The C-test Construct Validity

Khalid Addamigh, PhD

King Saud University, Saudi Arabia

C-test

This form of testing was first introduced in 1981 by Klein-Braley and Raatz and has given rise to much dispute, which makes it a good case to look at in detail as an example of construct validation. The format of “a C-test demands exact word gap-filling by mutilating words in regular ways rather than by omitting words at regular intervals” (McDonough, 1995: 114). The standard procedure is that the second half of every second word is deleted and the first and last sentences are left intact. However, Grotijahn (personal communication, 2003) reports that this format is flexible, so if the proficiency level of the testees is low, the deletion could take place with every third word, or more, and can be performed on the last third of a word. He also states that the mutilation can also be made at the beginning or in the middle of a word if the language of the test has some restrictions on deleting certain parts of the words, e.g. because of a characteristic morphological system of prefixes or suffixes.

Construct validity

The term ‘construct’ was first introduced formally in 1955 by Cronbach & Meehl, who considered it as a hypothesised trait of people that is supposed to be reflected in test performance (Kunnan, 1995). Although ‘test construct validity’ is commonly used and investigated by researchers, it is still, as Alderson (2003) puts it, a theory-free concept. Scholfield (2000) also claims that it is misleadingly named, since the notion of a 'construct' is involved in all validity work, not just when one examines
'construct validity'. The concept of ‘construct validity’ is, therefore, not exact in the literature. The traditional view of construct validity is the narrow view that it is just one way of checking a test, as mentioned in the preceding section. However, this type of test validation is now seen by some experts, in a more comprehensive way, as the overarching term for validity checking in general (Messick, 1990), especially since “testing theories have changed and it is validity rather than reliability that is now considered to be of prime importance” (Banerjee and Clapham, 2003: 115). Brown (2002) views it as the extent to which a test measures the psychological construct that it is supposed to measure, identical with our definition of validity in general.

Similarly, Clapham (2003) claims that

The term 'construct validity' refers to the overall construct or trait being measured. It is an inclusive term which, according to some testing practitioners, covers all aspects of validity, and is therefore a synonym for 'validity'. If a test is supposed to be testing the construct of listening, it should indeed be testing listening, rather than reading, writing and/or memory.

Given the prevailing ‘modern’ view and following Palmer and Groot (1981), we will, therefore, regard the methods used to check construct validity as “a process of investigating what a test measures” (Palmer and Groot, 1981: 4).

Construct validity can be checked by means of one or more of a number of quantitative and qualitative approaches. Following Cronbach, Scholfield (2000) describes three types of conventional quantitative construct validation:

1- Correlational procedure, which concerns the degree to which a supposed test of a particular trait correlates with measures of other relevant variables with which theory would predict it would correlate. E.g. one would assume that a valid vocabulary test should correlate highly positively with a general
proficiency test. Hence we will check the correlation of scores on our test for a
positive correlation with TOEFL and Nation’s test, see Chapter VI for details.

2- Experimental procedure, where a researcher could give a test, to be validated,
to subjects before and after they are exposed to different conditions, which
should induce change, or s/he might use repeated measures via pre and post
tests with certain pedagogical treatments between, which theory assumes must
produce a change in scores. In our case this could have been performed
(though we did not do this) by giving the test to be validated to some subjects,
then teaching the test words, then giving the test again. Obviously scores
should go up, if the test is a valid test of the knowledge of the words.

3- Non-experimental procedure, such as when an EFL test is given to native
speakers and EFL learners, where one would expect adult, educated, native
speakers to achieve high scores. Jafarpur (1995) used this procedure to test the
construct validity of an EFL test and found that native speakers did not
achieve perfect scores. Hence, arguably, the test was not valid.

The key point for us, however, is that none of 1, 2 or 3 would provide insight into the
test-taker’s mind, to uncover directly what s/he was doing during the test-taking
process and so explain why the result could come as it did for Jafarpur, for example.

By contrast, some researchers recommend using qualitative procedures to validate
tests. Levenston et al. (n. d.) claim that using discourse analysis to analyse the nature
of the interdependence between the omitted lexical item and the given passage in a
cloze test, would allow a better understanding of the processes involved in filling in
the gaps. They maintain that discourse analysis can uncover “how the reader actually
processes the text to arrive at his completion in a cloze test” (Levenston et al., n. d., 203). However, to explore ‘how the test-taker actually processes the test’, analysis of verbal protocols produced by testees would be a direct approach that would be more likely to reflect the actual process, rather than depending on ‘inferring’ it. McDonough (1995) extends this way of validating a test by comparing the learners’ strategies revealed in the test situation with those used for the same intended trait in a non-test situation. Validity is reflected in the degree of similarity or difference in the strategies used in the two situations. In the current study, while we embraced the use of think-aloud verbal protocols to investigate MCFGV tests, we did not perform a parallel study in a non-test situation.

**Construct validity of the C-test**

Different studies have advanced different arguments about what the C-test measures resulting in considerable dispute among the professionals in language testing and the need for several validation studies. A number of experts maintain that C-tests are a valid measure of general language proficiency (Klein-Braley, 1985; Raatz 1985; Dornyei and Katona, 1992; Klein-Braley, 1997; Eckes and Grotjahn, 2003). The validation of the C-test in these studies depended mainly on the traditional correlational procedure (II.4.3.c) that checked the degree to which testees’ C-test’s scores correlated with their scores on a general language proficiency test that had already been validated. In addition to the claim that this form of testing measures general L2 proficiency, other features are also claimed:

There is no denying that it is an attractive choice for school and department that are responsible for assessing the overall proficiency of their students efficiently and cost-effectively. C-tests are a convenient option in resource poor environments for a number of reasons: they have a relatively simple design, they are low-cost in terms of developing, piloting, or photocopying, they are short and easy to administer and can be marked
quickly and efficiently by just a few raters. (Kontra and Kormos, 2003: …)

In contrast, other professionals doubt the claim that the C-test is a valid measure of general language proficiency. Alderson (2002), for example, firmly asserts that

The notion that there is a Holy Grail of language testing, a magic procedure which could produce universally valid measures of language ability, had, I thought, been finally laid to rest… However, I worry that we have not learned from history: we risk reviving alchemists’ claims of universal validity for another methods, this time the C-Test procedure. We should be on our guard against this danger. (Alderson, 2002: 15)

Stemmer (1991) empirically investigated the question: what does the C-test measure? There were three French C-test texts and four methods were used to collect and analyse the data:

1. a questionnaire to collect some background information about the subjects (e.g. age, sex, L1, L2 of parents).
2. discourse analysis of the three texts used in the test to infer the particular type of language knowledge involved in producing answers for the test items.
3. the actual outcome, i.e. the test performance on each item analysed for the numbers of items answered correctly, incorrectly or left blank.
4. think-aloud reports intended to collect data on the test-taking process from 30 subjects. The testees were 17-25 years old, studying French as a foreign language and their mother tongues were 8 to 9 different European languages. Before a testee did the test individually, s/he had listened to a short tape-recorded of a person thinking-aloud while doing a part of a C-test, as a training session for 4. One may argue, however, that this phase might have influenced the data collected, as the subjects may have mimicked that sample when they were tackling the study test; therefore, we did not do this in our
study. The think-aloud reports were audio-taped and analysed to gain insight into the subjects’ problem solving behaviour, but the researcher did not report the language in which the think-aloud reports were performed. This could hide a major shortcoming, since it is unlikely the researcher understood all the L1s of the testees, and their responses may have been limited in L2. She also did not report what motivation was of interest to testees for taking the test.

The study revealed that more function words were answered correctly than content words. Since function words contribute less to the meaning than do content words, this finding was claimed to indicate that the C-test operates more on a grammatical level than on advanced reading comprehension ability or general language proficiency. Hence, “the higher level comprehension skills are not involved in C-test solving” (Stemmer, 1991: 329).

Jafarpur (1995) found a similar result. In his study, 20 versions of an English C-test were distributed to 325 Iranian non-native speakers of English and 202 undergraduate English native speakers. Each version was taken by about 10 native and 16 non-native test-takers. A questionnaire was used to collect data about the face validity (II.4.3.a) of the C-test from EFL learner and instructors. Subjects also took a conventional cloze test so that their scores on the C-test and cloze test could be compared using Pearson product-moment correlation coefficients. Moreover, the mean scores of each group were subjected to analysis of variance and t-tests (II.4.5.3). The result indicated that the C-test suffers from a number of problems that basically cast doubt on its validity as a measure of general L2 proficiency. It was found that native speakers did not achieve perfect scores and C-tests did not possess face validity. “To sum up, the
results of this investigation indicated that C-testing does not achieve the claim made on its behalf [that it is a valid measure for general language proficiency]” (Jafarpur, 1995: 209).

In a recent unpublished study, Kontra and Kormos (2003) investigated the construct validity of C-tests using think-aloud. The aim was to see the actual activity involved in answering the test items. The participants were 10, 1 male and 9 female, EFL university Hungarian students aged 19-22. Their English proficiency was upper-intermediate and advanced. A short training in thinking aloud was given to each participant, but the researcher did not provide details about the content of this session. The subjects then were asked to take a C-test with consisting of three texts, with 20 gaps each. They were asked to think aloud while they were doing the test. The language in which the participants commented on their thought-processes was up to them, but most of them used Hungarian. They performed the test individually, in the presence of one of the researchers, which could be considered as an unusual situation for taking a test. The verbal data was tape-recorded, transcribed and analysed, but there were no much details about these phases. The researcher conclude that conclude that in the investigated setting the C-test is a valid measure of a number of components of foreign language competence and that with the appropriate methods in the analysis of the results, it can be used reliably to test the proficiency of upper-intermediate and advanced students of English (Kontra and Kormos, 2003: ...)

All in all, despite its short history, the C-test has received great deal of serious attention regarding its construct validity, about which there is still no consensus. What is important for us is the light the debate throws on the varied ways in which validation can be attempted, as several dissimilar approaches have been used to check the construct validity of the C-test. Since we are concerned with the strategies used in
a test situation, the approach of most interest is the introspective method, which we have seen was not always used (Klein-Braley, 1997; Jafar, 1995; Eckes and Grotjahn, 2003). This method of data collection will be discussed in detail in II.8.