Construct Validity of Foreign Language Tests

Khalid Addamigh (King Saud University, Saudi Arabia)

Phil Scholfield (The University of Essex, UK)

A standard view for construct validity is that a test is valid if it measures just the ability that it is expected to measure, i.e. what is often called the intended ‘trait’ or ‘construct’ of it. Now, given that the purpose of designing vocabulary test items in foreign language tests is to test whether or not the testees possess lexical knowledge, the validity question for us is: to what extent do the test items measure vocabulary? e.g. could test takers use some strategies to arrive at the correct response even if they do not possess the knowledge intended to be measured. If that is the case, construct validity of the test is harmed, as the performances behind processing the right answer are inconsistent with the theory behind the construct.

The scope for the current qualitative exploration was to discover the actual knowledge behind scoring on EFL vocabulary sections. To gain insight into test takers' conscious process when working on test items, we called upon a mental process-based instrument, introspection. That was based upon a triangulated data collection approach using verbal self-report methods. The test used in this study was made up of two subtests, professionally-made and teacher-made tests. The sample was 32 university EFL learners of two general proficiency levels, as determined by TOEFL, and two lexical proficiency groups, as identified by Nation's test.

The data collected shows that when the testees did not know the answer they always used some type of test-taking strategies (TTS) to bridge the gaps in their knowledge and to help them to arrive at what was likely to be the correct choice. Among the correct responses of the 36 test item cases studied in depth in this study, we find that only 60% were answered correctly based on knowledge retrieved from the subjects' mental lexical, which would be the answering mode expected by the test maker. However, 40% of the item cases answered correctly the test words were not known. These items were answered correctly because the test takers used strategies that enabled them to arrive at the correct alternatives. In a multiple-choice test format of four alternatives, approximately 25% of the correct answers could have been chosen by a random selection. However, the analysis shows that the test-takers correctly answered a far greater percentage than this. This gives clear evidence that those correct answers were rational choices and not simply based on the random chance

"Elimination using pragmatic meaning unlikelihood" was generally the most effective strategy. It helped the testees to arrive at the unknown correct answers by eliminating the distractors and generally led to greater success, when the correct response was unidentified. It also, in many cases, narrowed down the correct choice to fewer alternatives. Using TTS, however, does not always lead to a positive result. They led in two item cases to selections of incorrect answers, although the right responses were words actually known to the subjects. The strategy used in those two items was solely "using the L1 wording of a collocation". This approach was not successful probably because the L1 Arabic and the L2 English are totally distinct languages with different historical roots, so do not share great similarity in that linguistic feature.
Model framework for the present study
(as adapted from Biggs (1987))
Problems could occur at any stage of processing a test item. Ideally, when an individual attempts a multiple-choice gap-filling vocabulary test item, one might imagine the test-taker would follow the following steps:

**Stage 1:**
Understanding the stem

**Stage 2:**
Inferencing what meaning the gap needs

**Stage 3:**
Retrieving the alternatives' meanings

**Stage 4:**
Choosing an answer based on matching 1, 2 & 3 or other features
An example of the relevance of the test-taker’s perspective has been noticed by the present writer, from when he sat a real IELTS test. The introspection of the process gave evidence of invalidity, which think-aloud elicitation by a researcher would be likely to access, whereas other methods of validation could miss this evidence. In some of the test items of the listening comprehension section in that test, each of the four multiple-choice response options was about three quarters of a line in length. Since the time provided for the response was very short, the researcher noticed that answering such items, in fact, needed speed-reading competence as well as listening comprehension ability. In other words, some examinees may have comprehended the given listening material very well, but they may have been unable to give the right responses because the written items, which supposedly tested listening comprehension, actually measured a second skill at the same time.

Cohen (1984) mentioned findings that support this claim. He reported that students were given reading comprehension questions with 4 multiple-choice options, without the reading passage to which they were supposedly related. The random possibility of selecting a correct answer in such a situation is 25%. However, the subjects scored nearly twice that, clearly relying on some shortcoming in the design of the items which allowed knowledge other than that derived from comprehending the text to be used.