MICHIGAN LANGUAGE ASSESSMENT

Revising the MET Technical Report

Contact Information

All correspondence and mailings should be addressed to:

Michigan Language Assessment
Argus 1 building
535 West William St., Suite 310
Ann Arbor, Michigan
48103-4978 USA

T +1 866.696.3522
T+1 734.615.9629
F +1 734.763.0369

info@michiganassessment.org
MichiganAssessment.org

Table of Contents

List of Tables

List of Figures

1. Introduction

The Michigan English Test (MET) is a standardized, multilevel examination of general English language proficiency. Developed and produced by Michigan Language Assessment, the test covers the four language skills: listening, reading, speaking, and writing.

The listening and reading sections measure listening, reading, grammar, and vocabulary skills in educational, public, and occupational contexts, with recordings and reading passages that reflect interactions in an American-English linguistic environment. The speaking section measures an individual's ability to produce comprehensible speech in response to a range of tasks and topics, and the writing section measures an individual's ability to write in English in response to two different tasks.

This report summarizes the results of recent revisions made to the MET, providing details on what changes were made to each section, as well as describing how a new seat time was determined for the reading and grammar section and the steps taken to ensure exam reliability was maintained. The revision process occurred over a two year period, and the revised MET was implemented in January 2019.

2. Increasing Academic Focus

The MET has always had good coverage of language as it is used in educational settings, but since many MET users are specifically in educational contexts, the decision was made to increase the test's academic focus.

In the reading and grammar section, the existing thematic reading task (Macmillan, Chapman & Stucker, 2014) already requires test takers to make meaning across texts, which is important in the academic domain. It also covers all the cognitive operations involved in reading (Khalifa and Weir, 2009), as well as testing both the ability to read carefully (detail-oriented) and quickly (skimming). Nevertheless, a new item type, extended reading passages, was added to further enhance the contextual validity of the test. The revised reading and grammar section contains two of these passages, which cover a range of topics that might be encountered in educational contexts, each of which are followed by five questions. Additionally, the proportion of thematic reading passages and grammar questions situated in the educational domain were also increased for the reading and grammar section.

Similarly, for the listening, writing, and speaking sections, the existing tasks already had good coverage of the educational domain. Research has shown that speaking and listening are more similar across academic and general contexts compared to reading and writing (Biber, Conrad, Reppen, Byrd, & Helt, 2002), so less revision was required for these sections. While no new item types were added to the listening, writing, or speaking sections, the proportion of conversations situated in the academic domain were increased for the listening section, and at least one writing task and two speaking parts now also focus on academic domain topics.

3. Reducing Test Length

Another focus of the MET revision project was to reduce the overall length of the listening and reading sections. Michigan Language Assessment had received feedback from MET stakeholders that the number of items and duration of the test were onerous, so work was undertaken to develop a shortened version of the test while maintaining high reliability and content validity.

Table 1: Summary of Changes to MET Test Length and Seat Time

Section	Previous MET		Revised MET	
	Length	Seat Time	Length	Seat Time
Listening	60 questions	45 minutes	50 questions	35 minutes
Reading & Grammar	75 questions	90 minutes	50 questions	65 minutes
Total (2-Skill MET)	135 questions	135 minutes	100 questions	100 minutes

Table 1 provides an overview of the changes to the overall length and number of questions in each section, while Tables 2 and 3 provide additional information on the exact changes to the number of items of each type.

Table 2: Summary of Changes to the Number of Listening Section Questions

Item Type	Previous	Revised
Short Conversations	22 questions	19 questions
Longer Conversations	21 questions (6 sets)	14 questions (4 sets)
Short Talks	17 questions (4 sets)	17 questions (4 sets)

Table 3: Summary of Changes to the Number of Reading & Grammar Section Questions

Item Type	Previous	Revised
Grammar	25 questions	20 questions
Extended Reading	N/A	10 questions (2 sets)
Thematic Reading	50 questions (4 sets)	20 questions (2 sets)

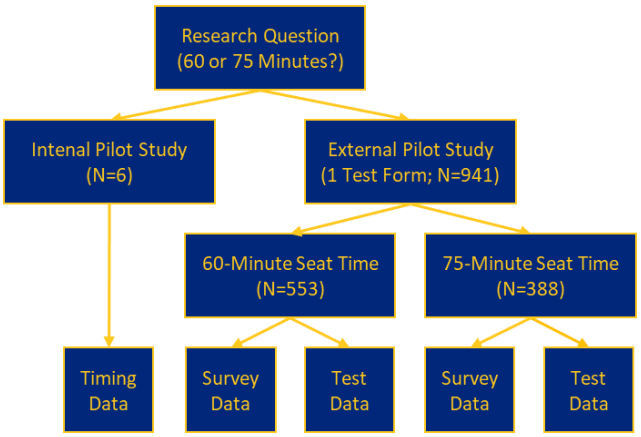
4. Determining Seat Time

Another challenge faced by the MET revision team was determining an appropriate new seat time for the revised reading and grammar section. In order to avoid introducing any undue speededness, a study was conducted to determine an appropriate seat time.

4.1 Study Design

Item response timing data from a previous research project was used to determine the upper and lower limits for two potential seat times to be trialed: 60 minutes and 75 minutes. Figure 1 describes the design of the research study. It shows that several sources of evidence were used to determine which seat time was most appropriate, including timing data from an internal pilot, test data from an external pilot, and survey response data from an external pilot.

Figure 1: Seat Time Study Design



4.2 Internal Pilot Timing Data

Prior to piloting the revised reading and grammar section on test takers, a small internal pilot was conducted with 6 Michigan Language Assessment employees (3 native speakers and 3 highly proficient non-native speakers). The purpose of this internal pilot was to use the information on the amount of time the participants needed to complete the 50 and 75 item reading and grammar sections, along with the seat time for the 75 item reading and grammar section, to estimate an appropriate seat time for the revised 50 item reading and grammar section. While the small sample size meant that the results must be interpreted with caution, it resulted in a seat time estimate of 71.25 minutes for the revised reading and grammar section. This estimate was obtained using the following ratio:

$$\frac{Average\ Time\ (75\ Item)}{Previous\ Seat\ Time} = \frac{Average\ Time\ (50\ Item)}{Estimated\ Seat\ Time}$$

Table 4: Summary of Raw Test Results by Seat Time

Seat Time	N	Mean	SD	Minimum	First Quartile	Median	Third Quartile	Maximum
60-Minutes	553	31.43	11.01	0	22	32	41	50
75 Minutes	388	33.55	10.44	10	26	34	43	50
Welch Two Sample T-Test: T=-2.9975, df=859.54, p-value=0.0028								

4.3 External Pilot Test Data

Following the piloting of the revised reading and grammar section, several different analyses were done to investigate the effects of the different seat times.

Table 4 summarizes the raw score results for the 60-minute and 75-minute seat time populations. It shows that there was a statistically significant difference in test taker performance between the two groups, with test takers who had the 60-minute seat time answering two fewer items correct, on average, than test takers who had the 75-minute seat time.

Three indices described in Lu & Sireci (2007) for evaluating test speededness in a single administration were used in this study to determine if either the 60-minute or 75-minute seat times resulted in any undue test speededness. Table 5 describes the three indices, providing information on their calculation and interpretation, while Table 6 presents the values of the different speededness indices for each seat time. The small power ratio (<0.25), large speededness ratio (close to 1), and small speededness quotient (close to 0) for both seat times suggests that neither seat time resulted in a speeded test.

Table 5: Description of Single Administration Speededness Indices

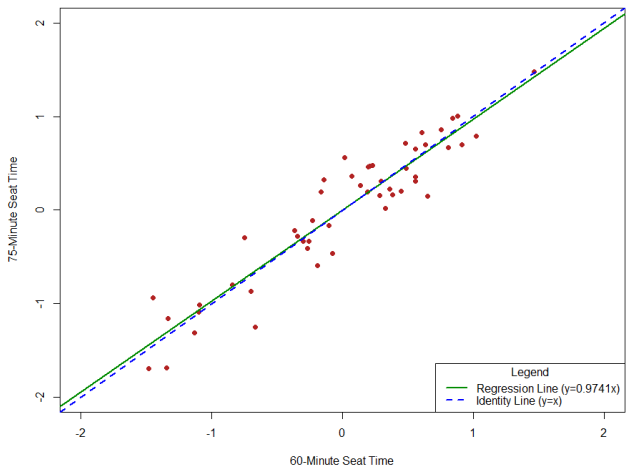
Indices	Formula	Description	Interpretation
Power Ratio	$\frac{S_U}{S_X}$	Ratio of the standard deviations of the number of items not reached (U) to the total number of items not given a correct answer (X).	Values less than 0.25 are indicative of an unspeeded test
Speededness Ratio	$\frac{S_W}{S_X}$	Ratio of the standard deviations of the number of items incorrectly answered or omitted (W) to the total number of items not given a correct answer (X).	Values less than 0.1 are indicative of a speeded test.
Speededness Quotient	$\frac{\sum U}{\sum W + \sum U}$	Proportion of items not reached (U) to the total number of items not given correct answers (W+U) summed across all test takers.	Values close to 0 are indicative of a power test. Values close to 1 are indicative of a speeded test.

Table 6: Summary of Single Administration Speededness Indices by Seat Time

Indices	60-Minutes	70-Minutes
Power Ratio	0.14	0.03
Speededness Ratio	0.97	1.00
Speededness Quotient	0.012	0.001

Differential item function analysis was also performed to determine if there was a significant difference in item difficulties between test takers who took the 60-minute and 75-minute seat times. Figure 2 presents a scatterplot of the item difficulties for the two seat times, along with regression and identity lines, in order to allow for a visual comparison of the item difficulties produced by the two seat times.

Figure 2: Comparison of Item Difficulties for 60-Minute and 75-Minute Seat Times



The clustering of the points near the identity line, and the similarity of the regression line to the identity line suggest that the item difficulties were not substantially different, regardless of the seat time allotted. Additionally, the coefficient of determination was high (r^2=0.8771), which means that most of the variation in item difficulties for the 75-minute seat time (87.71%) can be explained by the item difficulties for the 60-minute seat time.

4.4 External Pilot Survey Data

In order to collect information on perceptions of the seat times, surveys were administered to both the test takers and the proctors to collect information on whether they felt that the amount of time given for the reading and grammar section was "too much", "ok", or "too little". Tables 7 and 8 summarize the distribution of test taker and proctor responses, respectively, for each seat time.

Table 7 shows that seat time difference had a statistically significant impact on test takers' perceptions of having "too much", "ok", or "too little" time to complete the exam. By contrast, Table 8 shows that seat time differences did not have a statistically significant impact on the proctors' perceptions of the test takers having "too much", "ok", or "too little" time to complete the exam.

4.5 Final Seat Time

While the analysis of the survey data revealed that the seat time did have a significant impact on the test takers' perception of the amount of time they had to complete the exam, the results of the differential item function analysis indicate that there was not a significant difference in test taker performance on the 60-

Table 7: Summary of Test Taker Survey Responses by Seat Time

Seat Time	N	Too Much	OK	Too Little
60-Minutes	550	3.27	71.27	25.45
75 Minutes	385	5.71	87.79	6.49
Pearson's Chi-Square Test (Independence): X^2=57.21, df=2, p-value<0.001				

Table 8: Summary of Proctor Survey Responses by Seat Time

Seat Time	N	Too Much	OK	Too Little
60-Minutes	33	0.00	78.79	21.21
75 Minutes	16	6.25	81.25	12.50
Fisher's Exact Test (Independence): p-value=0.3887				

minute and 75-minute test forms. Furthermore, the Lu & Sireci (2007) indices for evaluating test speededness in a single administration indicated that neither seat time resulted in undue speededness. This evidence suggests that any seat time between 60 and 75 minutes would have been appropriate for the revised MET reading and grammar section. Using this information, the test revision team ultimately decided that a 65-minute seat time for the revised MET reading and grammar section provided a reasonable compromise between the practicality of a shorter seat time and the need to minimize construct irrelevant variance.

5. Maintaining Test Reliability

Finally, one of the most important considerations during the revision of the MET was ensuring that the revised test maintained a high level of reliability, despite the decrease in test length. The effects of different test lengths on the MET's reliability were estimated for different revision scenarios using response data from five operational test administrations to simulate responses to the revised test forms. Table 9 summarizes these reliability estimates for the revised MET. They show that the reliability estimates for both sections were above the minimally acceptable value of 0.80, which suggests that the revised MET still provides excellent consistency of measurement.

Table 9: Estimated Reliability of The Revised MET

Administration	Listening	Reading & Grammar
Test 1	0.854	0.827
Test 2	0.886	0.853
Test 3	0.845	0.849
Test 4	0.900	0.892
Test 5	0.886	0.864
Average	0.874	0.857

6. Conclusion

Overall, this report has provided a summary of the revisions that were made to the MET in 2019. It details the changes that were made and describes research that was done to determine a new seat time for the reading and grammar section and investigate the impact of the changes on test reliability. In addition to the research discussed in this report, additional analyses are routinely conducted to continuously monitor the impact of the MET revisions to ensure that the changes have had the expected effects.

7. References

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and Writing in the University: A Multidimensional Comparison. *TESOL Quarterly*, 36(1), 9-48. doi:10.2307/3588359

Khalifa, H., & Weir, C. J. (2009). Examining reading: Research and practice in assessing second language reading: *Studies in language testing*, 29. Cambridge, UK: Cambridge University Press.

Lu, Y. and Sireci, S. G. (2007), Validity Issues in Test Speededness. *Educational Measurement: Issues and Practice*, 26: 29–37. doi:10.1111/j.1745-3992.2007.00106.x

Macmillan, F., Chapman, M., & Stucker, J. R. (2014). A look into cross-text reading items: Purpose, development, performance. *Research Notes*, *55*, 12-15.