis of equal means has been rejected using an F-test.

Further Reading:

Fong, D. (2011). *The least significant difference (LSD)*. Retrieved from https://onlinecourses.science.psu.edu/stat502/node/37

GraphPad Software. (2011). *Fisher's Least Significant Difference (LSD) test*. Retrieved from http://www.graphpad.com/faq/viewfaq.cfm?faq=176

Lane, D. (2011). Differences between two means (independent groups). Retrieved from http://onlinestatbook.com/chapter10/difference_means.html

3. **Q:** What sort of test would be most appropriate to ascertain whether a large sample of male and female university students differed significantly in their attitudes towards study abroad according to a Likert scale with 4 or 5 options?

(a) An unpaired t-test (b) An ANCOVA (c) An ANOVA

(d) A MANOVA (e) A chi-square test

NOTE*: Two answers are viable, but one answer will yield more information than the other.*

A: Although a chi-square test could be used to ascertain this, an analysis of covariance (ANCOVA) would yield richer information because it includes information about regression. An ANCOVA allows us to compare two or more categorical variables (such as gender) with one

ordinal variable (such as Likert scale scores) to ascertain – and possibly correct for – other hidden variables, known as covariates (such as foreign language ability). Of all the descriptions of this test that I have read, perhaps the most lucid for general readers is McDonald's (2009, 232-237).
Other options not mentioned among these choices include a paired t-test and linear regression.

An unpaired t-test could be used to compare two independent sets of data from two different populations. It is often contrasted with a paired (or repeated measures) t-test, a topic discussed in depth by Hinton (2004) and Stephens (2008).

The ANOVA and MANOVA would be appropriate for determining whether significant differences exist between multiple sets of data. Harlow (2010) provides a clear description of how these two types of tests differ.

Further Reading:

Harlow, L. L. (2010). *The essence of multivariate thinking: Basic themes and methods* (Multivariate applications series). Mahwah, NJ: Lawrence Erlbaum.

Hinton, P. R. (2004). *Statistics explained: A guide for social science students* (2nd Edition). New York: Routledge.

Hopkins, W. G. (1997). *A new view of statistics*. Retrieved from http://www.sportsci.org/resource/stats/quiz.html

McDonald, J. H. (2009). *Handbook of biological statistics*. Baltimore, MD: Sparky House Publishing. Also available at http://udel.edu/~mcdonald/statancova.html

Stephens, L. (2008). *Schaum's outline of statistics in psychology*. New York: McGraw-Hill.

Wendorf, C. A. (2004). *Analysis of covariance (ANCOVA)*. Retrieved from http://www4.uwsp.edu/psych/cw/statmanual/ancovaoverview.html

4. **Q:** Which of the following points is <u>not</u> problematic about the avowed ranking of these universities in the chart at http://www.4icu.org/jp/?

(A) The criteria for the ranking is not explained in the title of the chart or anywhere on the same page as the chart.

(B) This ranking was made entirely on the basis of a secret web metric formula – presumably, universities with lots of web traffic according to some search engines will receive higher ratings than those with low web traffic.

(C) The fact that the different faculties at the same university may have different rankings is ignored.

(D) The point that the lesson quality might vary widely from teacher to teacher at the same school is also ignored.

(E) The location of some of the schools appears to be incorrect.

(F) There is no mention of the measurement error of this rubric or indication of how widely apart the various school rankings are. Moreover, the possibility that two schools may have essentially equivalent rankings according to a given criterion does not appear to be considered.

A: Simply put, this chart is a travesty in every way. It is tempting to describe this as "info-clutter" designed to generate revenue from placed advertisements rather than informing readers of the key

variables that should be considered when ranking universities. To ascertain how the actual ranking was made, readers must go at least two clicks away from the original web page. Since much Internet advertising is paid on a "per click" basis, perhaps that is the reason why the information is so decontextualized.

Item (E) is perhaps the least problematic aspect of this chart. Waseda University, for instance, has a campus in Kita Kyuushuu (southern Japan) as well as six other campuses in Tokyo. This fact, however, would not impact its ranking according to the scanty information about how this ranking is done at http://www.4icu.org/menu/about.htm.

This so-called university ranking illustrates why critical thinking skills are needed to sift through the uneven quality of the information on the Internet. Unfortunately, some people without such skills might actually consider this a valid ranking of how good a university is.

Further Reading:

Doherty, J. D. (1999). Teaching information skills in the information age: The need for critical thinking. *Library Philosophy and Practice, 1* (2). Retrieved from http://unllib.unl.edu/LPP/doherty.htm