toeic® COMPENDIUM STUDY

Design and development of the TOEIC® Listening and Reading Test

Multistage adaptive test version

Jaime Cid, Youhua Wei, and Feng Yu | *ETS, USA*
Chihiro Sampei | *IIBC, Japan*

August 2026

TC-26-01

Introduction

The TOEIC® Listening and Reading (TOEIC L&R) test is a standardized global English-language proficiency test for people whose native language is not English. It measures the English communication skills required in the workplace and everyday life. The TOEIC L&R test has typically been administered in a full-length linear paper-and-pencil or computer-based format (hereinafter referred to as "Linear Test") which includes 100 items each in the Listening and Reading sections, and the total testing time is approximately two hours. Operational evidence reported by ETS (2025) indicates that the Linear Test scores have demonstrated internal consistency reliability (KR-20) of approximately 0.90 or higher across forms in norming samples. Given this evidence, the Linear Test may be described as a widely used assessment for measuring English-language proficiency.

However, the Linear Test is not necessarily efficient in terms of testing time: two hours can be a little too long for some test takers. In fact, there has been a need, especially from institutional test users[1], for a more "practical" version of the test, requiring less testing time, for test takers, but also for administrators as well, while maintaining measurement quality comparable to that of the Linear Test version.

To meet this need and improve the test-taking experience, the TOEIC Program team at ETS designed and developed a Multistage Adaptive Test version of the TOEIC L&R test (hereinafter referred to as "MSA Test") based on advanced psychometric methodology, specifically Item Response Theory (IRT). The MSA Test version contains 45 items each in the Listening and Reading sections, and its total testing time is approximately one hour. This means that the MSA Test improves measurement efficiency by administering fewer test items compared to the Linear Test and reducing total testing time. Nevertheless, the MSA Test was deliberately developed to be comparable to the Linear Test in several fundamental quality aspects: a) the content and construct measured by the test, b) the interpretation and reporting scale of the scores (ranging from 5 to 495 in increments of 5 for both the Listening and Reading sections), and c) the overall measurement quality.

The purpose of this report is to provide a detailed account of the design and development of the MSA Test, including the theoretical considerations that informed its construction. The report outlines the overall test design, key features, and findings from the pilot study conducted during the development process.

[1] The TOEIC Test is offered via the Public Testing Program, where individuals arrange their own tests, or through the Institutional Program, which is organized by corporations or other groups.

Theoretical considerations for multistage adaptive testing

An MSA test is a type of computer adaptive test designed to improve efficiency and accuracy of measurement. One significant advantage of adaptive testing over linear testing is the high efficiency of the measurement process. In traditional linear testing, all test takers receive the same set of items (both easy and difficult) regardless of their actual ability levels. To accurately measure all test takers across a range of ability levels, linear tests require many test items to cover a wide range of difficulty. As a result, linear tests are generally lengthy, requiring a long testing time. In contrast, adaptive tests require relatively fewer items and less testing time, as they are designed to adjust test item difficulties to match test takers' abilities. In computer adaptive testing, if a test taker gives a wrong answer, an algorithm selects an easier item; if the test taker answers correctly, the next item will be more difficult. Generally speaking, test items that are matched to test takers' abilities (the items that are challenging but not overwhelming) can provide more informative measurement data about test takers' true ability. Conversely, unmatched items tell little about test takers' true ability. This is because it is highly likely that a test taker at a high ability level will answer an easy item correctly and a test taker at a low ability level will answer a difficult item incorrectly. By avoiding items that are unmatched to the test taker's proficiency, adaptive testing optimizes testing efficiency.

Despite these advantages, choosing between adaptive tests (including multistage adaptive tests, MSA) and linear tests involves important security and fairness tradeoffs. Adaptive approaches require a large, pre-calibrated item pool with psychometric parameters (e.g., difficulty) placed on a common scale (Lord, 1980; Weiss, 1982); strong calibration supports fair routing and scoring, but weak subgroup calibration or parameter drift can threaten fairness (Debelak et al., 2023). From a security standpoint, linear tests concentrate exposure on one (or a few) fixed forms—easier to monitor, but any compromised items can affect many test takers. Adaptive designs distribute exposure across items/modules, yet high-frequency routing modules in MSA may be widely seen and therefore require tight exposure control and monitoring for pre-knowledge.

Adaptive testing follows an iterative process that includes two basic steps: item selection and ability estimation. The item selection step is to select the most appropriate test items to ask, given what is currently known about the test taker's ability level. The ability estimation step is to re-estimate the test taker's ability based on the test taker's answers to all the items which have been given to the test taker so far. The newly estimated ability will be the input for the next iteration. This iterative process repeats until a target criterion—such as a specific number of items—is met.

Adaptive tests can be designed to adapt after each individual test item (item-level) or after a set of items (module-level). The latter design is known as Multistage Adaptive (MSA) testing. An MSA test starts by presenting an identical set of test items to test takers at all proficiency levels and then routes each test taker to subsequent stages of varying difficulty, depending on the test taker's performance in the previous stage.

The MSA approach is suitable for tests that include both discrete items and set items (i.e., several items associated with the same set leader, such as a reading passage), as is the case with the TOEIC L&R test. On the other hand, the item-level adaptive approach is more suitable for tests with only discrete items. Additionally, the MSA approach provides a good balance among practicality, flexibility, measurement accuracy, and control over test content coverage (Zenisky et al., 2010). Especially when stringent

psychometric requirements are met, MSA tests can offer practical advantages over item-level adaptive tests, such as better management of item pool usage, greater control over test content coverage, increased flexibility in test assembly, and the ability for test takers to move back and forth within items in the same testing stage (for the TOEIC L&R MSA Test, this applies only to the Reading Section).

TOEIC L&R MSA test design and development

ETS researchers and content experts followed best practices when developing the TOEIC L&R MSA Test. This section describes the overall design and some key features of the TOEIC L&R MSA Test: its content construct, items, routing mechanism, reported scores, and testing time.

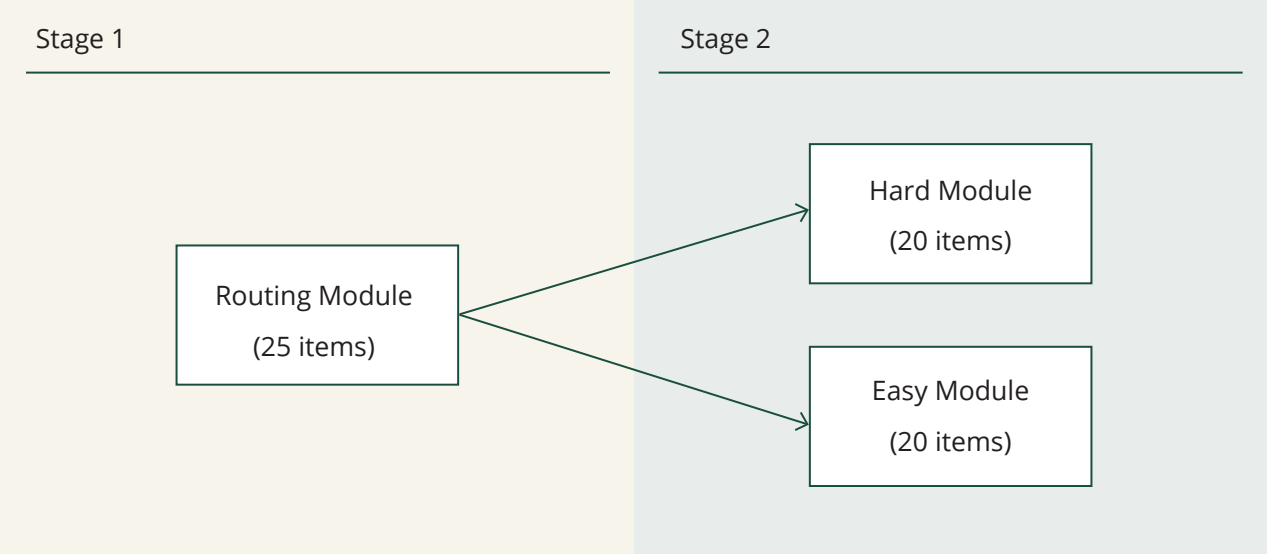
Overall design

Because the TOEIC L&R test includes both discrete items and set items, and due to the advantages stated earlier, the MSA approach was adopted as the basic design for the TOEIC L&R Multistage Adaptive Test. Additionally, to minimize the testing time while maintaining the highest possible content and construct comparability with the Linear Test, the final design was a two-stage MSA test with a total of 45 items in each of the Listening and Reading sections.

There are two testing stages for each test section: one Routing Module on Stage 1 and two modules—Easy Module and Hard Module—in Stage 2 (MSA 1–2 design). According to the design, all test takers are first given the pre-assembled Routing Module with 25 identical items at a medium difficulty level (Stage 1). Based on their performance in the first stage of testing, routing decisions are made. Test takers who perform at or above a predetermined proficiency criterion in the first stage of testing are routed to the pre-assembled Hard Module, which contains 20 items, in the second stage of testing (Stage 2). The Hard Module contains more difficult items, which more closely match the test takers' true abilities. Those test takers who perform below the criterion in the first stage of testing are routed to the pre-assembled Easy Module, which also contains 20 items, in the second stage of testing (Stage 2). The Easy Module contains easier items, which more closely match the test takers' true abilities. Overall, there are two adaptive paths, the Hard Path (25 items from the Routing Module and 20 items from the Hard Module) and the Easy Path (25 items from the Routing Module and 20 items from the Easy Module). The two-stage MSA Test ends with the completion of the second-stage module, and the final test score is calculated based on all the items included in the Routing Module and the second-stage module that the test taker completed (Davey, 2011). Figure 1 illustrates the TOEIC L&R MSA Test design for each test section.

Figure 1

MSA 1–2 Design for TOEIC L&R (45 Items Per Test Section)



Test features

The next sections describe key features of the TOEIC L&R MSA Test and present relevant data and findings from a pilot study conducted during the test development process. In this pilot study, 707 test takers who had recently taken the TOEIC L&R Linear Tests through Public Testing Program were selected as participants. Table 1 shows the distribution of the participants' Total scores (Listening and Reading combined scores) from the Linear Tests.

Two MSA Test forms (Form A and Form B) created for illustrative purposes were used in the pilot study; Form A was administered to 350 participants, and Form B was given to the remaining 357 participants. During the data cleaning process, a few participant cases were deemed outliers and were excluded from the final statistical analyses. Table 2 provides the final sample sizes for each test section, adaptive path, and test form.

Table 1

Frequency Distribution of TOEIC L&R Linear Test Scores by Pilot Study Participants

TOEIC L&R Total Score Range	Number of Participants (%)
10–100	0 (0.0%)
105–200	3 (0.4%)
205–300	48 (6.8%)
305–400	140 (19.8%)
405–500	128 (18.1%)
505–600	114 (16.1%)
605–700	68 (9.6%)
705–800	69 (9.8%)
805–900	78 (11.0%)
905–990	59 (8.3%)
Total	707 (100.0%)

Table 2

Sample Sizes for Each Test Section, Path, and Form

		Section	
Test Form	Path	Listening	Reading
Form A	Easy Path	81 (23%)	177 (52%)
	Hard Path	268 (77%)	166 (48%)
	Total	349 (100%)	343 (100%)
Form B	Easy Path	66 (18%)	179 (52%)
	Hard Path	291 (82%)	166 (48%)
	Total	357 (100%)	345 (100%)
Forms A and B	Easy Path	147 (21%)	356 (52%)
	Hard Path	559 (79%)	332 (48%)
	Total	706 (100%)	688 (100%)

Test content construct

To ensure that the underlying test construct measured by the Linear Test was maintained, the MSA Test was constructed as closely as possible to match the content specifications of the Linear Test. Table 3 shows a comparison of the content specifications between the MSA Test and the Linear Test for the Listening and Reading sections.

Table 3

Content Specification Comparison between TOEIC L&R Linear Test and MSA Test

			Number of Items (MSA Test)		
Section	Item Type	Number of Items (Linear Test)	Stage 1 Routing Module	Stage 2 Easy or Hard Module	Total
Listening	Photographs	6	3	0	3
	Question-Response	25	4	5	9
	Conversations	39	9	9	18
	Talks	30	9	6	15
	Total	100	25	20	45
Reading	Incomplete Sentences	30	5	7	12
	Text Completion	16	4	4	8
	Single Passages	29	16	9	25
	Multiple Passages	25			
	Total	100	25	20	45

Test items

All of the TOEIC L&R MSA Test items come from a large pool of high-quality items accumulated through the Linear Test administrations. Through the Item Response Theory (IRT) 2-parameter model, the items have been thoroughly analyzed and indexed in detail; for each item, there is an a-parameter for item discrimination and a b-parameter for item difficulty. This feature makes the adaptive structure of the MSA Test feasible, as it underlies the routing mechanism of the test (for details, see the later section "*Routing Mechanism*"). Also, since the items in the MSA Test have been analyzed through the Linear Test administrations, the IRT parameter estimates for the MSA Test items are on the same IRT scale as those for the Linear Test. Therefore, the test items in these two test versions are directly comparable in terms of item discrimination (a-parameter) and item difficulty (b-parameter), which is the theoretical foundation for the reported scores from both versions of the TOEIC L&R test being directly comparable and having the same psychometric meaning (for details, see the later section *Reported Score*). As a reference, Table 4 provides summary statistics of the IRT item parameter estimates for the two MSA Test forms created for the pilot study. For both sections in both forms, the Easy Modules have much lower b-parameter means than the Hard Modules; the b-parameter means for Routing

Modules are intermediate between the Easy and Hard Modules. This means that the Easy Modules are much easier than the Hard Modules and the Routing Modules have medium difficulty level. Combining Stage 1 and Stage 2 modules together, the Easy Paths are much easier than the Hard Paths. This is consistent with the MSA design described above. In addition, the summary statistics of corresponding paths in both Listening and Reading sections are similar, with minor differences.

Table 4

Summary Statistics of IRT Item Parameter Estimates for Pilot MSA Test Forms

		Stage/Module/ Path		Number of Items	a-parameter[2]		b-parameter[3]	
Test Form	Section	Stage	Module		Mean	SD	Mean	SD
Form A	Listening	1	Routing	25	0.57	0.20	-1.07	1.01
		2	Easy	20	0.78	0.26	-1.53	1.00
		2	Hard	20	0.68	0.23	-0.41	0.63
		Easy Path		45	0.67	0.25	-1.27	1.02
		Hard Path		45	0.62	0.22	-0.77	0.92
	Reading	1	Routing	25	0.53	0.20	-0.17	1.32
		2	Easy	20	0.61	0.19	-1.04	0.98
		2	Hard	20	0.50	0.19	0.52	1.10
		Easy Path		45	0.57	0.20	-0.56	1.25
		Hard Path		45	0.52	0.19	0.14	1.26
Form B	Listening	1	Routing	25	0.57	0.18	-0.91	0.89
		2	Easy	20	0.66	0.21	-1.65	0.78
		2	Hard	20	0.62	0.18	-0.25	0.65
		Easy Path		45	0.61	0.20	-1.24	0.91
		Hard Path		45	0.59	0.18	-0.62	0.85
	Reading	1	Routing	25	0.56	0.22	-0.29	1.01
		2	Easy	20	0.63	0.25	-1.12	0.94
		2	Hard	20	0.53	0.18	0.20	0.81
		Easy Path		45	0.59	0.23	-0.66	1.05
		Hard Path		45	0.55	0.20	-0.08	0.95

[2] In IRT, the a-parameter is a measure of item discrimination. It typically ranges from 0 to 2, with a higher value indicating better discrimination (how well an item differentiates low and high ability test takers).

[3] The b-parameter is a measure of item difficulty. It typically varies from -3 to 3, with a higher value indicating higher difficulty.

Routing mechanism

For each test section of each MSA Test form, there is a routing criterion by which test takers are directed to either the Easy or Hard Module in the second stage of testing (Stage 2). This routing decision ensures that the module selected for Stage 2 provides the most precise measurement of each test taker's estimated ability. In the MSA Test, the routing criterion is a specific raw score (the total number of correct answers) on the Routing Module (Stage 1), determined through a psychometric procedure developed by Luecht (2014). The procedure establishes the logical threshold for the routing decision by identifying the IRT theta value (which indicates ability level) at the intersection of the module information functions (MIFs) for the Easy and Hard Modules. MIF reflects the level of measurement precision at different ability levels, or IRT theta values. The point where the MIFs of the two adjacent modules intersect represents the theta level at which the two modules provide equal information about test takers' ability. Therefore, measurement precision is considered maximized by routing test takers according to this threshold. The Stage 2 Easy Module provides maximum information for test takers below the threshold, and the Stage 2 Hard Module provides maximum information for test takers above the threshold. To determine the cut score for the Routing Module, the theta value at the MIF intersection is converted to the corresponding raw score using the test characteristic curve (TCC) of the Routing Module. The expected raw score is then rounded up to the nearest integer, which is used as the predetermined routing criterion.

Reported score

As with the Linear Test, the reported score for each MSA Test section is a scaled score, calculated from the raw score for the entire test section. In the scoring process, raw scores are converted to corresponding scaled scores through raw-to-scale score conversion for each Path (Easy Path and Hard Path) of each MSA Test form. Therefore, the final scaled score for each test section depends on the path assigned to the test taker and the number of items answered correctly throughout the section. The conversion is created through the IRT true score equating method (Kolen & Brennan, 2004) in the MSA Test form assembly. The underlying IRT scale for the test items and the score reporting scale are common across both the MSA Test and the Linear Test, and the scaled score of the MSA Test is pre-equated to the Linear Test scaled score. Therefore, the scaled scores of these two versions of the TOEIC L&R test are on the same reporting scale and are directly comparable.

With IRT item parameter estimates and raw-to-scale score conversions available, the reliability of the MSA Test scaled scores can be directly estimated using an IRT methodology (Kolen et al., 1996). For the two MSA Test forms used in the pilot study, reliability estimates for each test section ranged from 0.86 to 0.88. Since the same test design framework and test specifications are used for all the MSA Test form assemblies, reliability estimates are expected to be within this range for all operational MSA Test forms. For comparison, the average reliability estimates are approximately 0.90 for each of the Listening and Reading sections of the Linear Test forms. In general, test length has a direct impact on test form reliability for linear tests. That is, all else being equal, a linear test with 45 items would have a much lower reliability estimate than a linear test with 100 items. It is clear that an adaptive testing approach with high-quality test items enabled the MSA Test to achieve measurement quality comparable to the Linear Test, while shortening the test length by 55% in terms of the number of items.

Testing time

Table 5 presents a summary of the testing time (in minutes) for each MSA module (Routing, Easy, and Hard), path (Easy and Hard), and section (Listening and Reading). The testing time is approximately 25 minutes for the Listening section and 37 minutes for the Reading section, regardless of which path a test taker is routed to. The total MSA testing time is approximately 62 minutes for all test takers.

Table 5

Summary of MSA Testing Time (in Minutes) for Each Module, Path, and Section

	Stage 1	Stage 2		Total	
Section	Routing Module	Easy Module	Hard Module	Easy Path	Hard Path
Listening	13.5	11.5	11.5	25	25
Reading	23	14	14	37	37

Because TOEIC Listening Comprehension is paced by an audio recording, speededness is not a concern. However, the TOEIC Reading section is designed to be a power test so that scores mainly reflect test takers' abilities on the construct, rather than their working speed. Therefore, the testing time should be sufficient to allow most test takers, regardless of adaptive path or proficiency level, to complete the test questions at a reasonable pace. To confirm this, the allocated testing time and speededness of the Reading section were validated through the pilot study[4]. Table 6 shows the average time (in minutes) the study participants spent at each stage, path, and the entire Reading section. The average total time per section differed modestly between the participants routed to the Easy Path (32.4 minutes) and those routed to the Hard Path (34.3 minutes).

Table 6

Average Time Spent on Reading Questions by Stage, Path, and Section

	Number of	Average Time Spent		
Path	Participants	Stage 1	Stage 2	Section
Easy Path	356	20.3	12.1	32.4
Hard Path	332	21.0	13.3	34.3
Total	688	20.7	12.7	33.3

Furthermore, Table 7 shows the data for different proficiency-level groups based on the participants' scaled scores on the MSA Test Reading section.

[4] The Listening section is time paced by audio recordings. The testing time fluctuates depending on the test items included in each test form, and there is no need to examine the speededness of the Listening section.

Table 7

Average Time Spent on Reading Questions by Proficiency Group

Scaled Score on MSA Test Reading Section	Number of Participants	Average Time Spent		
		Stage 1	Stage 2	Section
Low (195 and Below)	226	20.0	11.4	31.4
Medium (200–325)	230	**22.6**	12.2	34.9
High (330–495)	232	19.4	**14.3**	33.7
Total	688	20.7	12.7	33.3

On Stage 1 (Routing Module), participants at the medium score level (200–325) spent 22.6 minutes on average, which was the longest time among the three groups. A possible explanation is that the test takers in the high proficiency group did not need much time since the items in the Routing Module were relatively easy for them; at the same time, the items were relatively difficult for those in the low score group, and spending more time did not help them. On Stage 2, the test takers in the high score group needed 14.3 minutes on average, the longest among the three groups. This is reasonable because the Hard Path, which most of the high proficiency test takers (if not all) were routed to, included more lengthy reading passages than the Easy Path. The pilot study confirmed that the testing time of 37 minutes (22.6 + 14.3 ≈ 23 + 14) for the Reading section minimizes test speededness for any of these groups. Since the same test design framework and test specifications are used for the test form assembly of all MSA Test forms, this testing time is applied to all operational MSA Test forms.

Table 8 provides the speededness statistics for the Reading section of the two MSA Test forms. The statistics show that a) nearly all examinees completed 75 percent of the items in each MSA Path; b) all items were answered by at least 80 percent of the examinees; and c) the speededness index (the ratio of the variance of the number of items not answered to the variance of total raw scores) was below 0.15 for each MSA Path. As a criterion, if this index value is 0.15 or less, the test is considered unspeeded (Walker, 1981). The three statistics in the table confirmed that the MSA Test Reading section was not speeded, and these statistics were very similar to findings from typical paper-based Public Testing Program Linear Test administrations. The sample data are also provided in the table.

Table 8

Speededness Statistics for the MSA Test Reading Section

Speededness Statistics	Form A Easy Path	Form A Hard Path	Form B Easy Path	Form B Hard Path	Linear Test
Number of test takers	177	166	179	166	37,840
Number of Items	45	45	45	45	100
Percentage of test takers answering at least 75% of items	100.00	98.19	99.44	100.00	99.4
Number (%) of items answered by at least 80% of test takers	45 (100%)	45 (100%)	45 (100%)	45 (100%)	100 (100%)
Ratio of not answered variance / total variance	0.09	0.11	0.14	0.05	0.05

Conclusion

This report describes the design and development of the TOEIC L&R MSA Test, with particular attention to the test's content construct, item composition, routing mechanism, scoring procedures, and testing time. The MSA Test was developed to enhance the test-taking experience by increasing the efficiency of the test administration while maintaining measurement quality comparable to that of the full-length Linear Test. It is important to note, however, that the two versions of the TOEIC L&R test are not identical by design. The MSA Test leverages IRT methodology and a statistically calibrated item bank to reduce testing time while preserving the underlying test construct and measurement precision, rather than replicating the Linear Test in full. Findings from the pilot study conducted during the development process indicated that the MSA Test and the Linear Test fulfilled their intended purposes at similarly high levels of quality.

References

Davey, T. (2011). *Practical considerations in computer-based testing*. ETS. http://www.ets.org/Media/Research/pdf/CBT-2011.pdf

Debelak, R., Appelbaum, S., Debeer, D., & Tomasik, M. J. (2023). Detecting differential item functioning in 2PL multistage assessments. *Psych*, *5*(2), 461–477. https://doi.org/10.3390/psych5020031

ETS. (2025). *TOEIC® listening and reading examinee handbook.* https://www.ets.org/pdfs/toeic/toeic-listening-reading-test-examinee-handbook.pdf

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*, 129–140. https://doi.org/10.1111/j.1745-3984.1996.tb00485.x

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Luecht, R. (2014). Design and implementation of large-scale multistage testing systems. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 69–83). CRC Press.

Walker, R. C. (1981). *A reader's guide to test analysis reports*. ETS.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473–492. https://doi.org/10.1177/014662168200600408

Zenisky, A. L., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). Springer.

toeic®