Paradowski, Michał B. (2002) In: The Teacher 1: 40.

THE FEATURES OF GOOD LANGUAGE TESTS

Most of the progress and achievement tests that we as teachers run among our students we have to design ourselves. This is done in order to ensure that we are indeed testing what we intend to test, what we have taught, and what our students need to know. Tests found in teacher's books may of course be used as well, but they may be inappropriate or require adjustment to suit our particular purpose. There is also the risk that some students will procure the test and answer key beforehand, thus making the entire enterprise rather pointless. Moreover, several coursebooks lack accompanying testing materials, and after all teachers hardly restrict themselves to textbook use only. Thus the necessity of the self-design of progress and achievement tests seems obvious and well founded. Yet one test does not equal another, and it is therefore expedient to observe the following universally recognized criteria:

- **practicality**

  This criterion comes into concern in mass testing, when there is a need for evaluating the progress of several learners. Examining the practicality of the test planned involves, among others, looking at:
  - the preparation necessary to design the test (basically how long it is expected to take)
  - the administering of the test proper (arrangement of seating, distribution of the test among the learners, supervision, necessary equipment, timing, etc.; for instance, oral interviewing of several students will be time-consuming; moreover, the students taking the exam later may be better prepared knowing the questions)
  - scoring (marking e.g. essays, translation or dictation pieces may be a time-consuming and therefore inefficient process)
- **validity**

  This may denote both *face validity*, i.e. the way laymen (the learners, their parents, etc.) will appraise the test, whether the test appears to test what it is supposed to test, and *content validity*, i.e. the question whether the test reflects the content of the syllabus and whether it really measures what it is supposed to measure, and nothing else (e.g. summarizing a text heard from tape not only checks writing, but also listening comprehension and the ability to select, extract, and condense the most essential information; general knowledge, intelligence-testing and culturally-loaded questions do not test linguistic competence but extralinguistic knowledge or the analytical skills of the testee).

  We can also distinguish between *concurrent validity*, related to testing the learner's current command of the language, and *predictive validity*, i.e. assessing how well the learner will perform in future tasks, basing on his/her current level of linguistic attainment. Validity also implies that the tasks should be as realistic as possible and closely related to the situations in which the examinees will perform in real life.
- **reliability**

  In other words, the consistency and credibility of measurement; ensuring that the results of the test are not incidental. A differentiation is usually made between *test reliability* – whether the test measures language consistently (short tests are considered less reliable than ones covering a representative sample of the material taught and with a variety of testing formats), and *marker* (*examiner/rate/scorer*) *reliability* (closely connected with *objectivity*). The latter is low for testing speaking (the individual subjective tastes, norms and criteria of assessment of two examiners may differ considerably – the *inter-marker reliability* problem; moreover, the examiner's assessment may be affected by the speaker's physical appearance or other personal preferences) and writing (the norms assumed by different examiners may again be incongruous, and the final score may also be affected by the place the test occupies in the pile: an average composition marked directly after a good one will probably be given a lower mark than if it came after a poor one; moreover, a growing fatigue of the examiner may result in an increasing irritation or, quite the contrary, in a growing leniency – the *intra-marker reliability* problem). Marker reliability may be enhanced by increasing the number of examiners on the panel (in the case of oral interviews), developing a set of specified analytic criteria and standards instead of holistic (impressionistic) assessment, or grouping written works according to proficiency level prior to giving marks.

Practicality and reliability are particularly significant in norm-referenced placement and proficiency tests, whereas in criterion-referenced testing the most prominent role is given to validity. The testing technique that meets all the criteria above is the multiple-choice format, but its use is restricted to measuring receptive skills only. Tests of speaking and writing, on the other hand, although having high validity, may pose reliability (and occasionally also practicality) problems.

Beside bearing in mind the criteria discussed above, a test designer must also specify the *scoring criteria* (determine what to give and subtract points for) and *scoring weight* (how much weight should be given to an item; the number of points increases with the degree of format openness, the production necessary and the difficulty level; the proportion of points allotted to the different parts of the test should also reflect the relative importance of the skills tested in the syllabus). The tendency in achievement tests is to set pass at a relatively high level, between 60 and 75% of the total number of points. Apart from that, it is essential to provide clear, i.e. comprehensible and unambiguous instructions (without demanding the knowledge of metalinguistic terms and concepts; resorting to instructions in the learner's native tongue is quite justifiable in this respect) and to ensure that the examples clarify and do not confuse. The format of the test also ought to reflect the format of the activities in the classroom, in order to ensure that the learners are familiar with the tasks and rubrics. These directions together with the aforementioned criteria form a tool, a checklist against which not only a progress or achievement test, but also a placement and proficiency test can be evaluated. Finally, before running the test in class, you can present it to your colleague, who will perhaps notice and point out some ambiguities you may have failed to observe. A good test thus designed may bring you one more benefit: the positive *washback* (also referred to in some sources as *backwash*) *effect* on the teaching programme.

Recommended further reading:

Cohen, A. 1994. *Assessing Language Ability in the Classroom.* New York: Heinle and Heinle.

Harris, M., and P. McCann. 1998. *Assessment.* Oxford: Macmillan Heinemann ELT.

Heaton, J.B. 1990. *Classroom Testing.* London: Longman.

Hughes, A. 1993. *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Sobolew, L., and R. Gozdawa-Gołębiowski. 1994. *English through Practice Tests. Poradnik dla ucznia i nauczyciela.* Warszawa: PWN.