SHIKEN: JALT Testing & Evaluation SIG Newsletter. 8 (2) August 2004 (p. 12 – 15)

Statistics Corner

Questions and answers about language testing statistics:

Test-taker motivations

James Dean Brown (University of Hawai'i at Manoa)

QUESTION: This isn't exactly a statistical question, but it does have to do with gathering test data for statistical analyses. At my university, we just did a study comparing the scores of students on our English language placement test with their scores two years later on the same test. Their average scores went down nearly a standard deviation after two years of studying English. Do you think they unlearned English somehow over the years, or is it possible they weren't as motivated to do well in the second test?

ANSWER: Both of your ideas provide possible explanations for your results: the average English level of your students may indeed have declined over two years, but it seems more likely to me that their motivations for taking the test may have changed. I have seen a number of instances over the years of students who have either had good reasons to score poorly on a test or have simply not been motivated to do well on a test. In either case, such factors can cause students to score lower than their actual abilities. In this column, I will address three sub-questions related to what you are asking: Why would some students purposely perform poorly on a test? What factors can muddy the interpretation of gain scores? And, what strategies can we use to counter such factors?

Why Would Some Students Purposely Perform Poorly on a Test?

Quite reasonably, when taking a test, students will do whatever they feel is in their best interests, and sometimes, it is not in their best interests (at least from their points of view) to perform well on a test. For example, when I was an undergraduate, I once took a French pronunciation course. On the first day of class, my teacher took us to the language lab and asked us to read a French poem into a tape recorder. Before starting, she told us that she was going to compare our pronunciation at the beginning and end of the course and count improvement as part of our grade. As a consequence, during the pretest, I consciously decided to read the poem with my best American accent (it must have hurt her ears). However, when it came time for the posttest, I put real effort into getting the pronunciation right. The story ends happily, at least from my point of view: I improved a great deal in my pronunciation and got an "A". Note: it possible that I could have improved equally well minutes later by simply doing a serious reading of the poem, without ever taking the course. Was I an evil student trying to mess up my teacher's test results? No, I was a cynical undergraduate (much like undergraduates everywhere) doing what I perceived to be in my best interests.

Another example, again taken from my family, is that of my son who took the Japanese language placement test at the University of Hawai'i at Manoa (UHM). He had lived in Japan for several years and studied Japanese for a number of years. Yet, he performed very poorly (below chance as I recall) on the UHM Japanese Placement Test. Why? His explanation was that he had to take two years of Japanese one way or another, and it would be easier for him if he took the lowest levels possible. Sure enough, he took four semesters of Japanese 101, 102, 201, and 202 during which he really didn't need to study at all.

What Factors Can Muddy the Interpretation of Gain Scores?

The case that you bring up in your question at the top of this article may be an example of how students' motivations can change between two different administrations of a test. If the students at your school would benefit from scoring well on the placement examination at the beginning of their studies, say by needing to study fewer years of English, they might well do their best on the placement test. However, when they take the test two years later, the purpose (to determine how much they had learned) may not be important to them. As a result they might have no personal reason to care if they do well on the test, or alternatively, at that point, they might simply be distracted by their busy lives, family problems, or their other course work. Any of the above factors could noticeably affect the average scores on that second administration, even if those factors only applied to some of the students.

Another possible explanation for your results is that you may not have had exactly the same students taking the pretest and posttest. Say, for instance, that the students who performed well on your placement test by scoring 90% or higher were exempted from further English study and thus did not ever take the posttest. When comparing the pretest scores with the posttest scores, it would be important to eliminate the data from those students who were exempted on the basis of their high pretest scores so that the pretest/posttest comparison would be comparing the same people. If these exempted students were inadvertently left in the analysis, the effect might be to minimize the pretest/posttest gains or even to create losses on average.

What Strategies Can We Use to Counter Such Factors?

The bottom line is that student motivations need to be taken into consideration when planning studies involving tests, and indeed in interpreting the scores of any test, especially in light of the observation that many students will act in what they consider to be their best interests. To avoid such effects, we must, in one way or another, insure that students are motivated to do well on each and every test they take. Otherwise the resulting scores may be a meaningless waste of everybody's time.

Consider the following example of one way to motivate students to perform to the best of their abilities. In one situation where I was teaching, if students did very well (defined as 80% or better) on the test at the beginning of the course, I would move them out of my class to a higher level. Importantly, in order to motivate them to do their best, I told the students about this before they took the test. Since these were students who were eager to finish their English language training and get on with their educations, I believe they did their very best on the pretest. Then, when those remaining in the course took the final examination, they knew that a good portion of their grades depended on their scores, so I am sure they did their best on the posttest as well. Only under such conditions, where it is in the best interest of students (from their point of view) to perform well, will comparisons between pretest and posttest scores be meaningful. Note also that only the scores of students who took both tests were compared in this situation.

In the case that you bring up in your question at the top of this article, it seems to be in the best interests of the students to perform well on the placement test, but they seem to have no reason to perform well on the posttest. You need to figure out a way to get the students motivated to do well on that posttest as well. To that end, you will need to change your school's policies if that is possible.

Perhaps you could tell the students that their performances on the posttest will be included as a component of their grades in their final course, or that a certain score (or score gain) must be achieved in order to finally be exempted from English training, or that the scores will be reported to their parents, etc. In short, if you want the results of that pretest/posttest comparison to have any meaning, you probably need to make some policy change that will insure that performing well on the posttest is in the best interests of the students. [For more on other testing policy issues, see Brown, 2004].

Consider another example, that of my son who purposely performed poorly on the Japanese language placement test. To address that and other language policy problems at UHM, we formed a committee that ultimately changed the policies such that it was in the best interest of students to perform well on the placement test. Now instead of having to take two years of a language starting from whatever level the students place into, they must reach a certain level (at least the equivalent of four semesters), while taking at least one course. In addition, they get credit for all courses in the sequence that they did not have to take. Thus, it would now be to my son's advantage to place high in the sequence. If for instance he had placed into the fourth course (Japanese 202), he could have received credit for four courses by taking only one. My guess is that, under these new policies, my son and other students like him would do their absolute best on the test because that behavior would be in their best interests.

I once was asked about a similar situation at a university in Taiwan. The teachers complained that their students were purposely doing poorly on the placement test. Those students had to take two years of English classes starting from whatever level they placed into. So students were purposely performing poorly. The solution that I suggested was that the students be required to finish two years of English, and if they placed high in that training or placed out of the training so much the better. Under such conditions, it would be in the best interests of most students to perform at their highest levels of ability.

Conclusion

Clearly then, researchers and other test score users must put themselves in the students' shoes and think about what possible motivations the students might have during a test administration as well as how those motivations might affect the test results. These issues clearly seem to me to be part of what Messick called *consequential validity* (Messick, 1988, 1989, 1996; also see Brown, 1996, Brown & Hudson, 2002). Since addressing changes in student proficiency before and after two years of instruction, as you are doing, has important potential policy consequences, you will want to do the comparison a well as you can. At minimum, doing so probably means adopting policies that will create conditions under which students will be motivated to score well on both administrations of the test and insuring that the comparison only includes students who took both administrations.

References

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall.

Brown, J. D. (2004). Grade inflation, standardized tests, and the case for on-campus language testing. In D. Douglas (Ed.), *English language testing in U.S. colleges and universities* (2nd ed.) (pp. 37-56). Washington, DC: NAFSA.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language testing, 13*, 241-256.

HTML: http://jalt.org/test/bro_20.htm / PDF: http://jalt.org/test/PDF/Brown20.pdf