IMPLEMENTING LANGUAGE ASSESSMENT PRINCIPLES IN AN ONLINE TESTING SYSTEM

Wojciech Malec
Institute of English Philology, John Paul II Catholic University of Lublin
Al. Racławickie 14, Lublin, Poland
malew@kul.pl

Abstract: *Apart from reducing the time needed for test scoring and analysis, a web-based testing system can support language instructors in developing effective language tests. In particular, online technology has the potential to assist them in making valid score interpretations and dependable mastery/non-mastery decisions. However, technology alone will not guarantee high-quality tests – these must be properly developed in accordance with the fundamental principles of language assessment. This article takes a look at the basic assessment principles and demonstrates how they can be implemented in an online testing system.*

Keywords: e-testing, web-based test development, assessment principles

INTRODUCTION

Rather than being used solely for measurement and evaluation, language testing (whether traditional or web-based) should inform learning and teaching. Yet for this to be the case, the quality of measurement instruments must be as high as we can possibly make it. In order to ensure high quality language tests, it is essential that their construction be in keeping with the underlying principles of language assessment. Using the WebClass platform as an example, this article focuses on the application of assessment principles to web-based language testing, and emphasizes the facilitating role of online technology in the entire process of test development.

1. WEBCLASS

WebClass (Malec 2012) is an online, database-driven e-learning and e-testing system implementing PHP/MySQL and JavaScript/AJAX technologies. Its features include administration, basic communication, content authoring, and assessment. The system is used in blended learning environments with students of English

philology at the John Paul II Catholic University of Lublin and the State School of Higher Professional Education in Zamość.

The e-testing component of the platform can be used as an independent module to construct, administer, and analyze online assessments. The key part of test construction is the writing and evaluation of test items, whose types at WebClass range from selected response (true/false, multiple choice, multiple-choice cloze, multiple correct, matching), through limited production (cloze, gap-filling, transformations, error correction, short answer), to extended production (composition). When the items are compiled into a test form, they can be assigned to registered users, who respond to them online and submit their answers to the database. The answers can then be marked either automatically (selected and limited production types) or by the tester (compositions). At any time throughout the course, for both the instructor and the students, tabulated reports (with grades) can be dynamically generated upon access. Such reports include all test results and (optionally) attendance and self-assessment scores, as well as any other (e.g. paper-and-pencil) assessment results, which may be manually entered into the system. Instructors also have the option of analyzing the tests statistically in order to find out whether they are consistent with the quantitative principles of measurement.

2. ASSESSMENT PRINCIPLES

This section takes a look at the basic principles of language assessment (as identified and discussed by Brown 2004), including practicality, reliability, validity, authenticity, and washback. These principles can be used as criteria for evaluating the usefulness of an existing assessment instrument by means of qualitative and quantitative analyses.[1] More importantly, however, they permeate and impact on the entire test development cycle: from test design, through operationalization, to administration (for more on stages of test development, see, e.g., Bachman and Palmer 1996; Hughes 2003; Fulcher 2010). The overall aim here is to demonstrate that an online testing system can greatly facilitate the application of (at least some of) these principles to language test development.

2.1 Practicality

A test is considered to be practical as long as it "is not excessively expensive, stays within appropriate time constraints, is relatively easy to administer, and has a scoring/evaluation procedure that is specific and time-efficient" (Brown 2004: 19). Online assessments have definitely more to offer in terms of practicality than do traditional paper-and-pencil tests.

First, tests at WebClass do not always have to be constructed by writing new items from scratch: existing (previous) items can be imported from a bank, or from

[1] Qualitative analysis involves making subjective judgments about the effectiveness of items or tasks, and relies to a large extent on common sense. Quantitative analysis, on the other hand, consists in calculating and interpreting item and test statistics.

another test. It is also possible to speed up the process of item writing by editing the questions in an HTML editor and then converting them to test items proper using a text-to-items converter. A test creator wizard is also available, which makes the process of test construction fully automated – the wizard retrieves random items from one or more item banks and puts them in a user-defined number of item sets.

Moreover, web-based tests score highly in the area of easy delivery and scoring efficiency. It is a fact long recognized that a web browser is everything that students need in order to take online tests and quizzes "whenever and wherever it is convenient" (Roever, 2001: 88). On the instructor's side, test practicality is significantly enhanced by the possibility of using automated scoring procedures. At WebClass, in the case of limited production items, it is not only the (perfectly) correct responses that are given credit automatically – the test constructor can also specify other acceptable (e.g. partially correct) answers, as seen in Figure 1:

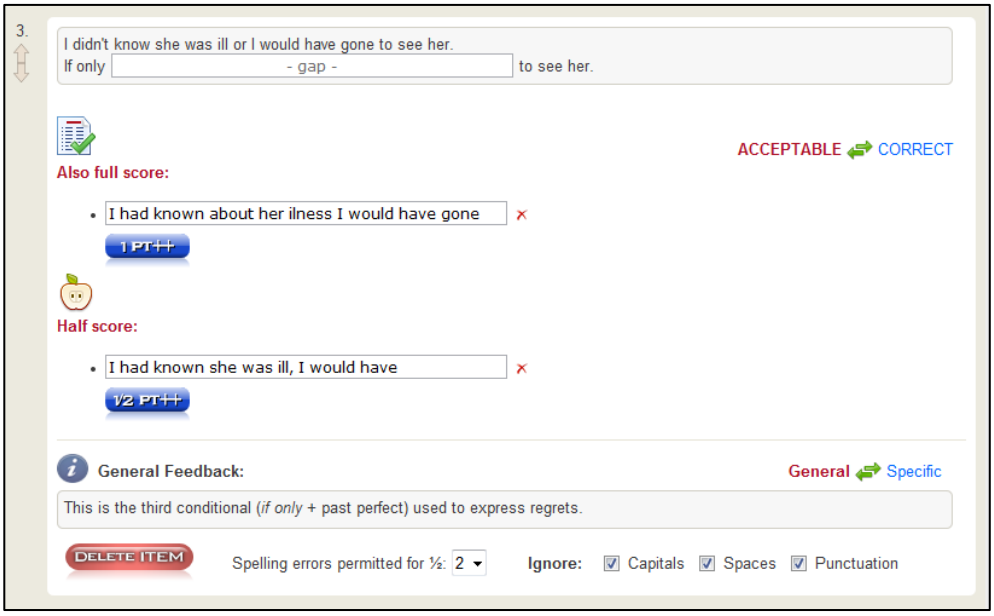

Figure 1. WebClass test editor showing part of a transformation item
Source: http://webclass.co/tests_edit/?test_id=1395

The figure illustrates a transformation item (taken from a grammar test) eliciting the completion of a paraphrase of the sentence *I didn't know she was ill or I would have gone to see her*. The expected response is *I had known she was ill, I would have gone* (not displayed in the figure). However, one student supplied *I had known about her ilness I would have gone*, which is not exactly the expected response and, additionally, one word is misspelt. Nevertheless, the instructor decided that, since the grammatical structure was fine, this particular answer merited full credit. Such a change to the key can be made with a single click of a button after the test has already been administered, while verifying the automatic marking. When the test is

opened in the editor, as in Figure 1, the acceptable answer is displayed under 'Also full score'. Similarly, answers which merit partial credit (half a mark) can be added to the scoring rubric either at the construction or at the verification stage of test development. Half a mark can also be awarded automatically for 'imperfect' responses as long as they do not contain more misspelt characters than the number selected by the test constructor (error level for ½), which equals 2 in the case of the item in Figure 1. In addition to this, the scoring algorithm can be set to completely ignore capital letters, spaces, and punctuation.

As far as extended responses are concerned, these have to be marked manually at WebClass in a window containing a simple HTML editor, where the tester can mark errors and insert comments. The features which can potentially contribute to test practicality include the possibility of looking up mouse-selected text in the BNC (British National Corpus) or in Google Books (useful for checking collocations) as well as a plagiarism detection script (see Malec, forthcoming, for more on this).

Finally, online testing constitutes a radical change in the way quantitative analysis can be carried out by school teachers. This type of test analysis requires individual item scores (ones, zeroes, half points, etc.), yet it is hard to imagine overworked teachers entering hundreds of values into, e.g., Excel spreadsheets on a daily basis. At WebClass, a number of statistics are calculated automatically (see below). Additionally, test scores can be downloaded as comma-separated variables (CSV files) for further analysis.

2.2 Reliability

In language testing, the higher the degree to which test scores are free from errors of measurement, "the greater the relative effect of the language abilities we want to measure, and hence, the reliability of language test scores" (Bachman 1990: 160). In other words, we want individual variability in test performance to be due to actual changes in ability rather than differences in test tasks, administrative procedures, raters, or factors such as mood, fatigue, lack of interest, etc. Generally speaking, reliability pertains to "consistency of pupil performance and consistency in assessing that performance" (Gipps 1994: 67).

In online testing, thanks to automated scoring procedures, the potential source of unreliability stemming from the fact the humans can make errors when marking tests (cf. Fulcher 2010: 46) is eliminated. The scoring algorithms make sure that all of the students who submit identical answers get identical scores. At WebClass, rater reliability is further ensured at the verification stage (see also the previous section), as illustrated in Figure 2:

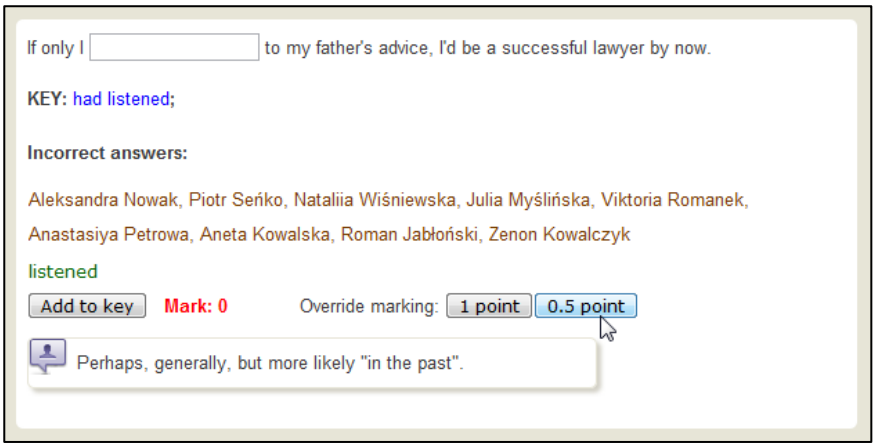

Figure 2. Verification of the marking
Source: http://webclass.co/tests_verify/?test_id=1395

In the test verification window, for each test item, incorrect answers are displayed together with all of the test takers who submitted them. If the tester overrides the automatic score (by clicking '1 point' or '0.5 point'), the change applies to all of the relevant test takers.

Reliability statistics are calculated at WebClass when a test analysis window is opened. An example is given in Figure 3:

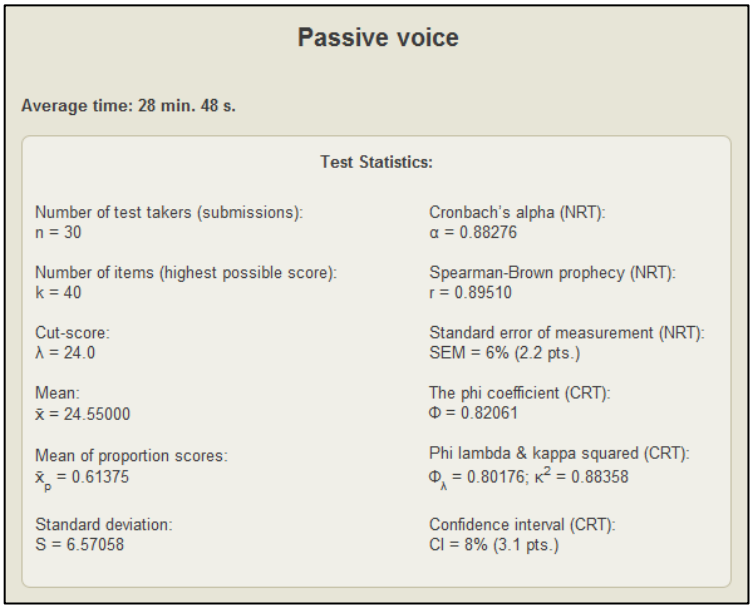

Figure 3. Test statistics
Source: http://webclass.co/tests_stats/analysis.php?test_id=1361

The statistics, which include both norm-referenced and criterion-referenced reliability estimates, can help the test developer to evaluate the quality of the assessment instrument. For example, the values of phi lambda and kappa squared indicate the degree to which mastery/non-mastery decisions made on the basis of test scores are dependable (for more on these and other statistics, their calculation and interpretation, as well as differences between norm-referenced and criterion-referenced testing, see Brown and Hudson 2002; Bachman 2004).

2.3 Validity

Reliability and validity should be seen as "complementary aspects of a common concern in measurement – identifying, estimating, and controlling the effects of factors that affect test scores" (Bachman 1990: 160). Alderson *et al.* (1995) explain the relationship between the two concepts in the following way:

> In principle, a test cannot be valid unless it is reliable. If a test does not measure something consistently, it follows that it cannot always be measuring it accurately. On the other hand, it is quite possible for a test to be reliable but invalid. A test can, for example, consistently give the same results, although it is not measuring what it is supposed to. Therefore, although reliability is needed for validity, it alone is not sufficient. (Alderson *et al.*, 1995: 187)

High reliability of a test is not an end in itself, and calculating it should be regarded as "part of a unified approach to establishing the overall validity of a test" (Weir, 2005: 43). It is now widely recognized that certain aspects of reliability and validity may be virtually indistinguishable from each other, as in the case of parallel forms reliability and concurrent validity (Alderson *et al.* 1995: 188). In this sense, it does not ultimately matter whether the evidence that is produced in support of the usefulness of a test pertains to reliability or validity, although the distinction between the two concepts is usually maintained simply for clarity and neatness of presentation. What does matter is that assessment instruments should be inspected from many different angles in order to make sure that the scores they produce are as accurate as possible a reflection of test takers' knowledge and ability. Obviously, no test is perfect, and test scores will always contain some error. Tests may also vary on certain aspects of validity from one administration to another, or from one group of test takers to another. In short, validity is a matter of degree and a relative rather than absolute concept (Messick 1989; Weir 2005).

In classroom testing, and arguably in most distance-learning contexts, the major sources of evidence for test validity are content validity and construct validity (cf. Brown and Hudson 2002: 212; Brown 2004: 32). In short, the former refers to the extent to which test items are representative of the course content and learning objectives, whereas the latter relates to the degree to which the test actually taps into the psychological construct that it purports to be measuring. Evidence for content validity is sought through qualitative analysis, while evidence for construct validity may come from correlational approaches such as the multitrait-multimethod matrix

(Campbell and Fiske 1959) and factor analysis, as well as from intervention and differential-groups studies (Brown and Hudson 2002).

As mentioned above, the testing system at WebClass offers the possibility of exporting test scores to a .csv file, from which they can be easily transferred to a statistical software package such as SPSS for advanced analyses. However, basic item analysis is carried out in the same window which displays reliability estimates. It includes one statistic that is particularly useful for detecting items which might not measure what they are supposed to measure. Consider the results of item analysis given in Figure 4:

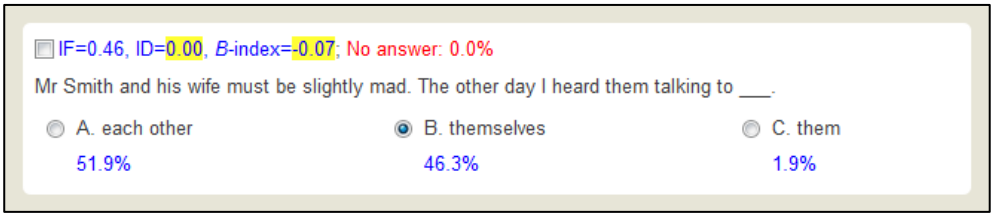

Figure 4. Multiple-choice item analysis
Source: http://webclass.co/tests_stats/analysis.php?test_id=171

The *B*-index is a simple statistic which tells us how successful a given item was at distinguishing masters (students who passed the test) from non-masters (those who failed it). A negative value of this statistic indicates that the item analyzed was more difficult for masters than for non-masters, contrary to what we would actually expect. On closer inspection, it turns out that the item in Figure 1 does not measure grammatical knowledge only, because option A is grammatically correct, even if it does not make sense. It follows, then, that the item taps logical thinking, in addition to grammatical knowledge. Since the test as a whole was supposed to measure knowledge of lexico-grammatical structures, the item in question was perhaps a bit of a misfit.

2.4 Authenticity

Authentic test tasks are those which are similar to some tasks in the target language use domain (real world). According to Brown (2004), test authenticity can be increased by using natural language, contextualized items, meaningful topics, by connecting items through a story line, and by making sure that they simulate real-world tasks. In addition, they can be made more authentic through the provision of multimedia (Chapelle and Douglas 2006).

Certain web-based methods of assessment are arguably more authentic than their paper-and-pencil counterparts. A case in point is the assessment of writing ability. One of the things that the test developer has to consider is the fact that these days hardly any text is hand-written. Official documents, for example, are either submitted using online forms, email, etc. or typed on the computer and printed out. Indeed, written communication is nowadays mostly computer-mediated (emails, social networking services, chat rooms, etc.). It follows then that test tasks requiring

students to enter the text on the computer are somewhat more authentic than analogous paper-and-pencil ones. By the same token, web-based assessments of writing ability (consisting of, e.g., tasks eliciting an extended response, such as a letter) satisfy the criterion of authenticity by virtue of simulating real-world tasks.

2.5 Washback

Washback in language testing generally refers to the influence of testing on learning and teaching (e.g. Bailey 1996). The common belief that "good test scores equal good education" (Popham, 2001: 16) often gives rise to 'teaching to the test', i.e. improving test results instead of diagnosing learners' strengths and weaknesses. Brown (2004) points out that "[o]ne way to enhance [positive] washback is to comment generously and specifically on test performance" (p. 29).

Online testing readily lends itself to the requirements of assessment *for* learning, defined by Black *et al.* (2004: 10) as "any assessment for which the first priority in its design and practice is to serve the purpose of promoting students' learning" (see also, e.g., Chappuis and Stiggins 2002; Cauley and McMillan 2010; Wiliam 2011). One of the key components of such assessment is the provision of effective feedback on students' progress towards the learning objectives. Using the WebClass testing system, two types of feedback can be provided: general and response-specific. General feedback contains some extra information on the content of the test item; it is always provided irrespective of the response. Response-specific feedback, on the other hand, is the comment that students receive when their response matches the one for which the feedback is intended. Test takers can view and study the feedback immediately upon completion of the tasks, or later as a homework assignment.

Brown (2004: 37) suggests that beneficial washback can also be achieved by encouraging students to set their own learning goals and engage in self-assessment. Both of these practices are formalized at WebClass: students can define their own learning objectives (in addition to those set by the instructor) and then evaluate their learning outcomes (by indicating how well they have done on each objective) on a predefined scale. The results of student self-assessment can optionally be included in the calculation of the final semester grade.

CONCLUSION

This article has presented a short review of the five principles of language assessment outlined by Brown (2004). Of all the principles in question, discussed here in the context of web-based testing, validity is arguably the most fundamental one: test developers need to make sure, first and foremost, that test scores are accurate indicators of whatever the test claims to be testing. Guidelines on achieving this apply in equal measure to both online and paper-and-pencil testing:

> For validity to be high, the assessments should sample students' performance on each objective, an appropriate mix of assessment methods should be used, and

> assessment methods should be selected on the basis of providing the truest picture possible. (Morgan and O'Reilly, 2006: 95)

Conceptually, web-based testing as exemplified by the WebClass system does not constitute a major departure from traditional paper-and-pencil testing: online test development needs to follow the same well-known principles of assessment. The difference that web-based technology really makes is in the area of practicality: time saving (with respect to test construction and quantitative analysis), easy delivery, and scoring efficiency.

REFERENCES

Alderson, J., C., Clapham, C., and Wall, D., 1995: *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press, ISBN: 9780521478298.

Bachman, L., F., 1990: *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press, ISBN: 9780194370035.

Bachman, L., F., 2004: *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press, ISBN: 9780521003285.

Bachman, L., F., and Palmer, A., S., 1996: *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press, ISBN: 9780194371483.

Bailey, K., M., 1996: Working for washback: a review of the washback concept in language testing. *Language Testing* 13, pp. 257–279, ISSN: 02655322.

Black, P., Harrison, C., Lee, C., Marshall, B., and Wiliam, D., 2004: Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan* 86(1), pp. 9-21, ISSN: 00317217.

Brown, H., D., 2004: *Language Assessment: Principles and Classroom Practices*. White Plains, NY: Pearson Education., ISBN: 9780130988348

Brown, J., D., and Hudson, T., 2002: *Criterion-referenced Language Testing*. Cambridge: Cambridge University Press, ISBN: 9780521000833.

Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56, 81–105, ISSN: 00332909.

Cauley, K., M., and McMillan, K., H., 2010: Formative assessment techniques to support student motivation and achievement. *The Clearing House* 83(1), pp. 1-6, ISSN: 00098655.

Chapelle, C., A., and Douglas, D., 2006: *Assessing Language through Computer Technology*. Cambridge: Cambridge University Press, ISBN: 9780521840217.

Chappuis, S., and Stiggins, R., J., 2002: Classroom assessment for learning. *Educational Leadership* 60(1), pp. 40-43, ISSN: 00131784.

Fulcher, G., 2010: *Practical Language Testing*. London: Hodder Education, ISBN: 9780340984482.

Gipps, V., C., 1994: *Beyond Testing: Towards a Theory of Educational Assessment*. London: Falmer Press, ISBN 9780750703291.

Hughes, A., 2003: *Testing for Language Teachers* (2nd edition). Cambridge: Cambridge University Press, ISBN: 9780521823258.

Malec, W., 2012: *WebClass* [online learning management system]. Available at http://webclass.co (Accessed on 15 June 2013).

Malec, W., (forthcoming). On the potential of web-based assessment of language skills. In H. Chodkiewicz and M. Trepczyńska (eds.), *Language Skills: Traditions, Transitions and Ways Forward*. Newcastle upon Tyne, UK: Cambridge Scholars Publishing, ISBN: 9781443853187.

Messick, S., 1989: Validity. In R. L. Linn (ed.), *Educational Measurement* (3rd edition, pp. 13–103). New York: Macmillan, ISBN: 9780029224007.

Morgan, C., and O'Reilly, M., 2006: Ten key qualities of assessment online. In M. Hricko and S. L. Howell (eds.), *Online Assessment and Measurement: Foundations and Challenges* (pp. 86-101). Hershey: Information Science Publishing, ISBN 9781591404989.

Popham, W., J., 2001: *The Truth About Testing: An Educator's Call to Action*. Alexandria, VA: Association for Supervision and Curriculum Development, ISBN: 9780871205230.

Roever, C., 2001: Web-based language testing. *Language Learning & Technology* 5(2), pp. 84-94, ISSN: 10943501.

Weir, C., J., 2005: *Language Testing and Validation: An Evidence-based Approach*. Basingstoke: Palgrave Macmillan, ISBN: 9781403911896.

Wiliam, D., 2011: What is assessment for learning? *Studies in Educational Evaluation* 37(1), pp. 3-14, ISSN: 0191491X.