Suggested Answers for Assessment Literacy Self-Study Quiz #1

by Tim Newfields

Here are some possible answers to the questions about testing, statistics, and assessment which were raised in the December 2006 issue of SHIKEN. Please note that some of these answers are not definitive - often a variety of solutions are possible. If you feel an answer is unclear or disagree with a conclusion, please contact the editor.

Part I: Open Questions

1. Q: What's the formula for the ubiquitous *hensachi* used in ranking schools in Japan?

A: According to the Japanese version of Wikipedia, it is:

$$\text{偏差値} = \frac{10(y-\bar{x})}{\sqrt{\frac{1}{n}\sum_{i-1}^{n}(x_i-\bar{x})^2}} + 50$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

n:標本数　Σ:総和　x_i:個々の標本値　$\bar{x}$:平均値　y:偏差値を求めたい標本の値

Figure 1. The formula for standardized rank score [hensachi].

NOTE: n = sample size, ∑ = sum, x_1 = individual scores, mean = mean, and y = standard deviation

Hence a school precisely at the center of its group would have a *hensachi* rating of 50. A school that's one standard deviation above the mean, roughly in the 85th percentile for its group, would have a *hensachi* rating of 60. A school one standard deviation under the mean, which should be near the 16th percentile, would have an expected *hensachi* rating of 40. The entire *hensachi* range is from 20 - 80 and 95.4% of all schools fall within the 30 - 70 range.

Each year the leading cram schools calculate the *hensachi* scores for educational institutions throughout Japan. The term *hensachi* does not appear to be well-understood by those in the West. It has been mistranslated as "standard score", "T-score", "standard deviation", "deviation value", or simply "scores". A better translation of this term might be "standardized rank score". It is somewhat akin to the CEEB scale used by ETS. Brown (1995, p. 25, quoted in Poole, 2003) describes it as the "abstract notion of a national norm-referenced person-indexed score." Since cram schools have different score data in their survey samples, *hensachi* ratings vary slightly from school to school.

Considering the shoddy nature of many school entrance exams, the whole *hensachi* rating system should be critically questioned. It may not be too far to suggest that it is merely a pseudo-scientific method of evaluating student performance and the predictive validity of *hensachi* ratings is questionable.

Further reading:

Poole, G. (2003) Assessing Japan's Institutional Entrance Requirements. Asian EFL Journal, 5 (1). Articles 4. Retrieved March 11, 2007 from http://www.asian-efl-journal.com/march03.sub5a.php

Tomoya. (2000). Algorhytm. Retrieved March 11, 2007 from http://www5d.biglobe.ne.jp/~tomoya03/shtml/algorithm/Hensa.htm

Wikipedia. (2007). Hensachi [in Japanese]. Retrieved March 11, 2007 from http://ja.wikipedia.org/wiki/PO]t�$

2. Q: At the RateMyProfessors.com (www.ratemyprofessors.com) website claiming to have over six million ratings over 700,000 professors in North America, the British Isles, Australia, and New Zealand two different professors teaching the same course received these ratings:

		Total Ratings	Total Quality
Professor A	🙂	4	4.3
Professor B	😐	18	3.1

What can we assume about these teachers based on this information? And what are some of the validity issues to consider when examining any data from the web sources such as RateMyProfessors.com?

A: This sort of rating has no more validity that cafeteria gossip. The survey samples are generally small and it's difficult to verify precisely who is inputting the data. Moreover, the persons who do bother to input that data will tend to be either more positive or negative than the entire student sample. In short, this sort of evaluation offers only a very fuzzy picture of what's going on.

3. Q: A teacher wants to select a random sample of high school students out of a total population of 400 students (*N*=400). How big does her sample need to be (*n*=?) to have a 95% confidence interval for that population?

A: Actually, this question is misleading. "Confidence level" and "confidence interval" refer to different phenomena and these terms are sometimes confused. Confidence levels are set a priori and sample size doesn't need to be an issue in their determination. In most social sciences, confidence levels of 95% or 99% are widely used. The question probably should have been phrased, "With a .95 confidence interval and 5% margin of error for a population of 400, what is the required sample size?" Mathematically, this could be expressed this way:

$$\mathbf{ss} = \frac{Z^2 \times P \times (1-P)}{CI^2}$$

in which "Z" represents the Z value, which is 1.96 at 95%. "P" represents the decimal value of responds who answer a survey item a specific way. For surveys which has many items, a safe choice would be to assume a low response such as .1. "CI", as you might guess, is the confidence interval.

Confidence intervals can be calculated from either a t-test or a normal distribution. If you're dealing with a sample size of less than 30, a t-distribution table should be consulted to obtain the confidence interval. Simply multiply the confidence interval (often expressed as the Greek lower case symbol µ or the abbreviation "CI") with the standard error:

$$CI = z \text{ x } [p \text{ x } (1 - p) / n]^{-1/2}$$

in which "p" represents the decimal value of the assumed accuracy of a system, and "n" represents the number of times a test record was obtained, and "z" represents a value derived from a Gaussian curve known as the "level of confidence".

Plugging in this data, here is what is obtained a test given only once:

$$CI = 1.96 \text{ x } [.95 \text{ x } (1 - .95) / 1]^{-1/2}$$

We arrive at a confidence interval of 1.33 and survey sample size of 2.86. That means if the total population is 400, a random sample of about three persons is needed. (Isn't this too small??)

Further reading:

Sauro, J. (n.d.). Measuring usability: You don't need a large sample of users to obtain meaningful data. Retrieved March 11, 2007 from http://www.measuringusability.com/sample_continuous.htm

Ruggles, T. (2002). Calculation of the confidence interval. Retrieved March 11, 2007 from http://bio-tech-inc.com/calculation_of_the_confidence_interval.htm

Zucker, H. (1998). Sample size calculator. Retrieved March 24, 2007 http://www.surveysystem.com/sscalc.htm

4. Q: A survey was conducted of student attitudes towards a placement test. Three hundred questionnaires were handed out, but the response rate was 60%. The 120 students who did not respond were ignored and all research findings were based on the remaining 180 responses. When 144 students (80% of the survey sample) indicated support for a given idea, the administrator of the survey claimed "80% of the students were in favor of X." Any problem?

A: This is an all-too-common way of skewing statistics . A more ethical reporting procedure would indicate that only 44% (i=144) of the respondents expressed support for X and that 40% (*n*=120) of the survey recipients did not respond at all. Be sure to indicate precisely how the respondents were selected from the larger population and avoid undue generalizations from your small sample to the population at large.

Further reading:

Center for Media & Democracy. (2004). Disinformation. Retrieved March 11, 2007 from http://www.sourcewatch.org/index.php?title=Disinformation

5. Q: How widely accepted is Yerkes-Dodson Law (1908) today? What pedagogical applications, if any, does it have?

A: The Yerkes-Dodson law postulates a correlation between stimulus strength (sometimes interpreted as "arousal") and habit formation (sometimes interpreted as "performance") for tasks of varied difficulty. It suggests that there is an optimal level of arousal after which performance degrades and is a good example of "folklore" psychology which sounds convincing in general principle, but perhaps naive in suggesting their is just one kind of "performance" in each situation and only one type of "arousal".

Further reading:

Teigen, K. H. (1994). Yerkes-Dodson: A law for all seasons. Theory & Psychology, 4 (4). Retrieved March 11, 2007 from http://tap.sagepub.com/cgi/content/abstract/4/4/525

Yerkes-Dodson law. (2007, March 15). In Wikipedia, The Free Encyclopedia. Retrieved March 17, 2007 from http://en.wikipedia.org/wiki/Yerkes-Dodson_law.

6. Q: What information should technically be contained in an informed consent form? For EFL research would simple verbal agreements suffice?

A: There seems to be a big gap between what is technically supposed to happen and what commonly takes place. Technically, a survey form should at least include: (1) some general statement of the purpose of the research project, (2) what sort of information will be disclosed, (3) whom the information will be disclosed to, (4) the rewards and potential hazards of participating in any research, and (5) who to contact if there is a concern or grievance. The Japan Psychological Association has adopted a number of the guidelines regarding disclosure recommended by the APA. However, as Nichigai (1994) suggests there is something "paternalistic" about lot the way much research is conducted in Japan. The term "consent form", which is literally translated as "*[Menseki] Dou-i-sho*", itself is more often apt to be "*Go-kyouryouku no O-negai*" - which essentially amounts to a "request for cooperation". In the medical field, particularly when clear risks are involved in any treatment, explicit consent forms are widely used. In most linguistic research studies with supposedly low risks, however, generally only a tacit verbal agreement to participate in research is obtained.

Further reading:

Nagoya University. (2007). Kojin jouhou hogo seido ni tsuite.[About standards for protecting private information]. Retrieved March 17, 2007 from http://www.nagoya-u.ac.jp/info/personal_info/file.html#a_ken_kokusai

Ninchoji, T. (1994, July). The effect of an introduction of a new consent form on settlement of the doctrine "informed consent". Japan Hospital Journal (13), 45-7.

Pious, S. (n.d.) Tips on informed consent. Retrieved March 11, 2007 from http://www.socialpsychology.org/consent.htm

Part II: Multiple Choice Questions

1. Q: To calculate the reliability coefficient for some data, which of the following variables are needed?

(A) Standard Deviation (SD) from a specific measure
(B) Standard Error of Measurement (SEM)
(C) Standard Error of Estimation (SEE)
(D) Sampling Distribution of the Mean

A: There are actually several different types of reliability coefficients, the most common of which is Cronbach's alpha reliability coefficient. This is calculated through the following formula:

$$\alpha = \frac{N \cdot \bar{r}}{1 + (N - 1) \cdot \bar{r}}$$

To do this computation, we need to know the variance of individual items on a test as well as the variance for the sum of all items. What this question boils down to how to compute variance. Since variance can be defined as "the average of squared differences around the mean" (Mousavi, 2002, p. 819) technically we do not need any of the four choices (A-D) mentioned in Question 1: if the scores, number of test items, and means are known the rest of the information can be calculated.

Cronbach's alpha reliability coefficient can be described as "the average inter-item correlation among . . . items" (UCLA Academic Technology Services) and it ranges from 0 to 1.

Further reading:

Brown, J.D. (2002, February). The Cronbach alpha reliability estimate. Shiken: JALT Testing & Evaluation SIG Newsletter. 6 (1) 16 - 18. Retrieved March 11, 2007 from http://jalt.org/test/bro_13.htm

Variance. (2002). In S. A. Mousavi An Encyclopedic Dictionary of Language Testing. (3rd Ed.). (pp. 819-820). Taipei: Tung Hua Book Company.

StatSoft Inc. (1984). Reliability and Item Analysis. Retrieved March 11, 2007 from http://www.statsoft.com/textbook/streliab.html

UCLA Academic Technology Services. (n.d.). SPSS FAQ: What does Cronbach's alpha mean? Retrieved March 11, 2007 from http://www.ats.ucla.edu/STAT/SPSS/faq/alpha.html

2. Q: To calculate the Kuder-Richardson 21 internal consistency reliability for some data, three variables are needed: (1) the mean, (2) the standard deviation squared, and (3) _________.

(A) the number of respondents (C) the range of scores for that test
(B) the number of items in a test (D) the number of "misfitting" items in that test

A: Option (B) is needed. For a good explanation of the KR-21, refer to Dörnyei (2007).

Further reading:

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics* (Oxford Applied Linguistics) Oxford: Oxford University Press.

Kuder-Richardson Reliability Coefficient. (2002). In S. A. Mousavi, *An Encyclopedic Dictionary of Language Testing*. (3rd Ed.). Taipei: Tung Hua Book Company. p. 374-375.

3. Q: To calculate a chi-square statistic with one degree of freedom for two groups, which of the following is NOT needed:

(A) the mean score of each group (C) the size of each sample
(B) the standard deviation of each group (D) the range of scores for the group

A: The short answer is "C", but this warrants some explanation. Since this statistic tells us about the difference between the observed and hypothetical distribution for two or more variables, a natural question is, "How can I know what the hypothetical distribution of some phenomena should be?" With classical computational methods, strictly speaking you would need a random sample and continuous interval data to know what a hypothetical distribution should be. However, rightly or wrongly, statisticians will sometimes assume that data from a large convenience sample or nominal data will have a gaussian distribution. To know the values for that distribution, one should consult a F-distribution table. Many are available online.

Further reading:

Lemmon, A. C. (n.d.). EvoTutor. Retrieved March 22, 2007 from http://www.evotutor.org/Statistics/St4A.html

Wikipedia. (2007). Chi-square distribution Retrieved March 11, 2007 http://en.wikipedia.org/wiki/Chi-square_distribution

4. Q: Which of the following is NOT a characteristics of a curve with a perfectly normal distribution:

(A) It is platykurtic. (C) It is bell-shaped.
(B) It is asymptotic. (D) The mean, median, and mode are the same.

A: A perfectly normal curve would not be platykurtic (with a negative kurtosis) - so (A) is the correct answer. Platykurtic curves have a relatively flat peak and more rounded "shoulders".

Further reading:

Heffner, C. L. (2004). Statistics Primer (Revised). Retrieved March 18, 2007 from http://allpsych.com/stats/unit1/17.html

5. Q: To find out how the total score on a test correlates with the chance of getting a single item on that test correct, a ___________ should be ascertained.

(A) point biserial correlation
(B) Pearson product-moment correlation
(C) point estimation
(D) Spearman's Rho

A: The correct answer is (A). This could be done through by determining the point biserial correlation coefficient for each item. Another way is to measure the item-total correlation.

Further reading:

Kehoe, J. (1995). Basic item analysis for multiple-choice tests. Practical Assessment, Research & Evaluation, 4 (10). Retrieved March 11, 2007 from http://PAREonline.net/getvn.asp?v=4&n=10.

HTML: http://jalt.org/test/SSA1.htm PDF: http://jalt.org/test/PDF/SSA1.pdf