*SHIKEN: JALT Testing & Evaluation SIG Newsletter.* March 2011. *15* (1) 19-30.

# Suggested Answers for Assessment Literacy Self-Study Quiz #10

by Tim Newfields

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

## Part I: Open Questions

1. **Q:** What does a *regression fallacy* refer to? How can it be avoided?

**A:** To answer this, it is essential to understand what is meant by "regression toward the mean". This term is also known as a "regression effect" or "regression artifact" and was first described by Galton in 1875. Basically, it refers to a tendency for data samples to move closer towards the mean as additional samples are obtained.

Let's illustrate this concept with a concrete example. Suppose that university students take a standardized test as they enter school, then a different version of the same test two semesters later. What will likely happen is that those who scored below average on that test will be more likely to score higher when retaking it, even if they didn't study or actually learn anything in the interval. Conversely, those who scored high the first time will have a greater chance of a subsequent score drop, purely for statistical reasons. In other words, high or low pretest scores to tend to move toward the mean on the posttest regardless of treatment. Smith and Smith (2005) explain this by stating:

> Because observed test scores are an imperfect measure of ability, high scores are typically an overestimate of ability and low scores are typically an underestimate — causing high and low scores to regress to the mean in subsequent tests. (p. 395)

If we understand regression toward the mean, it should be easy to guess what is meant by a *regression fallacy*. Ascribing random tugs towards the norm to non-random causes is known as a regression fallacy. In testing contexts, this occurs when test score gains or score drops are mistakenly attributed to some external factor such as "improved ability".

This begs the question: how can one tell whether a genuine improvement has occurred, or a score change is merely due to stochastic fluctuations? There are many different ways to resolve this, depending on the type of data involved. Let's say we are dealing with pretest/posttest scores, both of which should be regarded as ordinal data. Trochim (2006) provides an easy to follow online explanation of how to calculate the regression effect in such a scenario. Another way to calculate regression effects is described by Ostermann, Willich, and Lüdtke (2008). Whereas Trochim's method can be employed by any teacher, Ostermann, Willich, and Lüdtke's approach requires a more sophisticated understanding of algorithms.

How can regression fallacies be avoided? Poulton (1994, p. 128, 134) underscores the need for researchers to be educated more about regression in general. It should be emphasized that regression itself is not problematic - incorrectly ascribing data shifts to non-random variables is the problem. Still, it should go without saying that as the gap between a person's **true score** and **observed score** widens, so does the regression toward the mean. Obviously, regression effects can be attenuated if a test's observed score approximates its true score. This occurs when the measurement error of a test is minimal. In other words, if a test measures what it purports to for the sample it was designed for, regression effects will be attenuated – but practically speaking, regression artifacts are a feature of all experiments.

**Further Reading:**

Dallal, G. E. (2000). The regression effect - The regression fallacy. Retrieved March 11, 2011 from http://www.jerrydallal.com/LHSP/regeff.htm

Lohman, D. F. & Korb, K. A. (2006). Gifted today but not tomorrow? Longitudinal changes in ability and achievement during elementary school. *Journal for the Education of the Gifted, 29*(4) 451-484. doi: 10.4219/jeg-2006-245

Ostermann, T., Willich, S. N., & Lüdtke, R. (2008). Regression toward the mean – A detection method for unknown population mean based on Mee and Chua's algorithm. *BMC Medical Research Methodology, 8*(52) n.p. doi:10.1186/1471-2288-8-52

Poulton, E. C. (1994). *Behavioral decision theory: A new approach*. Cambridge, UK & New York, NY: Cambridge University Press.

Smith, G. & Smith, J. (2005). Regression to the mean in average test scores. *Educational Assessment, 10*(4) 377-399. doi: 10.1186/1471-2288-8-52

Trochim, W. (2006). *Research methods knowledge base: Regression toward the mean*. Retrieved March 10, 2011 from http://www.socialresearchmethods.net/kb/regrmean.php

2. **Q:** What is *test maintenance*? What sort of *test maintenance* procedures should schools creating their own entrance exams employ?

**A:** Test maintenance refers to the way that an organization maintains the reliability, validity, and security of its tests over time. One way to conceptualize test maintenance is in terms of a systems development life cycle. This way of viewing tests is heavily influenced by the field of systems engineering. In that paradigm, test maintenance occurs after an existing test is implemented, providing a basis for subsequent revisions.

This concept of "test maintenance" has also been described in terms of a test development cycle by Breen (1989) and subsequently by Wigglesworth and Elder (1996) as well as Weir and Milanovic (2003). Although the process is cyclic, a conceptual final step consists of "evaluation and

revision" and much of those processes overlap with procedures described in the previously mentioned "test maintenance" phase of a systems development life cycle.

For a typical school entrance exam, what specific factors should be monitored? JLTA's *Code of Good Testing Practice* (2007) makes it clear that an ethical test is transparent in that each stage of the test construction process is open to scrutiny. Obviously, the item facility and item discrimination of each test item should be analyzed and if a test item is not measuring what it purports to, the weighting of that item should be adjusted. Moreover, the descriptive statistics for each test - its mean, skewness, kurtosis, and standard deviation as well as its reliability and criterion validity - should be considered carefully and if some area is found to be problematic, test designers should consider ways of changing it.

A good example of a test maintenance study within a classical test theory framework appears in Ito (2005). Aline and Churchill (2006) also provide a commendable validation study of a university entrance exam from a Rasch perspective. The recommendations they make about subsequent versions of the test examined illustrates how test maintenance should be conducted.

Needless to say, test maintenance is expensive and schools in Japan appear to be rather skimpy about it. The Japanese schools I have observed tend to focus heavily on their standardized rank scores (Jp: *hensachi*) as well as test security, paying minimal attention to issues concerning test validity and reliability.

**Further Reading:**

Aline, D. & Churchill, E. (2006). Analyzing entrance exam item types with Rasch. *Kanagawa Daigaku Gengo Kenkyuu,* 28, 125-142. Retrieved on March 8, 2011 from http://hdl.handle.net/10487/3846

Breen, M. (1989). The evaluation cycle for language learning tasks. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 187-206). Cambridge: Cambridge University Press.

Ito, A. (2005) Validation study on the English language test in a Japanese nationwide university entrance examination. *Asian EFL Journal, 7*(2) 6. Retrieved on March 8, 2011 from http://www.asian-efl-journal.com/June_05_ai.php

Japan Language Testing Association. (2007). *JLTA code of good testing practice*. Retrieved on March 8, 2011 from http://www.avis.ne.jp/~youichi/COP.html

Weir, C. J. & Milanovic, M. (Eds.) (2003). *Continuity and innovation: The history of the Cambridge Proficiency Exam 1913-2002,* Studies in Language Testing 15. Cambridge: Cambridge University Press/UCLES.

Wigglesworth, G. & Elder, C. (Eds). (1996). *The language testing cycle: From inception to washback*. Canberra, Australia: Australian National University.

3. **Q:** What is *questionnaire acquiescence*? Why should it be of concern to survey designers? How can it be reduced?

**A:** Questionnaire acquiescence occurs when respondents attempt to finish a survey, test, or a voting ballot as quickly as possible and avoid making more time-consuming or controversial responses. Let us focus on surveys. Questionnaire acquiescence occurs under two conditions: (1) if respondents feel indifferent about the task and therefore speed through it, or (2) if respondents worry about expressing their opinions candidly about an issue because they do not feel their confidentiality and/or freedom to respond honestly is adequately assured. In both cases, the data becomes skewed.

Questionnaire acquiescence can be reduced the following ways: making surveys short and easy to complete, providing equitable incentives for respondents to complete survey tasks fully, proving a comfortable time frame for respondents to do all of the tasks, and providing clear confidentiality guarantees. Questionnaire acquiescence can also be reduced by masking the researcher's personal attitudes regarding an issue being explored. For example, when designing a questionnaire about the effectiveness of a program, it is advisable to mix positive statements such as "I enjoyed this program" with negative ones such as "this program was a waste of time". In that way, the researcher's agenda is less obvious. All too often, the sub-text message in amateur surveys is easy to discern and respondents have a tendency to agree with the researcher's expectations out of politeness.

**Further Reading:**

Ray, J. J. (1990). Acquiescence and problems with forced-choice scales. *Journal of Social Psychology*, 130(3), 397-399. Retrieved on March 4, 2011 from http://jonjayray.tripod.com/forcho.html

O'Muircheartaigh, C., Krosnick, J.A., & Helic, A. (2000). Middle alternatives, acquiescence, and the quality of questionnaire data. *The Harris School Working Papers Series, 1*(3) n.p. Retrieved on March 13, 2011 from http://harrisschool.uchicago.edu/about/publications/working-papers/abstract.asp?paper_no=01.03+++

Saris, W. E., Krosnick, J. A., & Shaeffer, E. M. (2005). *Comparing questions with agree/disagree response options to questions with construct-specific response options*. Unpublished manuscript, Political, Social, Cultural Sciences, University of Amsterdam.

4. **Q:** In a fixed choice test, what is the difference between an *illegal value* and an *outlier*? How can test designers reduce both of these?

**A:** An *illegal value* is a response that is outside of the range of valid options available. This term is most widely used in computer programming, but also is relevant to test analysis. In multiple-choice tests, the most common type of illegal value occurs when more than one response is selected under conditions when only one response is permitted. It is harder to judge illegal values with open response test items. However, if a test asks respondents to describe in detail how they would respond to a specific situation, and an examinee pumps out lots of fluff without indicating

any clear response, that could be considered an illegal value.

Illegal values can be reduced by giving clear test instructions, providing one sample answer at the start of each new task type, and also indicating the consequences of illegal responses. For example, if only one correct response is permitted to the multiple-choice test problem, that should be specified in the instructions.

An *outlier* is simply an unusual (but in most cases legal) response. For example, a test item that behaves in a markedly different way from other test items should be considered an outlier. Interestingly, there is no single standard for determining whether or not a datum is an outlier. Renze (1999) suggests that items 1.5 times below the first quartile or above the third quartile be considered outliers, but in different contexts other cutoff points may be more appropriate. Practical ways of dealing with outliers are discussed at length by Rousseeuw and Leroy (2003).

Let's briefly consider two sample outliers, and then reflect on why they may have occurred. One situation would be if a test item was "difficult" to the majority of examinees, but somehow several low-scoring students got it right. In such a scenario, the possibility that those students merely guessed the correct answer should be considered. The converse scenario is more problematic: if high-scoring students who did well on a test as a whole did poorly on one item that most other students got right, that item needs to be examined closely. There's a good chance that such an item may be interpreted in more than one way.

Outliers are especially prone to occur if the data distribution is heavily tailed in either direction, if there are recording errors, or if there is ample measurement error (Osborne & Overbay, 2004). Another common source of outliers is when two or more distinct sub-groups exist within a sample. For instance, if we examined the EFL tests scores of incoming students at one university, those who have lived in an English-speaking country for several years or more would likely have different score patterns from those who have not.

As Taleb (2007, 2010) has eloquently expressed, outliers can be likened to "black swans" in that they are often hard to predict, and their impact is often disproportionate to their numbers. They should not be lightly dismissed, and often produce interesting research questions that lead to fresh discoveries.

**Further Reading:**

NIST/SEMATECH e-Handbook of Statistical Methods. (n.d.). What are outliers in the data? Retrieved on March 15, 2011 from http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

Osborne, J. W. & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation, 9*(6). n.p. Retrieved March 14, 2011 from http://PAREonline.net/getvn.asp?v=9&n=6

Renze, J. (1999). MathWorld, A Wolfram Web Resource created by E. Weisstein: Outlier. Retrieved on March 10, 2011 from http://mathworld.wolfram.com/Outlier.html

Rousseeuw, P. J. & Leroy, A. M. (2003). *Robust regression and outlier detection* (Wiley Series in Probability and Statistics). New York: Wiley-Interscience.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. New York, NT: Random House.

Taleb, N. N. (2010). *The black swan: The impact of the highly improbable* (2nd Edition). New York, NT: Random House & Penguin.

5. **Q:** What steps should be taken to discourage the leakage of entrance examination test items onto the Internet while the exams are being held in a manner similar to what happened during the Kyoto, Doshisha, Rikkyo and Waseda university entrance examinations of February 2011?

**A:** A government panel is being set up to explore that question. It seems likely that they will recommend restricted access to restrooms during exams and require all cell phones to be placed in sealed pouches during that period. They might also recommend that exams undergo data forensic methods to statistically detect when cheating is likely occurring. In the USA companies such as Caveon Test Security have sprung up to systematically analyze high-stake tests and scan for anomalies that indicate probable cheating. Elsewhere in the world we are likely to see more forensic companies using Bayesian principles to detect when fraudulent responses are occurring.

A more creative solution to high-tech cheating might be to consider having more "open book exam" test items in which all test takers are permitted to use any data source they wish (citing those sources appropriately in the same way all academic research should be cited) to obtain answers to complex questions. In some ways, that would simulate real life situations far better than the archaic university exams.

**Further Reading:**

Gabriel, T. (2010, December 27). Cheaters find an adversary in technology. *New York Times: Online Edition*. Retrieved on March 10, 2011 from http://www.nytimes.com/2010/12/28/education/28cheat.html

Univ. entrance exam cheats go online. (2001, March 12). *Japan Times Weekly*. Retrieved on March 10, 2011 from http://weekly.japantimes.co.jp/nn/univ-entrance-exam-cheats-go-online

High-tech cheating in entrance exams. (2011, March 2). *Asahi Shimbun: English Web Edition*. Retrieved on March 10, 2011 from http://www.asahi.com/english/TKY201103010401.html

Taroni, F., Bozza, S., Biedermann, A., Garbolino, P., Aitken, C. (2010). Data analysis in forensic science: A Bayesian decision perspective (Statistics in Practice). Chichester, West Sussex: John Wiley & Sons, Ltd.

## *Part II: Multiple Choice Questions*

1 **Q:** In surveys and structured interviews, which statement is true about *filter questions*?

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

(a) Not all respondents are expected to answer the question(s) following a filter question.

(b) They are designed to assess whether or not respondents are being honest.

(c) They are designed to detect whether or not some type of guessing is occurring.

(d) Their primary purpose is to prime respondents for ensuing questions.

**A:** Option (a) is the correct answer. Filter questions, which are also known as *contingency questions*, allow different sets of questions to be asked depending on how each filter question is answered. Let's illustrate this with a concrete example. Suppose you were designing a survey on attitudes towards foreign language learning such as the one by Tsuda (2003). One filter question that would likely appear in such a survey is whether or not the respondents had lived overseas. Those who answered "no" would next be asked a different, unrelated question. Those who answered "yes" would be asked a follow-up question(s) about the length or location of their overseas experience. In other words, only respondents answering a filter question positively would be asked a follow-up question.

Filter questions have two basic designs. One is to raise all filter questions initially, then all follow-up questions. Another procedure is to ask a follow-up question after each positively answered filter question. Not surprisingly, these two different formats sometimes yield different response patterns (Henning, 2010).

Questions whose primary purpose is to assess whether or not respondents are being honest are generally known as *norming* or *calibrating* questions. Used primarily in legal investigations, they often rely on a complex baseline calibration systems.

Systematic exploration of guessing would likely examine hedging protocols and rely on detailed verbal transcriptions. An example of one such study can be found in Yu (1999). Questions designed to elicit hedges (such as asking the exact population of a given city) would be known as *elicitation* questions.

Option (d) describes *priming questions*. Priming is a phenomenon in which a previously mentioned survey or test item influences the response to a latter item. Priming is sometimes described as a question-order effect and according to Lasorsa (2003), it can be a source of significant context variance. For this reason surveys generally seek to reduce – or at least control for – priming effects. One strategy is to use several randomized alternative forms of the same survey. Another is to add "buffer questions" between core questions (Wänke & Schwarz, 1997). Yet another is to recognize question-order effects as inevitable and simply try to be consistent and explicit about the order.

**Further Reading:**

Albrecht, S. A., Albrecht, C. C., Albrecht, C. O., & Zimberland, M. (2009). Fraud examination (3rd Edition). Mason, OH: South-Western Cengage Learning.

Henning, J. (2010). *Sequential vs. grouped placement of filter questions*. Retrieved March 15, 2011 from http://blog.vovici.com/blog/bid/28235/Sequential-vs-Grouped-Placement-of-Filter-Questions

Lasorsa, D. L. (2003). Question-order effects in surveys: The case of political interests, news attention, and knowledge. *Journalism & Mass Communication Quarterly, 80*(3) 499-512. Retrieved March 17, 2011 from http://www.aejmc.org/_scholarship/research_use/jmcq/03fall/*lasorsa.pdf*

Tsuda, S. (2003). Attitudes toward English language learning in higher education in Japan: Raising awareness of the notion of global English. *Intercultural Communication Studies, 12*(3) 61-75. Retrieved March 18, 20011 from http://www.uri.edu/iaics/content/2003v12n3/06%20Sanae%20Tsuda.pdf

Trochim, W. (2006). *Research methods knowledge base: Types of questions*. Retrieved March 15, 2011 from http://www.socialresearchmethods.net/kb/questype.php

Wänke, M. & Schwarz, N. (1997). Reducing question order effects: The operation of buffer items. In L.E. Lyberg, et al. (Eds.) *Survey measurement and process quality* (Wiley Series in Probability and Statistics). (pp. 115-139). New York: John Wiley & Sons, Inc.

Yu, S. (1999). The Pragmatic Development of Hedging in EFL Learners. Unpublished Ph.D. Thesis. City University of Hong Kong. Retrieved March 15, 2011 from http://lbms03.cityu.edu.hk/theses/ftt/phd-en-b23749398f.pdf

2. **Q:** Which statements are true about *fundamental attribution errors*?
   NOTE: *Two of the statements below are considered correct.*

   (a) They generally pertain to the behavior of single individuals.
   (b) They occur if some type of stereotype is made about outgroup members as a whole.
   (c) They over-emphasize personality trait variables.
   (d) They over-emphasize situational or environmental variables.

**A:** The correct answers are (a) and (c).

Individual-environmental interactions are complex and controversies regarding the extent that behaviors should be ascribed to personality or to environmental conditions have been perennial. According to Ross (1977), the tendency of people to ascribe the behaviors of others to personality variables such as "character" rather than situational variables such as "interlocutor power gaps" is known as a *fundamental attribution error*. The opposite tendency, to ascribe behaviors to environmental factors rather than to individual personality traits represents a different type of cognitive error. Needless to see, different academic disciplines (and researchers) tend to focus on different parts of the individual-environmental spectrum.

Statement (b) pertains to *ultimate attribution error*s (Pettigrew, 1979). In many ways ultimate attribution errors are similar to fundamental attribution errors, but they involve

stereotyping outgroup members as a whole rather than single individuals. Let's highlight this with an example. Suppose a teacher sees one Japanese student with bleached hair sleeping in a class, then promptly decides that the student is lazy. That would be a fundamental attribution error because it is possible that the student is not lazy, merely working several different part-time jobs to pay for his education. Now if the teacher thought, "All Japanese with bleached hair are lazy" after this incident, that would be an ultimate attribution error. (**NOTE**: Sharp readers might notice an implicit ultimate attribution error in the previous sentence: some Japanese teachers might have bleached hair and social theory predicts they would be less likely to make this sort of over-generalization.)

**Further Reading:**

Harper, M. (2009). *Fundamental attribution error*. Retrieved March 15, 2011 from http://www.knowledgerush.com/kr/encyclopedia/Fundamental_attribution_error/

Pettigrew, T.F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin, 5*(4) 461-476. doi: 10.1177/014616727900500407

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. (Ed.), *Advances in experimental social psychology*. (pp. 173-220). New York: Academic Press. doi: 10.1016/S0065-2601(08)60357-3

3. **Q:** If you were using OCR sheets in a large scale survey, which of the following would be considered *data processing errors*? NOTE: *At least two statements below are correct*.

(a) A response sheet is rejected because too many items were left blank.
(b) A response sheet was not collected because an examinee failed to turn it in.
(c) A response sheet was not marked because it stuck to the previous response sheet.
(d) A poorly marked response sheet was misread.

**A:** The last two options are data processing errors that can be ascribed to *machine error*.

Statement (b) is a non-sampling error that is usually described as a "non-response". In a sense this might be regarded as a "human data processing error" if we consider survey administration as closely linked to the data processing.

However, the first option represents a different type of non-sampling error. Here were have an intentional elimination of a response sheet because it failed to fulfill minimum criteria. That is entirely valid, as long as the criteria for deletion is clear to readers. What sometimes happens is that researchers reject some surveys because of unspecified criteria – minimal acceptance criteria should be specified. Some researchers might decide to reject surveys if more than 30% of the items are not completed – others might accept all surveys even if only 1 item has been completed. Decisions

about which surveys to accept/reject can sometimes have a big impact on survey results.

**Further Reading:**

Groves, R. M. (2004). *Survey errors and survey costs* (New edition), New York: Wiley-Interscience.

Statistics Canada - Statistique Canada. (2010). *Non-sampling error*. Retrieved March 16, 2011 from http://www.statcan.gc.ca/edu/power-pouvoir/ch6/nse-endae/5214806-eng.htm

**4. Q:** Assuming that we are dealing with data from a normally distributed curve, which of these statements would be true according to the *three-sigma rule*?

(a) 1 in 22 observations will likely be at least one standard deviation above or below the mean.

(b) 1 in 68 observations will likely be at least two standard deviations above or below the mean.

(c) 1 in 95 observations will likely be at least three standard deviations ± the mean.

**A:** The correct answer is Statement (c). The three-sigma "rule" is a very rough and quick way of understanding a Gaussian curve. Assuming a curve has a perfectly normal distribution – which is a chancy assumption with small sample sizes – about 68% of the data will fall within one standard deviation above or below the mean. 95% of the data will lie within two standard deviations above or below the mean, and 99.7% will occur within three standard deviations. It should be emphasized that this applies only if a distribution curve is perfectly normal - a condition that's frequently approximated, but seldom completely realized. When interpreting data, a normality test such as the D'Agostino-Pearson omnibus test or a Rasch goodness-of-fit measure should be applied before attempting to use the three-sigma rule.

**Further Reading:**

Analyse-it Software, Ltd. (2008). *Testing the assumption of normality*. Retrieved March 18, 2011 from http://www.analyse-it.com/blog/2008/8/testing-the-assumption-of-normality.aspx

Drexel University Math Forum. (2008). *Testing a set of data for normal distribution*. Retrieved March 18, 2011 from http://mathforum.org/library/drmath/view/72065.html

Laerd Statistics. (n.d.). *Testing for Normality using SPSS*. Retrieved March 18, 2011 from http://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics.php

Motulsky, H. (2009). *Normality tests - use with caution*. Retrieved March 18, 2011 from http://www.graphpad.com/library/BiostatsSpecial/article_197.htm

**HTML**: http://jalt.org/test/SSA10.htm / **PDF**: http://jalt.org/test/PDF/SSA10.pdf