Shiken: JALT Testing & Evaluation SIG Newsletter. 13(2). May 2009 (p. 28 – 34)

Assessment Literacy Self-Study Quiz #7

by Tim Newfields

Here are some suggested answers for the questions about testing, statistics, and assessment from the May 2009 issue of SHIKEN. If any answer seem unclear or you have further questions, contact the author at newfields55 at yahoo dot com.

Part I: Open Questions

1. Q: (1) What skills is this task from South Korea's 2008 College Scholastic Ability Test likely tapping into? (2) In what ways - if any - might this item be inappropriate for a general university undergraduate EFL entrance exam?

A: To know with reasonable certainty what skills a test item is tapping into would require a validation study. Though some of its information is a dated, Kunnan's *Validation in Language Assessment* provides some useful guidelines on how to validate test items, particularly in terms of all-important construct validity.

Even without a time-consuming validation study, however, we can speculate about what skills this test item might utilize. In addition to English reading and paragraph-level syntactical skills, it's quite possible that this test taps into psychological knowledge. Those interested in psychological matters are likely to do better on this item than those who aren't.

Is this item appropriate for 18-year-old EFL students trying to enter university? That's hard to say conclusively, but I suspect it might be better suited to graduate level psychology students. This subject matter might be beyond the reach of most high school students, even in their native languages. A topic closer to their realm of experience would probably be better.

Another issue to explore is whether item has any gender bias. It is quite likely that women will significantly outscore men on this item, but solid data is needed to say for sure.

Further reading: Kunnan, A. J. (Ed.) (1998). *Validation in Language Assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Westen, D. & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology, 84* (3) 608-618. Retrieved April 10, 2009 from http://www.psychsystems.net/lab/Quant_Const.pdf

2. Q: Explain the difference between local test item independence and test item unidimensionality.

A: Both these terms are widely used in item response theory (IRT) and Rasch research. Both can also be said to represent ideals that are seldom, if ever, fully actualized.

Test item independence is realized when the only thing that two test items have in common is the given factor that is being examined. Technically, Baghaei (2008) explains it this way, "if you partial out the test common factor from any two items, their residual covariance is zero." That means that performance on one test item is unrelated to performance on the other item, except for the trait that is under scrutiny.

Unidimensionality is the flip side of this coin: it implies that only one latent trait is involved in each given analysis. If more than one trait is somehow involved in an analysis (which generally happens to some degree in real life), the principle of unidimensionality is

compromised. Many IRT and Rasch measures are robust enough to tolerate small violations of this principle, but if significant violations occur, then more confounding errors will arise.

Further reading:

Baghaei, P. (2008). Local dependency and Rasch measures. *Rasch Measurement Transactions, 21* (3) 1105-6. Retrieved April 11, 2009 from http://www.rasch.org/rmt/rmt213b.htm

Beguin, A. A. (2000). Robustness of equating high-stakes tests. Retrieved April 11, 2009 from http://www.cito.nl/share/poc/dissertaties/dissertationbeguin2000.pdf

Brannick, M. T. (n.d.) Item Response Theory. Retrieved April 11, 2009 from http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm

3. Q: (1) How do the two box-percentile plots online at http://jalt.org/test/Graphics/SSQ7b.gif differ?

A: A quick glance suggests both groups are reasonably mesokurtic (peaked around the mean) and more or less evenly skewed. Though Group 1 seems to have a slightly stronger positive skew than Group 2, it's not prominent. The mean values for both groups are nearly uniform, as are their upper and lower quartile ranges as well as their max./min. ranges.

The fact the mean for both groups is below zero merits further investigation. Without knowing more about the data or doing any number crunching, it's hard to interpret this.

The trait measured by this scale doesn't vary widely between the groups. A quick glance reveals the similarities of both groups far outweigh their differences. As a consequence, this chart illustrates the sort of pattern researchers aspire for when exploring the impact of a non-construct relevant factor on a research design, such as how well males and females performed on a given foreign language test.

Q: (2) Also, what advantages do these projections have over common boxplots?

A: Ordinary boxplots (also known as "box-and-whisker diagrams") generally reveal the lower quartile, median, and upper quartile range of a sample. The sample minimum and maximum observations are also often included, and sometimes the mean is as well. Although this information is helpful, it's difficult to assess the full-range frequency distributions from boxplots. Some anomalies can easily be hidden in boxplots. Box-percentile plots communicate more information in a way that is also easier for most persons to understand.

Q: (3) Finally, how can box-percentile plots such as these be calculated?

A: If you are using R (a free statistical analysis package developed by two researchers at the University of Auckland) you can download a codex to do box-percentile plots from http://lib.stat.cmu.edu/R/CRAN/. Another option is to utilize the web version of the software at http://www.math.montana.edu/Rweb/.

Other programs such as Aabel, fBasics, KaleidaGraph, Mathematica, MatLab, Stata, and S-PLUS are also reputed to handle box-percentile plots, but I have not tired them. Two good sources of information about statistical programs are Wikipedia's *List of statistical packages* (http://en.wikipedia.org/wiki/Statistical_software) and the list by Pezzullo of free products (http://www.statpages.org/javasta2.html)

Further reading: Brown, J.D. Skewness and kurtosis. *Shiken: JALT Testing & Evaluation SIG Newsletter, 1* (1) 20 - 23. Retrieved April 13, 2009 from http://jalt.org/test/PDF/Brown1.pdf

Esty, W. W. & Banfield, J. D. (2003, October). The Box-Percentile Plot. *Journal of Statistical Software, 8* (17). Retrieved April 13, 2009 from http://www.jstatsoft.org/v08/i17/paper

Pezzullo, J. C. (2009). Free Statistical Software. Retrieved April 13, 2009 from http://www.statpages.org/javasta2.html

Wikipedia. (2009). *List of statistical packages.* Retrieved April 13, 2009 from http://en.wikipedia.org/wiki/Statistical_software

4. **Q:** What are some advantages and disadvantages of Hirsch's ranking proposal?
   A: Advantages described by Egghe (2006) and Rousseau (2008) include:

* The index is easily calculated and can be applied to any group of researchers in any field.
* By combining two metrics, it is superior to a mere list of publications.
* Small data collection errors are reputed to have little (if any) impact on *h*-values.
* Minor publications as well as single top publications are ignored.

Some disadvantages include:

* As a single numerical measure, the actual quality of the publications is not assessed - nor are other aspects of faculty performance such as teaching skill that should be considered when making faculty promotion decisions.
* It does not measure *recent* academic output - some researchers can easily rest on their laurels.
* Young researchers are disadvantaged since it takes time for citation lists to grow.
* This measure can be inflated through extensive self-citation. The measure does not distinguish between self-citations and citations by others.

Because of these disadvantages, a number of alternative indices have been proposed. The *g*-index (Egghe, 2006), *R*-index (Jin, Liang, Rousseau, and Egghe, 2007) and Maxprod-index [Kosmulski, 2007] are among the more common. However, the whole notion of relying on a single numerical value to quantify achievement is flawed in many respects: multiple measures are necessary to accurately assess performance.

Further reading:

Egghe, L. (2006) An improvement to the h-index: The g-index. I*SSI Newsletter 2*(1) 8-9. Retrieved April 14, 2009 from http://stat-athens.aueb.gr/~jpan/Egghe-ISSI-2006.pdf

Jin, B-H., Liang, L., Rousseau, R. and Egghe, L. (2007). The R- and AR- indices: Complementing the h-index. *Chinese Science Bulletin, 52,* 855-863. Retrieved April 14, 2009 from http://dx.doi.org/10.1007/s11434-007-0145-9

Kosmulski, M. (2007). MAXPROD - A new index for assessment of the scientific output of an individual, and a comparison with the h-index. *International Journal of Scientometrics, Informetrics and Bibliometrics, 11* (1). Paper 5. Retrieved April 14, 2009 from http://cybermetrics.cindoc.csic.es/articles/v11i1p5.pdf)

Panaretos, J. & Malesios, C. (2009, January 18). Assessing scientific research performance and impact with single indices. *MPRA Paper No. 12842*. Retrieved April 14, 2009 from http://mpra.ub.uni-muenchen.de/12842/

Rousseau, R. (2008, June). Reflections on recent developments of the h-index and h-type indices. In H. Kretschmer & F. Havemann (Eds.). *Proceedings of WIS 2008, Berlin*. Retrieved April 14, 2009 from http://www.tarupublications.com/journals/cjsim/7-Rousseau.pdf

Part II: Multiple Choice Questions

1. Q: Which of the following statements is **true** about Poisson distributions?

(A) They approximate binomial (discrete probability) distributions if the sample size is at least 200 and significance level is 0.01 or less.
(B) As the number of events they describe decrease, they resemble normal bell-curve distributions more.
(C) As the number of occurrences they describe increase, the sum of probabilities for each distribution approaches 0.
(D) They are useful in predicting the frequency of periodic events.

A: Option (A) is too conservative. Even with smaller sizes of 100+ with $p < .05$ we should obtain a close approximation to a normal curve (Di Raimondo, 2007). Other authors such as El Sherbiny (2007) and Nandamurar (n.d.) suggest even smaller sample sizes of 20+ offer adequate standard bell-curve approximations in most cases. Since there are several rules of thumb for deciding what a minimum *n*-size is, estimates vary. However, Option (A) seems a bit too restrictive.

Option (B) describes the opposite of what actually happens: as a sampling becomes smaller, chances increase that the distribution will not approximate a normal curve. Conversely as the *n*-size approaches infinity, the distinctions between a Poisson distribution, binominal distribution, and normal distribution gradually become moot.

Option (C) is incorrect since the sum of probabilities approaches 1 rather than zero.

The remaining option is generally true. For this reason Poisson modeling is used to calculate phenomena such as how frequently a word will likely appear or how often a particular type of error occurs.

Further reading: Baytekin, O. (2002) A x^2 analysis of the Poisson approximation to binomial distribution. *Marmara University Journal of Pure and Applied Sciences, 18*, 33-36. Retrieved April 15, 2009 from http://fbe.marmara.edu.tr/dergi/pdf/inga02004.pdf

Di Raimondo, T. et al. (2007). Discrete distributions: hypergeometric, binomial, and poisson. Retrieved April 15, 2009 from http://controls.engin.umich.edu/wiki/index.php/Discrete_Distributions:_hypergeometric,_binomial, and_poisson

El Sherbiny, M. M. (2007, November 4). Discrete Probability Distributions. Retrieved April 15, 2009 from http://faculty.ksu.edu.sa/73212/Publications/Discrete%20Probability%20Distributions.ppt

Nandamurar, K. (n.d.). Poisson Distribution. Retrieved April 15, 2009 from http://www.cse.msu.edu/~nandakum/nrg/Tms/Probability/poisson.htm

West Virginia University Department of Statistics. (2006). The Poisson distribution. Retrieved April 15, 2009 from http://ideal.stat.wvu.edu:8080/ideal/resource/modules/1/Poisson/poisson.html

2. Q: The graph at http://jalt.org/test/Graphics/SSQ7c.gif is an example of a _____.

(A) pareto chart
(B) discrete frequency polygon
(C) cumulative distribution function (a.k.a ogive)
(D) Q-Q plot (a.k.a QQ plot or quartile-quartile plot)

A: Option (C) is correct. A pareto chart is the marriage of two charts: a descending-frequency histogram with a cumulative frequency line-graph. This enables readers to quickly surmise a small set of categorical responses.

Since the SSQ7c.gif plot curve is smooth, at first glance it does not appear to be a discreet frequency polygon – although discrete frequency polygons that are been statistically rounded off can indeed resemble this shape. Unrounded discrete frequency polygons are usually more angular. Option (B) is hence unlikely.

In a Q-Q plot the quantile ranges for two sets of data are compared with each other. Q-plots can exist in many formats. One of common form is when theoretical data is contrasted

with observed data. Another type is when two samples are juxtaposed in the same chart. Usually two distinct distributions are evident in Q-Q plots, but if both distributions are nearly in sync, one distribution might not be visible. So Option (D) seems unlikely, thought not impossible.

Further reading: Cramster, Inc. (2009). Q-Q plot. Retrieved April 17, 2009 from http://www.cramster.com/reference/wiki.aspx?wiki_name=Q-Q_plot

Simon K. (n.d.) Pareto Chart. Retrieved April 17, 2009 from http://www.gate2quality.com/quality-tools_2.html

United Stated Department of Commerce Information Technology Laboratory: Statistical Engineering Division. (2006, July 16). NIST/SEMATECH e-Handbook of Statistical Methods: 1.3.3.24. Quantile-Quantile Plot. Retrieved April 19, 2009 from http://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm

3. Q: Which of the following is least likely to improve the reliability of an examination?

(A) lengthening the number of test items.
(B) standardizing all conditions under which the test is administered, including instructions.
(C) deleting any items from the test that did not correlate with other items.
(D) creating multiple forms of the test.

A: The first three options are widely recognized ways to make tests more reliable. Option (D), however, is unlikely to impact test reliability, at least on the first administration. Arguably, using multiple forms could reduce the likelihood of cheating and hence enhance reliability. Although Option (D) is the least likely way of improving test reliability in the short term, even this option could help make subsequent revisions of the test more reliable as more information about which items function well becomes available.

Further reading: Jacobs, L. C. (1991). Test Reliability. Retrieved April 22, 2009 from http://www.indiana.edu/~best/test_reliability.shtml

Winsteps. (2009). Winsteps Help for Rasch Analysis: Reliability and separation of measures. Retrieved April 22, 2009 from http://www.winsteps.com/winman/index.htm?reliability.htm

4. Q: One difference between the KR-20 and KR-21 is _____.

(A) Only the KR-21 is robust if the unifactor trait is violated, provided the test is longer than 18 items.
(B) Reliability estimates obtained from the KR-21 are generally lower than those obtained from the KR-20.
(C) The former requires just one administration to measure a test's internal consistency, but the later requires several administrations.
(D) The former is used for dichotomous data (i.e. "right or wrong answers"), but the later is used for non-dichotomous data (i.e. partial credit answers).

A: Option (D) describes the "correct answer" from the perspective of classical test theory. Whereas the KR-21 reputedly can handle both whole-credit and partial-credit answers, the Kuder-Richardson 20 formula only works with test items that are either right or wrong.

Actually, both the KR-20 and KR-20 measures have significant limitations. For this reason more and more researchers are adopting IRT and Rasch based models of reliability. However, partly because of ease of calculation, and perhaps because of tradition, the KR-21 formula is not likely to entirely disappear soon. In fact, to cover both bases, it is not uncommon to offer both classical statistics along with IRT and/or Rasch measures of reliability (generally the "information function" for IRT or "separation index" for Rasch).

Concerning robustness, both the KR-20 and KR-21 are reported to be robust even if the unifactor trait is violated (Iacobucci and Duhachek, 2003), so Option (A) is false.

Option (B) is the inverse. As Bodner (1980, par. 4) states, the K-21 "severely underestimates the reliability of an exam unless all questions have approximately the same level of difficulty." In other words, in many situations the KR-21 yields higher reliability estimates than the KR-20.

Option (C) is false since both these tests require just one administration.

Further reading:

Andrich, D. (1982). An Index of Person Separation in Latent Trait Theory, the Traditional KR-20 Index, and the Guttman Scale Response PatternEducation Research and Perspectives, 9 (1) 95-104. Retrieved April 24, 2009 from http://www.rasch.org/erp7.htm

Bodner, G. (1980). Statistical Analysis of Multiple Choice Exams: Coefficients of Reliability. *Journal of Chemical Education, 57*, 188-190. Retrieved April 24, 2009 from http://chemed.chem.purdue.edu/chemed/stats.html

Brown, J. D. (2002). Do cloze tests work? Or, it is just an Illusion? *University of Hawaii Working Papers in Second Language Studies, 21* (1).Retrieved April 26, 2009 from http://www.hawaii.edu/sls/uhwpesl/21(1)/BrownCloze.pdf

Halle, C. D. (2009). Active Teaching, Learning, and Assessment: Unit 4: Validity and Reliability. Retrieved April 24, 2009 from charlesdennishale.com/books/eets_ap/ 3_Psychometrics_Reliatility_Validity_Sampling.pdf

Iacobucci, D. & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology, 13* (4),478 - 487. Retrieved April 26, 2009 from http://mba.vanderbilt.edu/vanderbilt/data/research/2190full.pdf

Linacre, J. M. (1997). KR-20 or Rasch Reliability: Which Tells the "Truth"? Rasch Measurement Transactions, 11 (3) 580 - 581. Retrieved April 26, 2009 from http://www.rasch.org/rmt/rmt113l.htm

5. Q: What is an ideal item facility index range for a 4-choice, multiple-choice norm-referenced test item? What about a true-false norm-referenced test item?

(A) Probably between .4 and .6.
(B) Probably around .63.
(C) Probably around .75.
(D) Probably above .8.
(E) Probably near the total test mean score difficulty.
(F) Actually, more information is needed before deciding that.

A: An argument for Option (F) – getting more information about the test context and intended test use – can certainly be made in situations like this. Many norm-referenced test designers dealing with diverse populations want a wide spread of easy and difficult items to assess persons with differing ability ranges.

Let us assume, for the moment, that a decision must be made with the limited information provided. In 2003 Brown suggests that .50 was the ideal figure for norm-reference tests, regardless of the number of distracter items. Other authors, however, have suggest that the number of distracter choices should be taken into account when determining optimal item difficulty. Lord (1977), offers several different formulas to calculate the optional difficulty level. In one relatively simple procedure first example cited, a perfect score is divided by the number of distractors to ascertain the random guessing level. This is then subtracted from a perfect score and divided by 2. When that sum is added to the random guessing level, the optimal difficulty level is obtained. Let's consider how this would work with a 4-option MC question. The random guessing level is 1.00/4 = 0.25 and hence the optimal difficulty level would be .25 + (1.00 - .25) / 2 = 0.625. In the second example cited, the random guessing level is (1.00/2 = .50) and hence the optimal difficulty level would be .50+(1.00-.50)/2 = .75.

What is important to remember is that there is no single "correct" answer regarding difficulty level since several formulas exist.

Q: What if both these items were designed for a criterion-reference test (CRT)?

A: Again, we could make the argument about the need to know more about the test context and purpose in this case.

If the purpose of a given test is to discriminate between "masters" and "non-masters" of a given skill, then theoretically the ideal solution would be to have all in the so-called masters get the item right and all those who aren't masters get it wrong. Such a test item would have a strong *item discrimination index*: a capacity to distinguish between those who reputedly possess a given skill and those who don't. So it could be argued that item discrimination (ID) is more important than item facility for a CRT test.

If we reflect on the social consequences of test washback, this whole line of thought breaks down. The traditional approach of favoring items only with high ID values might demotivate some learners. For this reason, many teachers try to have a sampling of easy and difficult items to give weaker students some sense of achievement – but not so many that the test itself looses all discriminating value.

Further reading:

Brown, J. D. (2003). Norm-referenced item analysis (item facility and item discrimination). *Shiken: JALT Testing & Evaluation SIG Newsletter, 7* (2) 16 – 19. Retrieved April 16, 2009 from http://jalt.org/test/PDF/Brown17.pdf

Lord, F. M. (1977). Optimal number of choices per item: A comparison of four approaches: Journal of Educational Measurement ,14 (1), 33-38.

The University of Texas at Austin Division of Instructional Innovation and Assessment. (2007, July 16). Analyzing Multiple-Choice Item Responses. Retrieved April 16, 2009 from http://www.utexas.edu/academic/mec/scan/analysis.html

Whatley, M. A. (2007). Item Analysis Worksheet. Retrieved April 16, 2009 from http://chiron.valdosta.edu/mawhatley/3900/itemanalysis.pdf

HTML: http://jalt.org/test/SSA7.htm / **PDF**: http://jalt.org/test/PDF/SSA7.pdf