Shiken: JALT Testing & Evaluation SIG Newsletter 12 (2) April 2008. (p. 46-54)

Suggested Answers for Assessment Literacy Self-Study Quiz #5

by Tim Newfields

Here are some possible answers to the questions about testing, statistics, and assessment raised in the April 2008 issue of SHIKEN. Please note that other possible interpretations of many of the questions is possible. If you feel an answer is unclear or conclusion is incorrect, please contact the editor.

Part I: Open Questions

1. Q: A person with a hearing disability is asked to take an EFL placement test . . . . what's the most ethical way to rate this individual?

A: First, we need to be sure of the test purpose and context. If the sole purpose is classroom streaming and the teachers administering the test already are familiar with level of each class, then an informal placement interview might work well in lieu of the listening portion of this test. Basing the placement choice entirely on the reading test scores would not a wise option because EFL reading skills do not necessarily correlate highly with listening skills (Basabas-Ikeguchi,1988).

If we are talking about a high-stakes test for which comprehensive scores are needed, several options exist. If an examinee's hearing ability is only partly impaired, a assistive listening device (ALD) could be employed. If the hearing loss is total, the best option might be to use some type of captioning system. The STEP-Eiken provides a captioning services for those unable to hear, but unfortunately ETS doesn't (ETS, 2007). Since an estimated .48% of the population is partly deaf and a further .18% is completely so (Holt, Hotto, & Cole, 1994), it is worth considering how to handle test accommodations for those with auditory impairments in advance.

Further Reading:

Basabas-Ikeguchi, C. (1988). Analysis of reading and listening comprehension skills in different language environments. Unpublished Master's Thesis, Dokkyo University. ERIC #: ED355807.

Burns, E. (1998). *Test accommodations for students with disabilities*. Springfield, IL: Charles C. Thomas.

ETS. (2007) *2007-2008 Bulletin supplement for test takers with disability*. Retrieved April 12, 2008 from www.ets.org/disability/

Holt, J., Hotto, S., Cole, K. (1994). Demographic aspects of hearing impairment: Questions and answers. (Third Edition). Retrieved April 12, 2008 from http://gri.gallaudet.edu/Demographics/factsheet.html#Q1/

2. Q: How should an oral proficiency interviewee with a possible a stuttering disorder be rated?

A: Arguably, this would be a valid case for breaking the policy of using only the target language during the interview since it is important to ascertain whether the stuttering is a pervasive speech impediment or simply a manifestation of nervousness due to undertaking a foreign language exam. If it is a persistent, global phenomena then there is little doubt that a handicap is present for which accommodations are due. If the person speaks their native language fluently, then the speech impairment may be a result of social anxiety rather than a defined impairment. In that case, no special accommodations would be justified and the examinee's fluency rating would subsequently drop.

The issue, however, is actually more complex since some forms of stuttering are episodic and oral interviewers are not qualified to provide clinical diagnoses. For such reasons the main criteria for identifying a handicap among adults should be self-diagnosis. If a person indicates that they have a stuttering disorder, then they are legally entitled to "reasonable accommodation" or "adaptive measures" from the agency in question (ELSA, 2000). When completing a test application, space should be provided for examinees to indicate whether they have any disabilities requiring special accommodation.

What specific accommodations should be made when rating the oral fluency of those with stuttering disorders? Here the issue becomes complex because disorders vary widely. One option would be to listen to the output as if no disorder existed – to essentially ignore the features of the output that could be ascribed to stuttering and try to rate the remaining speech features. This is not an easy process and it seems that oral proficiency raters vary widely in their responses to stammering.

Further Reading:

ELSA. (2000). ELSA Links – Discrimination. Retrieved April 13, 2008 from http://www.stuttering.ws/links/discrim_eu.htm

Tyrer, A. (2007, September 23). Oral assessments, and assessed presentations. Retrieved April 13, 2008 from http://www.stammeringlaw.org.uk/education/oral.htm

3. Q: One EFL instructor of a basic "English communication" class awards credit if his students indicate that they've recently donated blood. Any content validity issues here?

A: Teachers often use grades as levers to induce desired behaviors. If the grading process is ethical and in line with the curricular goals and the criteria for performance are communicated clearly to all stakeholders, there is no problem.

In the scenario presented in this question, however, several problems arise. First, the curricular goals are not expressed clearly – the syllabus is far too vague about expected outcomes. Moreover, the relevance of donating blood to those curricular goals is not established. How does donating blood pertain to English proficiency? Finally, this grading system presumes all students are healthy and able to donate blood. That might not be the case. The teacher is penalizing those whose health condition (or religious belief) does not enable them to make blood donations. This case illustrates how teachers need to be very cautious about offering incentive points to induce students to undertake specific behaviors: it is all too easy to dish out points for actions not directly relevant to the curricular goals.

Further Reading:

Anderson, L. W. (2002, November) Curricular alignment: A re-examination. *Theory Into Practice, 41* (4) 255 - 260. ERIC Document #: EJ667162.

Barrie, S., Brew, A., McCulloch, M. (1999). Qualitatively different conceptions of criteria used to assess student learning. Paper presented at the 1999 Australian Association for Research in Education. Retrieved April 14, 2008 from http://www.aare.edu.au/99pap/bre99209.htm

4. *Q:* What further information should be provided to end users of *ExpertRating's* English Speaking Test (online at *www.expertrating.com/english-speaking-test.asp)*?

A: Let's start by considering the construct that's reputedly being measured. The test claims to measure "correct pronunciation in [American] English". However, this claim implies that there is only one "correct" American English pronunciation. According to the University of Arizona Language Samples Project (2001) and Kun (2007) that is simply not the case. There are many regional and ethnic varieties of American English and no single dialect can be regarded as "correct". Hence this exam seems to have a serious design flaw at the basic construct level of this test.

Secondly, this test does not specify is how pronunciation ability is measured. Are there trained human raters or is the rating entirely based on a computer speech recognition system? If human raters are used, how many raters are employed and what are their evaluation criteria? The rating criteria for this exam is far too opaque.

Another major lacuna is this test completely neglects to mention what validation criteria, if any, it employs. No descriptive statistics about its reliability or validity are provided end users. Examinees have a right to know how well the scores on the given exam correlate with other widely used measures of English proficiency.

In short, this examination has a long way to go before it can be considered a valid, professional, or ethical measure of the ability to speak English. Commercial test developers need to be careful that they devote at least as much energy to test validation as they do to marketing.

Further Reading:

Garcia, P. A. (1987). *The competency testing mine field: Validation, legal and ethical issues with implications for minorities*. ERIC Document # ED336967

Kun, T. (2007). *American regional accent map.* Retrieved April 15, 2008 from http://freeshells.ch/~xavier/accentmap/

Saar, H. (2005, January 17). Validation guidelines for test developers. Retrieved April 15, 2008 from http://www.qalspell.ttu.ee/Validation%20Guidelines%20for%20Test%20Developers.doc

University of Arizona Language Samples Project. (2001). *Varieties of English.* Retrieved April 15, 2008 from http://www.ic.arizona.edu/~lsp/main.html

Part II: Multiple Choice Questions

1. Q: Which of the following is not a feature traditional conversation analysis?

(a) using authentic, recorded data which is fully transcribed
(b) analyzing single cases or deviant cases
(c) using turns as units of analysis
(d) codifying and quantifying the data

A: Data quantification is not a feature of traditional conversation analysis. The focus of conversation analyses is generally on the descriptive features of specific interactions rather than their frequency. Weider and Lawrence (1993) argue against any attempt to quantify conversations because of the idiolectic nature of human communication and the small sample sizes generally involved in CA studies. Despite this, CA studies make frequent use of pseudo-quantifying terms such as 'regularly', 'often', 'commonly', 'rarely' etc. (Ten Have, 2000) . A few researchers such as West (1984) go further and actually quantify their data to the extent of mentioning percentiles when describing male/female discourse patterns. The question of whether (and how) to quantify conversational data is an ongoing controversy in the field. Citing works by Stivers (2001, 2002) *TESOL Quarterly* advises writers wishing to use quantification that "ensure that it only follows careful analysis of the individual cases that are being quantified, with categories for quantification emerging from this analysis of individual cases" than any a priori decision.

Further Reading:

Stivers, T. (2001). Negotiating who presents the problem: Next speaker selection in pediatric encounters. *Journal of Communication, 51,* 252-282.

Stivers, T. (2002). Presenting the problem in pediatric encounters: "Symptoms only" versus "candidate diagnosis" presentations. *Health Communication*, 14, 299-338.

Ten Have, P. (2000, July 3). *Methodological issues in conversation analysis*. Retrieved April 16, 2008 from http://www2.fmg.uva.nl/emca/mica.htm

TESOL Quarterly. (n.d.). *Qualitative research: Conversation analysis guidelines*. Retrieved from April 16, 2008 from http://www.tesol.org/s_tesol/sec_document.asp?CID=476&DID=2154

West, C. (1984) *Routine complications: Trouble with talk between doctors and patients*. Bloomington: Indiana University Press.

Wieder, D. L. (1993). On the Compound Questions Raised by Attempts to Quantify Conversation Analysis' Phenomena, Part 2: The Issue of Incommensurability. *Research on Language and Social Interaction, 26* (2) 213-26. ERIC #: EJ464150.

2. If a person takes a multiple reading choice test and selects an answer simply because the other choices do not seem correct, it is a _____ strategy.

(a) testwiseness (b) test-management (c) language learner

According to Cohen (2007, p. 93) the case above would be an example of a test-management strategy. Other examples of such strategies include using a clock during an exam, re-reading a text passage, or guessing answers on the basis of background knowledge. Test-management strategies represent attempts to maximize personal resources to score well on a test. Cohen and Upton (2006) specify 28 different test-management strategies among TOEFL examinees.

Testwiseness is said to occur when examinees rely on secondary cues from test passages to perform above their actual abilities (Millman, Bishop, & Ebel 1965, cited by Edwards, 2003). First proposed by Thorndike in 1951, sample test-wiseness strategies include avoiding answers with words such as "all" or "none" or selecting test items which have more detail without knowing whether the answer is actually correct. Ideally, well-designed tests should not be susceptible to test-wiseness strategies. In actually, most examinations do contain at least some faulty items unduly favoring test-wise examinees (Rogers & Bateson, 1991; Mahamed, Gregory, Austin, Dan, 2006).

Language learner strategies are not related to testing per se; they represent broader attempts to gain linguistic and sociolinguistic competence in a target language (Tarone 1983, cited by Lessard-Clouston, 1997) . Sample language learning strategies might include asking questions when information isn't understood or modifying L2 output to better accommodate accepted social-cultural norms.

It might be worth mentioning that bifurcation between test-management strategies and testwiseness strategies is not entirely satisfactory and not all authors favor this distinction (Gu, 1996, cited by Bremner, 1997).

Further Reading:

Bremner, S. (1997, Autumn). Language learning strategies and language proficiency: Causes or outcomes? *Perspectives, 9*. Retrieved from April 18, 2008 from http://sunzi1.lib.hku.hk/hkjo/view/10/1000125.pdf -

Cohen, A. D. & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks* (TOEFL Monograph No. MS-33). Princeton, NJ: ETS. Retrieved from April 17, 2008 from http://www.ets.org/Media/Research/pdf/RR-06-06.pdf

Cohen, A. D. (2007) The coming of age for research on test-taking strategies. In J. Fox, et al (Eds.) *Language testing reconsidered*. Ottawa, Ontario: University of Ottawa Press., pp. 89 – 112.

Edwards, B. (2003, August). An examination of factors contributing to a reduction in race–based subgroup differences on a constructed response paper–and–pencil test of achievement. Unpublished Ph.D. thesis at Texas A&M University. Retrieved from April 17, 2008 from http://txspace.tamu.edu/bitstream/handle/1969.1/128/etd-tamu-2003B-2003062513-Edwa-1.pdf?sequence=1

Gu, P.Y. (1996). Robin Hood in SLA: What has the learning strategy researcher taught us? *Asian Journal of English Language Teaching*, 6, 1-29.

Lessard-Clouston, M. (1997, December) Language Learning Strategies: An Overview for L2 Teachers. *The Internet TESL Journal, 3* (12). Retrieved from April 17, 2008 from http://iteslj.org/Articles/Lessard-Clouston-Strategy.html

Mahamed, A., Gregory, P., Austin, Z., & Dan, L. (2006, December). Testwiseness among international pharmacy graduates and Canadian senior pharmacy students. *American Journal of Pharmaceutical Education, 70* (6), p. 131. Retrieved from April 17, 2008 from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1803693

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test wiseness. *Educational and Psychological Measurement, 25*, 707–726.

Rogers, W. T.; Bateson, D. J. (1991, April). The influence of test-wiseness on performance of high school seniors on school leaving examinations. *Applied Measurement in Education, 4*, 159 – 183.

Tarone, E. (1983). Some thoughts on the notion of 'communication strategy'. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 61-74). London: Longman.

3. Which of the following statements are <u>true</u> about *p*-values?

(a) They indicate the *likelihood* a correlation between two or more variables.
(b) They indicate the *direction* of a correlation between two or more variables.
(c) They indicate the *strength* of a correlation between two or more variables.
(d) Actually, none of these.

The best answer is probably (d) because no single statistic by itself can provide us with enough information to meaningfully interpret an entire set of data. P-values statistics, if used at all, should only be used along to be a wide range of other statistics to discern the likelihood of a result being due to random noise or some significantly different.

According to Dixon (2000), p-value results are often misused and in fact they might not be the best tool for describing whether research results arise from random chance. He argues that likelihood ratios (often expressed with the Greek letter λ - lower case lamba), expressed in the formula below, offer a better way to gauge significant research results.

$$\lambda = \left[\frac{(1 - R_1^2)}{(1 - R_2^2)} \right]^{n/2}$$

Dixon's model does not appear to be widely used today, but another alternative to classic p-values proposed by Killeen (2005) which is approximated in the formula below is becoming more widely accepted:

$$p_{rep} \approx \left[1 + \left(\frac{p}{1 - p} \right)^{2/3} \right]^{-1}$$

According to Killeen (2005), r-rep values avoid the parametric inference inherent in traditional p-values and provide a viable way to detect random noise. The procedure for calculating this in SPSS is described by Wright (2008).

P-values, which are indeed flawed do not indicate the direction or the strength of a correlation, nor give us any clues about the causality. Under best conditions, they might offer some clue about the likelihood of some result being due to random chance if a test is well-designed and the sampling is also done well. However, the best of conditions is seldom met and most tests we encounter have some types of design flaws. For such reasons, considerable caution needs to be used when interpreting p-values. JD Brown (2008, p. 36-41) offered two examples of how p-value results could be misleading in this issue of SHIKEN.

Further Reading:

Brown, J.D. (2008, April). Statistics Corner. Questions and answers about language testing statistics: Effect size and eta squared. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12* (2) 36 – 41. Retrieved from April 18, 2008 from http://jalt.org/test/bro_28.htm

Dixon, P. (2003, September). The p-value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology, 57*, 189-202. Retrieved from April 18, 2008 from http://www.psych.ualberta.ca/~pdixon/Home/Preprints/pValue.pdf

Dixon, P. (2000, July). The p-value fallacy: Why inferential statistics don't describe results. Paper presented at the joint meeting of the Experimental Psychology Society of Great Britain and the Canadian Society for Brain, Behaviour, and Cognitive Science, Cambridge, UK. Retrieved from April 18, 2008 from http://www.psych.ualberta.ca/~pdixon/Home/Presentations/pValues/pValues.htm

Killeen, P. R. (2005, May). An alternative to null-hypothesis significance tests. *Psychological Science, 16* (5) 345–353. Retrieved from April 18, 2008 from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1473027

Wright, D. (2008, March 3). *Killeen's prep.* Retrieved from April 18, 2008 from http://www.sussex.ac.uk/Users/danw/masters/statistical%20analysis/killeen.htm

4. Q: Which of the following usually does not lead to score inflation?

(a) Coaching effect from teachers who know what a given test will probably cover.
(b) Exempting low-performing students being from taking the test.
(c) Random marking errors by those marking the test.
(d) Narrowing the test focus: having the test focus on just a few aspects of the target curriculum.

A: Since random marking errors (c) work both ways it would not lead to score inflation with a large sample. The likelihood of someone benefiting from a random marking error is as great as the possibility that they might be hurt by the error. All of the other factors mentioned can lead to test score inflation. So can poorly designed test questions which are vulnerable to testwiseness.

5. Q: Which of the following statements is true about power in a statistical sense?

(a) It ranges from -1 to 1.
(b) It should be used *post-hoc* and tailored to the data configuration.
(c) It depends in part on effect size.
(d) It reveals the likelihood of a Type I error.

A: According to Trochim (2006) and Jacobs (2006) effect size does have an impact on statistical power.

In many parts of the testing literature, effect size does not a single measure, but rather a host of indices to gauge the strength of the relationship between two variables. Common measures of effect size include Pearson's *R*, Cohen's *d*, Cramer's *V*, and Hedge's *g*. Some researchers such as Mousavi (p. 413) , however, define effect size more narrowly as the mean score for a experimental group minus the mean score for a control group divided by the standard deviation for the control group. That could be likened to a Z-score from ranging from 0 to 1. The more statistically powerful a test is, the less prone it is to a Type II error – falsely rejecting a null hypothesis.

Further Reading:

Becker, L. (2000, March 21). Effect size. Retrieved on April 19, 2008 from http://web.uccs.edu/lbecker/Psy590/es.htm

Jacobs, R. (2006, December 19). The concepts of statistical power and effect size. Retrieved from April 19, 2008 from http://www83.homepage.villanova.edu/ richard.jacobs/ EDU%208603/lessons/stastical%20power.html

Trochim, W. M.K. (2006). Research Methods Knowledge Base: Statistical Power. Retrieved on April 19, 2008 from http://www.socialresearchmethods.net/kb/power.php

Assessment Literacy Self-Study Quiz #5

HTML: http://jalt.org/test/SSQ5.htm / **PDF:** http:// jalt.org/test/PDF/SSQ5.pdf

Suggested Answers for Quiz #5

HTML: http://jalt.org/test/SSA5.htm / **PDF:** http:// jalt.org/test/PDF/SSA5.pdf