

Finding time in language assessments: Maximizing measurement properties by unit of testing time

Sarah Goodwin, Geoffrey T. LaFlair, J.R. Lockwood,
Steven W. Nydick, & Alina A. von Davier

American Association for Applied Linguistics

March 21, 2023



Test usefulness (Bachman & Palmer, 2010)

- Validity
- Reliability
- Practicality
- Impact
- Authenticity
- Interactivity



Practicality and impact affect access



Bandwidth considerations



Given in 213 countries and regions



Short administration time

Timing considerations

To avoid fatiguing or overwhelming examinees, we need to know what effects timing has on test items, all while ensuring **reliability** and **validity**



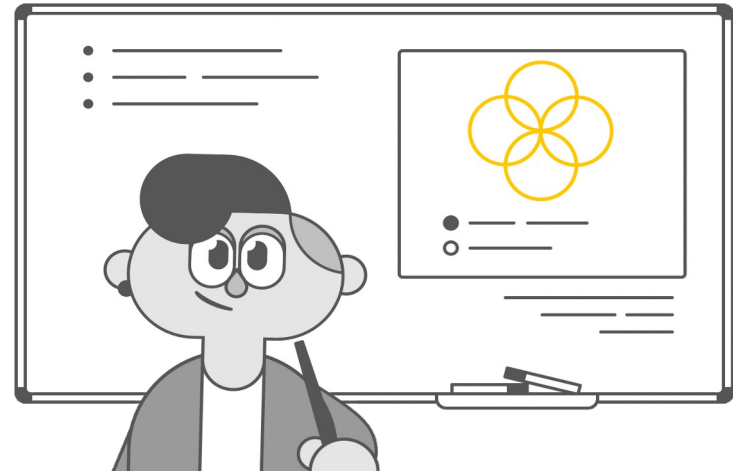
How critical is timing to a test construct?

- Imposed time limits, which are the norm on large-scale standardized tests, reduce sources of random error
- Time constraints associated with real-world tasks (speaking, writing) reflected as much as possible in assessment tasks
- Though response time does not affect scoring of performance, some degree of automaticity is expected for proficient test takers to complete test tasks

(Kane, 2020; Sireci & Botha, 2020; Wise & Kuhfeld, 2020)

Computer-adaptive test (CAT)s

- CATs are shorter than fixed-form tests but must have enough items to achieve a reliable final ability estimate
- **Efficiency:** the amount of measurement information per unit of time an item takes to administer (Cheng et al., 2017; Fan et al., 2012)
- With some exceptions (e.g., Crabtree, 2016; Hughes & Clesham, 2021; Kim et al., 2020, 2022), **timing** has sparingly been used to support test validity arguments



Finding Time: Steps we followed

- Test development/Literature review of factors affecting timing
 - Discussion of how to proceed
- Simulation of existing test taker results recalculating what would happen if items removed
 - Discussion
- Items removed to make enough time for a new item type

1:32



Type the missing letters to complete the text below

Sam bought a new coat yesterday. He **i s** often **c o l d**, **a n d** he **c h o s e** a **c o** that
i warmer **t h** his **o** jacket. His **c o** is **g r** and **r**. It **l o**
good **w i** his **r** hat. He is going to wear it every day this winter.

NEXT

Item type	Measurement opportunities	Maximum time / item
Conversation and Comprehension	9 per item set	1:30
Literacy and Comprehension	18 per item set	1:00
Conversation, Comprehension, Production	[depends on scoring features]	0:20
Literacy, Comprehension, Production	10 to 18 per item set	3:00
Literacy, Conversation, Production	[depends on scoring	1:00

Affecting timing: Test delivery and format

- Whether a recording can be heard multiple times
- At what rate a voice plays in listening input
- Length, e.g., single image versus a video, or the number of characters or words of input text
- Types of interaction between/among participants and items
- Degree of speededness, e.g., whether candidates are meant to process a 100-word passage in the same time as, or different timing allotted to, a 200-word passage
- User interface/navigation, e.g., whether candidates must scroll or all information is viewable on a single screen
- Local Internet bandwidth

Simulations



- 294,456 test takers (their first or only test)
- Only those who took at least 3 of each CAT item type
- Compared three conditions:
 1. Unchanged test
 2. Effect on scores/timing if one of each CAT item type (five items) removed
 3. Effect on scores/timing if two of each CAT item type (ten items) removed

Needed to decide among actions:

1. Decrease the number of measurement opportunities (MOs) from items but keep the same number of items
2. Keep the numbers of items intact but trim the time allotted to test takers to complete the same tasks
3. Remove whole items

Or another option? Documented all consequences ("If we changed x, then y...")

Old
exam

CAT_{old}

Production

How to go from old CAT to new CAT, to make room for new items?

New
exam

CAT_{new}

New
items

Production

Contribution of Item Types to Duolingo English Test Subscores

		Literacy	Conversation	Comprehension	Production
Reading	C-test	X		X	
	Yes/no (text)	X		X	
	Interactive reading	X		X	
Listening	Yes/no (audio)		X	X	
	Dictation		X	X	
Writing	Writing (image)	X			X
	Writing (text)	X			X
	Writing sample	X			X
Speaking	Elicited imitation		X	X	
	Speaking (image)		X		X
	Speaking (text)		X		X
	Speaking (audio)		X		X

Timing

- Condition 1 [leaving test unchanged]: CAT section median 28.39 minutes
- Condition 2 [removing one of each CAT item type]:
 - ~4.5 minutes gained, not including interstitial screens
- Condition 3 [removing two of each CAT item type]:
 - ~9 minutes gained, not including interstitial screens
- New CAT section: median ~19 minutes, with 10 fewer items

Caveats

- Balance in considering relationships among test length, psychometrics, and construct coverage
- This is part of a larger project where we sought to find time before adding new items
 - Adding new items was the step after this project

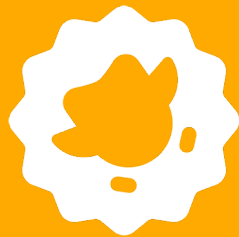
Reminder: importance of collaboration

- Factors affecting timing
 - Applied linguist and language assessment expertise
- Simulating what would happen if items removed
 - Psychometric considerations
- Items removed to find time for a new item type

Consider the time examinees spend as part of test validation:

Could we operationalize test efficiency as a "velocity" measure (e.g., measurement opportunities / time)?





thanks! questions?

sarahg@duolingo.com

References

- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford.
- Cheng, Y., Diao, Q., & Behrens, J. T. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behavior Research Methods*, 49(2), 502-512.
- Crabtree, A. R. (2016). *Psychometric properties of technology-enhanced item formats: An evaluation of construct validity and technical characteristics* [Doctoral dissertation, University of Iowa]. <https://doi.org/10.17077/etd.922fbj4d>
- Fan, Z., Wang, C., Chang, H. H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, 37(5), 655-670.
- Gutierrez, S. (2009). *Examining the psychometric properties of a multimedia innovative item format: Comparison of innovative and non-innovative versions of a situational judgment test*. [Doctoral dissertation, James Madison University].
- [Hughes & Clesham, 2021 - research summary of shortened test form](#)
- Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40(1), 1–15.
- Kane, M. (2020). The impact of time limits and timing information on validity. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 19-31). Routledge.
- Kim, A., Yumsek, M., Chapman, M. & Cook, H. G. (2019). *Investigating K-12 English learners' use of universal tools embedded in online language assessments* (WIDA Technical Report No. TR-2019-2). Board of Regents of the University of Wisconsin System.
- [Kim, A. A., Tywoniw, R. L., & Chapman, M. \(2022\). Technology-Enhanced Items in Grades 1–12 English Language Proficiency Assessments. *Language Assessment Quarterly*, 19\(4\), 343-367.](#) – based on [WIDA research report](#)
- Russell, M. (2016). A framework for examining the utility of technology-enhanced items. *Journal of Applied Testing Technology*, 17(1), 20–32.
- Sireci, S. G., & Botha, S. M. (2020). Timing considerations in test development and administration. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 32-46). Routledge.
- Wan, L. & Henly, G.A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education*, 25(1), 58–78. doi: 10.1080/08957347.2012.635507
- Wise, S. L., & Kuhfeld, M. R. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 150-164). Routledge.