

Test review

The Michigan English language assessment battery (MELAB)

Purpose: To assess English language ability of adults applying to North American colleges and universities or professionals who need to use English in their work.

Price: \$60 plus \$15 for oral interview

Author and Publisher: English Language Institute, Testing and Certification, MELAB Testing Program, University of Michigan, Ann Arbor, MI, USA

Contact information: English Language Institute, Testing and Certification, MELAB Testing Program, 3020 North University Building, University of Michigan, Ann Arbor, MI 48109-1057, USA; phone: +1 734 764 2416/763 3452; fax: +1 734 763 0369; email: melabelium@umich.edu; websites: <http://www.lsa.umich.edu/eli/testing.htm>; <http://www.lsa.umich.edu/eli/policychange.htm>

I Introduction

The Michigan English Language Assessment Battery (MELAB) is a standardized test of English proficiency whose purpose is 'to evaluate the advanced level English competence of adult non-native speakers of English' (MELAB technical manual, 1996: 1). The MELAB is designed for adults applying to North American colleges and universities or for professionals who need to use English in their work. It is accepted by many institutions in lieu of the TOEFL exam.

II Description

The MELAB has three sections. Part 1 is a 30-minute impromptu essay on the test taker's choice between two topics. New topic sets are introduced annually. The essays are scored by at least two trained raters on a locally developed 10-step holistic scale. The writing scale is set at nearly equal intervals between 50 and 100 (53 and 97, to be exact). It was originally designed to conform to the equated listening

and GCVR (grammar, cloze, vocabulary and reading comprehension) scales, so that the three sections of the exam are on the same scale and can therefore be averaged to determine the final score. The rating scale descriptors focus on topic development, organization, and range, accuracy and appropriateness of grammar and vocabulary.

Part 2 of the MELAB is a 50-item listening test delivered via audiotape. The listening test contains 25 to 35 discrete items based on questions, statements or short dialogs, and two to three extended listening texts (e.g., radio broadcasts, lectures or conversations of 3–5 minutes), each followed by several questions. Note taking is allowed for the extended texts. The listening test takes about 25 minutes to administer. All listening items are multiple choice with three options, and the reported score for listening is scaled from 30 and 100. For both Part 2 and Part 3, scaled scores for different forms of the test are based on normative information (primarily percentile rank) from approximately 100 examinees who take two forms of the test.

Part 3 of the MELAB, called the GCVR, includes sections on grammar (30 items in a two-turn conversational format), cloze (20 items taken from one passage), vocabulary (30 single-sentence items) and reading comprehension (four short, unrelated passages followed by five items each, for a total of 20 items). Test takers have 75 minutes to complete the GCVR. The reported score is scaled from 15 to 100. At any given time, three to four alternate forms of Parts 2 and 3 are in use.

The final MELAB score is calculated as the average of the three parts and reported as a scaled score between 33 and 99. In addition, the MELAB includes an optional 10–15 minute speaking test (oral interview). The speaking test is scored from 1 to 4 on a holistic scale and is reported separately from the main MELAB scores.

Total testing time for the MELAB is $2\frac{1}{2}$ to $3\frac{1}{2}$ hours, depending on whether the oral interview (speaking test) is administered. The MELAB is administered as an individual or group test by approximately 100 authorized examiners in the USA and Canada. All tests are sent to the English Language Institute, University of Michigan (ELI-UM) for scoring, with the exception of the speaking section, which is scored locally. Outside the USA and Canada, there is limited availability of the MELAB. Outside the USA and Canada, the MELAB is available only as a sponsored group test arranged by ELI-UM, and the Speaking section is not available outside the USA and Canada.

III Strengths and weaknesses

The MELAB appears to be a carefully designed, reliable test that provides useful information about candidates' language proficiency.

While the listening and GCVR parts of the test are susceptible to some of the criticisms of discrete-point testing (see below), these are balanced by the composition test and the oral interview, which provide a more comprehensive picture of a test-taker's interaction skills and ability to produce extended written and oral discourse in English.

The trade-off between the various aspects of test usefulness (Bachman and Palmer, 1996) is apparent in the listening and GCVR sections of the MELAB. Both parts of the test are highly reliable, with reliability coefficients (KR-21 and Cronbach's alpha) ranging from .82 to .95. Furthermore, the test authors have gone to great pains to demonstrate various aspects of validity in the technical manual. For each section of the test, the authors provide content-related evidence of validity, describing the nature of the skill that the test is intended to measure, the process of test development and a thorough description of the prompts and item types. Construct-related evidence for validity presented in the manual includes a consideration of the item types on the MELAB in relation to a theory of communicative language ability (Bachman, 1990; Bachman & Palmer, 1996); factor analysis and native-speaker performance, although this last category may be more appropriately considered as criterion-related. Criterion-related evidence of validity includes comparison of MELAB scores with productive tests of language (the MELAB composition and speaking test), comparison of MELAB scores with the TOEFL and comparison of MELAB scores with teacher assessments of students' proficiency.

On the other hand, the very features of the test that lead to high reliability and contribute to the authors' considerations for validity are those that have led to criticisms of discrete-point testing, particularly in terms of what Bachman and Palmer (1996) call authenticity and interactivity, and in terms of washback, or the effect of the test on instruction. Bachman and Palmer define authenticity as 'the degree of correspondence of the characteristics of a given language test task to the features of a T[arget] L[anguage] U[se] task' (p. 23) and interactivity as 'the extent and type of involvement of the test taker's individual characteristics [specifically, language knowledge, strategic competence, topical knowledge and affective schemata] in accomplishing test tasks' (p. 25). Given these definitions, the test tasks of the listening and GCVR are quite limited in terms of both authenticity and interactivity. For example, choosing the most grammatical way to complete 30 unrelated sentences is a task that is highly inauthentic outside of a language test. The same can be said of listening to a series of unrelated, decontextualized questions and statements, even if the micro-skills involved in these tasks may be relevant for academic writing and listening. Similarly, these test tasks

are low in interactiveness, as they involve very limited aspects of language knowledge and strategic competence.

The composition test is particularly noteworthy for its scoring system, which combines the efficiency of a holistic rating scale with procedures that take into consideration the multi-faceted nature of second language (L2) writing. While the reported score is a single number, raters may, if they choose, add codes to the numerical score indicating that one or more aspect of the composition (e.g., topic development, syntax or vocabulary) is particularly strong or weak. This system allows raters to acknowledge strengths and weaknesses within the writing that can be useful both for decision makers and for test takers.

A weakness of the composition test is that it consists of a single writing task, which limits the generalizability of the results and may disadvantage test takers who happen to have little interest or background in the two assigned topics. However, the trade-off between adequate sampling of the domain of academic writing and concerns of feasibility and practicality remains one of the great unsolved – and possibly insoluble – dilemmas in writing assessment (for discussions of this issue, see, for example, Hamp-Lyons, 1991; Camp, 1993). The authors of the MELAB can hardly be faulted for not having found a way around this dilemma. Nevertheless, the results of the composition test should be viewed with caution, in light of the fact that the score is based on a single writing sample. It should be noted that the prompts themselves go through a rigorous pre-testing and evaluation procedure before they are used operationally.

Perhaps a more serious limitation of this type of testing is its impact on instruction and curricula. While the MELAB technical manual stresses that active use of the language is the best way to prepare for a proficiency test such as the MELAB, the fact that so many of the items are based on recognition of or application of grammatical rules rather than authentic communication makes it likely that instruction will focus on linguistic accuracy rather than, for example, pragmatic appropriateness. As Bardovi-Harlig and Dornyei (1998) point out, for many L2 learners there is an imbalance between their knowledge of grammar and their knowledge of pragmatics. They discuss the potential role of language tests that privilege micro-level accuracy over macro-level appropriateness as one reason for this imbalance, as learners tend to be motivated to learn the skills that they will be tested on. The MELAB claims to assess pragmatic knowledge in most sections of the test, yet it was difficult to identify items in any section of the listening or GCVR test that seemed to tap knowledge of appropriateness outside of the use of certain idioms or collocations.

Similarly, the use of short reading passages (under 300 words) with

multiple-choice questions may encourage reading strategies that are not always applicable to academic reading in practice. Most of the items depend on comprehension of a single sentence within the passage rather than the passage as a whole. As Bernhardt (1991: 193) notes, many L2 readers are able to deal with 'units of language as separate entities' but this does not necessarily lead to comprehension of a coherent message within a text. Focusing on close, careful reading of short, unrelated passages to find the answers to specific detailed questions may not prepare students to cope with large amounts of reading in academic courses, to consider their own or the author's purposes in reading, or to integrate what they read with their background knowledge.

IV Fairness

A growing concern in language assessment is the issue of equity, or fairness, to examinees in terms of such issues as content familiarity, bias in favour of particular groups of examinees, and access. On this score the MELAB generally fares well. While the MELAB technical manual does not explicitly address the issue of test bias or differential item functioning, it does present comparative statistics for examinees grouped by reason for testing, sex, age and native-language groups. In terms of test content, the test writers have attempted to include content on a wide variety of subjects that would appeal to many different kinds of examinee, thus minimizing the risk that some examinees would be advantaged or disadvantaged by unequal content knowledge. Finally, the MELAB is considerably less expensive than the TOEFL, putting the test within reach of a broader spectrum of potential examinees. My only fairness-related quibble with the MELAB is the use of language some might consider sexist in the test instructions: in one sample multiple-choice question, the options include 'a married man' and 'a married lady': why not use the more neutral term 'woman'? Apart from this lapse, however, the test developers seem to have addressed fairness issues reasonably well.

V Documentation

Decisions about test use are made easier by thorough and clear documentation, and this is in fact one of the great strengths of the MELAB. Great care has been taken to communicate important information to the test-taker, the administrator and the test-user. This care is evident in all of the documentation related to the MELAB, from the score report form to the administration manual to the technical manual. For example, the score report contains not only the students' scores, but

also a brief description of each part of the test, along with score ranges, means and standard deviations for each test part and for the final score. The standard error of measurement (SEM) is reported for the final score as well, with a note explaining the use of the SEM in making decisions about test takers. Furthermore, the score report also reprints the composition descriptions, so that test takers and test users alike are informed of the criteria used to judge the writing samples.

The technical manual for the MELAB is another testimony to the desire on the test authors' part to communicate information about the test clearly and completely to potential and current test users. The manual provides a wealth of information about the design, administration, scoring and interpretation of the test and is written in a straightforward manner to be accessible to an audience with limited background and expertise in testing. The section on interpreting MELAB scores is particularly useful for non-testing experts considering adoption the MELAB, as it provides concrete examples of how the MELAB is used in various settings.

The technical manual also includes considerable information about the test that is of interest to a more specialized audience. There is an extensive section on test statistics that provides a variety of statistical information about the test, and an equally thorough section devoted to discussions of reliability and validity, as mentioned above.

The administration manual and the manual used to train oral interview examiners are equally well written and thorough, anticipating potential problems and ensuring appropriate standardization of administration and the maintenance of test security. One particularly noteworthy feature of the manual for oral interviewers is the inclusion of audio-taped excerpts from successful and less successful interviews along with annotated transcripts. These excerpts help to raise interviewers' awareness of interactional features that may promote or hinder examinees' opportunities to express themselves fully in English.

VI Summary

In summary, the MELAB is a thoughtfully constructed, reliable and particularly well documented test. The test does not avoid the limitations common to many standardized tests; in particular, most of the listening and GCVR sections of the MELAB emphasize comprehension of short, unrelated texts and easily testable points of grammar and vocabulary at the expense of extended discourse and more interactive, communicative aspects of language use. This not only limits the authenticity of the test but may also have a negative effect on curricula and instruction, as learners may not be motivated to learn what they will not be tested on. Despite these limitations, however,

potential test users are given ample information to decide for themselves whether the many strengths of the test outweigh any disadvantages.

VII References

- Bachman, L.F.** 1990: *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L.F.** and **Palmer, A.S.** 1996: *Language testing in practice*. Oxford: Oxford University Press.
- Bardovi-Harlig, K.** and **Dornyei, Z.** 1998: Do language learners recognize pragmatic violations? Pragmatic versus grammatical awareness in instructed L2 learning. *TESOL Quarterly* 32, 233–262.
- Bernhardt, E.B.** 1991: *Reading development in a second language: theoretical, empirical, and classroom perspectives*. Norwood, NJ: Ablex.
- Camp, R.** 1993: Changing the model for the direct assessment of writing. In Williamson, M.B and Huot, B.A., editors, *Validating holistic scoring for writing assessment*. Cresskill, NJ: Hampton Press, pp. 45–78.
- Hamp-Lyons, L.** 1991: Basic concepts. In Hamp-Lyons, L., editor, *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex, pp. 5–15.
- MELAB technical manual** 1996: Ann Arbor, MI. University of Michigan Press.

Sara Cushing Weigle
Georgia State University