The Duolingo English Test: Psychometric considerations

duolingo research

Duolingo Research Report DRR-20-02
February 27, 2020 (11 pages)
www.duolingo.com/research

Gunter Maris*

Abstract

The Duolingo English Test is a computer adaptive test of English language proficiency. In this paper, a measurement model for the Duolingo English Test is proposed, and its formal characteristics are derived. A scalable real-time rating system, with favorable statistical properties, is proposed and applied to data from the Duolingo English Test to address issues related to differential model fit with respect to item and person characteristics. The results indicate that the signed residual time model fit the test data well, and that there is not evidence of statistical bias toward different groups of test takers

This commissioned study is part of the Duolingo English Test external research collection.

Introduction

The Duolingo English Test is a large-scale, high-stakes, online test of English language proficiency that can be delivered anywhere and at any time. This approach to (language) testing inherently comes with a cold start problem: one needs a large item bank calibrated to an appropriate item response theory (IRT) model before one can start administering tests. The traditional solution to this problem, large scale pretesting, is not only expensive but suffers from a number of drawbacks. One needs large numbers of test takers to respond to items under realistic testing conditions. Both ensuring realistic testing conditions and getting test takers from the appropriate populations are challenging.

To support anywhere anytime testing, the Duolingo English Test is assembled uniquely for every test taker, in the form of a computer adaptive test (CAT). For the Duolingo English Test an alternative solution to the cold start problem is pursued. Items are automatically generated, using machine learning and natural language processing algorithms, to match the Common European Framework of Reference for languages (CEFR, Council of Europe, 2001), with item difficulty parameters that are predicted using machine learning. Item responses are automatically scored on a continuous zero to one scale. The development and scoring of the Duolingo English Test are documented in Settles, LaFlair, and Hagiwara (in press) and LaFlair and Settles (2019). The Duolingo English Test contains 10 item types, five of which are delivered in the CAT administration. These five item types include a ctest, audio yes/no vocabulary, text yes/no vocabulary, dictation, and elicited speech. The other five item types are open-ended speaking and writing tasks. The speaking tasks include a picture description prompt, a text-based prompt, and an audio prompt. The writing tasks include a picture description task and a text-based prompt. Duolingo English Test test takers receive a minimum of three and a maximum of seven of each of the CAT items; they receive four each of the open-ended writing and speaking tasks.

Once real tests are being administered using the pre-calibrated item bank, actual response data under high stakes testing conditions from actual test takers become available. This response data can be used to a) update where needed the pre-determined item difficulties, b) evaluate the fit of the IRT model, and c) to further train the machine learning and natural language processing algorithms to obtain ever better pre-calibrated item difficulties.

A CAT administration benefits both test security and measurement accuracy. At the same time, however, it complicates statistical analyses of item response data. Below we introduce an appropriate measurement model for the Duolingo English Test, and an appropriate scalable algorithm for statistical inference on the parameters of the measurement model.

A Measurement Model for the Duolingo English Test

The basic observations collected with the Duolingo English Test are continuous item responses between zero and one (x_i), with one (zero) indicating a fully (in)correct response. These observations are used to find that value of ability (θ) that

*ACTNext

Corresponding author:
Gunter Maris, PhD
Senior Director of Advanced Psychometrics, ACTNext
Iowa City, IA USA
Email: Gunter.Maris@act.org

minimizes the following cross entropy:

$$LL(\theta) = \sum_i x_i \log(p_i(\theta)) + (1 - x_i) \log(1 - p_i(\theta))$$

in which $p_i(\theta)$ is the item response function of the Rasch model:

$$p_i(\theta) = \frac{\exp(\theta - \delta_i)}{1 + \exp(\theta - \delta_i)}$$

with δ_i the difficulty of item i. As we'll see later on this approach works, but being based on information theory rather than on a statistical model, ignores that responses are inherently random. That is, if the same person were to take the same test again, we would not expect to see the exact same responses. A statistical model for the item responses accounts for this, and allows for looking at model fit, determining standard errors, and the like.

An appropriate item response theory (IRT) model that goes with these data is the signed residual time model of Maris and van der Maas (2012). Even though this model was originally proposed for a particular scoring rule that combines response accuracy with response time, the basic model is agnostic to this fact, and just needs observations in a finite interval. The SRT model is characterized in the following way for item scores x_i from 0 to 1:

$$\begin{aligned} f(\mathbf{x}|\theta) &= \prod_i f(x_i|\theta) \\ &= \prod_i \frac{\exp(x_i(\theta - \delta_i))}{\int_0^1 \exp(s(\theta - \delta_i))ds} \\ &= \prod_i \frac{\exp(x_i(\theta - \delta_i))}{\frac{\exp(\theta - \delta_i) - 1}{\theta - \delta_i}} \end{aligned}$$

This model is the straightforward extension of the Rasch model to continuous responses, and it shares with the Rasch model the property that the (person or item) sum scores are sufficient statistics for the model parameters (person ability θ or item difficulty δ).

To illustrate how the statistical model relates to the information theory based estimation procedure of the Duolingo English Test, we look at two simulated examples. For both examples, we have 1000 students with standard normally distributed ability and 50 items per student. For one simulation we have the same 50 items with standard normally distributed difficulty for all students, and for the other simulation the item difficulties are normally distributed with a standard deviation of 0.5 around the true ability of the student. The latter one mimics an adaptive test. Data are simulated from the statistical model, and ability is estimated by minimizing the cross entropy. For both simulations the correlation between true and estimated ability is high (0.82 for the linear simulation and 0.98 for the adaptive simulation), with the one for the adaptive test being significantly higher (as expected). However, looking at the scatter plot of true versus estimated ability (Figure 1) we see significant shrinkage for the linear test. For the adaptive test the shrinkage completely disappears. In practice, the information theory based approach works for the purpose of estimating ability, but as said before the statistical approach provides additional benefits.

The statistical model is not only a straightforward extension of the Rasch model, but the relation between the two runs quite a bit deeper. We show there is a one-to-one correspondence between a continuous response and an infinite sequence of binary responses, each from a (slightly different) Rasch model. The correspondence allows us to use algorithms, originally developed for binary responses, for continuous responses as well.

To demonstrate this we define two new variables as follows:

$$\begin{aligned} y_{i1} &= (x_i > 0.5) \\ z_i &= \begin{cases} x_i - 0.5 & \text{if } y_{i1} = 1 \\ x_i & \text{if } y_{i1} = 0 \end{cases} \end{aligned}$$

This construction separates the original continuous response into two conditionally independent sources of information on ability: $Y_{i1} \perp\!\!\!\perp Z_i|\theta$, from which the original observations can be reconstructed. Moreover, it is readily found that the implied measurement model for $\mathbf{Y}$ is the Rasch model:

$$p(Y_{i1} = 1|\theta) = p(X_i > 0.5|\theta) = \frac{\exp(0.5(\theta - \delta_i))}{1 + \exp(0.5(\theta - \delta_i))}$$

For the other variable (Z_i) we readily find its distribution to be (over the interval 0 to 1/2):

$$f(z_i|\theta) = \frac{\exp(z_i(\theta - \delta_i))}{\frac{\exp(0.5(\theta - \delta_i)) - 1}{\theta - \delta_i}}$$

That is the distribution of Z_i and X_i belong to the same family, with a different range for the values of the random variable. As a consequence we can use the same approach to split up Z_i into two new variables, and hence recursively turn the continuous response x_i into a set of conditionally independent Rasch response variables, with a discrimination that halves in every step of the recursion.

To better understand this construction, let's look a the second step in the recursion:

$$\begin{aligned} y_{i2} &= (z_i > 0.25) \\ &= \begin{cases} x_i > 0.75 & \text{if } y_{i1} = 1 \\ x_i > 0.25 & \text{if } y_{i1} = 0 \end{cases} \end{aligned}$$

We see that among responses that are correct ($Y_{i1} = 1$), those that are more correct are indicative of higher ability. Similarly, among responses that are incorrect ($Y_{i1} = 0$), those that are less incorrect are indicative of higher ability. These two indicators of

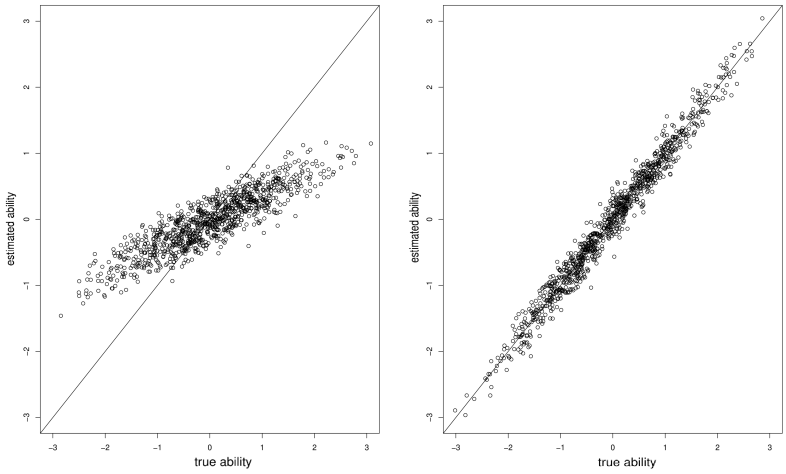


Figure 1. Parameter recovery for a linear test (left) and an adaptive one (right) based on minimizing cross information.

ability are under the model on the same footing, and statistically independent of the correctness of the response itself.

If we denote the binary response variable obtained in the j-th step of the recursion by Y_{ij}, we obtain the dyadic expansion of the continuous response variables into conditionally independent binary response variables, as depicted in Figure 2.

Another way to look as this is captured in the following formula:

$$\sum_{j=1}^{J} \frac{Y_{ij}}{2^j} \underset{J\to\infty}{\to} X_i$$

As 2^j tends to zero rapidly, it only takes a few of the binary response variables to closely approximate the original continuous response.

In the next section we'll make good use of this relation to develop a way to deal with statistical inference at scale for the proposed statistical model.

Method: Rating System for the Duolingo English Test

The Duolingo English Test being an adaptive online test with both a large number of test takers and a large number of items makes direct likelihood based inference a challenge, as such approaches don't scale very well.

Rating systems, such as the Elo (1978) rating system (originally developed for tracking ability in chess) are highly scalable, but come with their own shortcomings. The main shortcoming is that their statistical properties are not very well understood, making it difficult to assess standard errors or evaluate model fit. Brinkhuis and Maris (2019) provides a general introduction to tracking systems, and the minimal properties they should have.

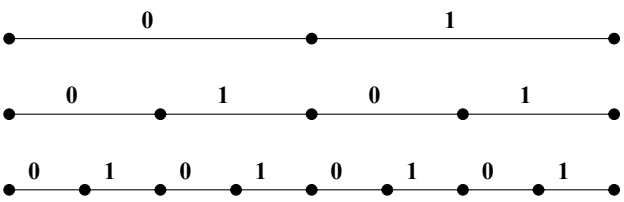
The urnings rating system provides an alternative. It's statistical properties are well understood and it is highly scalable (with person and item ratings being updated after every response). In equilibrium, urnings are known to be binomially distributed variables, with the urn size as a design parameter (similar to the K-factor in Elo ratings, and the logit of the probability being the ability/difficulty in the Rasch model).

In order to better understand the conceptual underpinnings of the urnings system, we consider the following game of chance where two players draw a ball from their own (infinite) urn containing a proportion π_p of green balls (the others being red) until they have drawn balls of different color. A simple derivation shows that the probability with which player p wins the game (i.e., ends up with a green ball, which we denote by $X_{pq} = 1$) is given by the Rasch model:

$$p(X_{pq} = 1|\theta_p, \theta_q) = \frac{\pi_p(1-\pi_q)}{\pi_p(1-\pi_q) + (1-\pi_p)\pi_q}$$

where $\theta_p = \ln(\pi_p/(1-\pi_p))$. Up to some technical details, the urnings rating system comprises of mimicking this game of chance with a finite sized urn and simply swapping the results. Urnings ratings are the number of green balls (U_p) in these finite sized urns.

Every person and item has an urn with red and green balls. Whenever a person responds to an item, a correct response is coded as a green ball for the person, and a red ball for the item, whereas incorrect responses are coded as a red ball for the person and a green ball for the item. From the person and item urns we draw a ball from each, with replacement, until they are of different colors. These are removed and replaced with the coded response balls. For technical reasons, to ensure that urnings are binomially distributed, with some (small) probability, the replacement does not take place.



$$P(Y_{i1} = 1|\theta) = \frac{\exp(0.5(\theta-\delta_i))}{1+\exp(0.5(\theta-\delta_i))}$$

$$P(Y_{i2} = 1|\theta) = \frac{\exp(0.25(\theta-\delta_i))}{1+\exp(0.25(\theta-\delta_i))}$$

$$P(Y_{i3} = 1|\theta) = \frac{\exp(0.125(\theta-\delta_i))}{1+\exp(0.125(\theta-\delta_i))}$$

Figure 2. (First three steps of a) dyadic expansion of continuous responses into conditionally independent binary response variables.

Figure 3 gives an overview of the urnings updating scheme, and the interested reader is referred to Maris, Bolsinova, Hofman, van der Maas, and Brinkhuis (2019) for the formal underpinning as to why this updating scheme produces binomially distributed ratings, when in equilibrium.

The technical details have to do with the match making probabilities, or the adaptive engine, which assigns probability $M_{pi}(\mathbf{u})$ to person p answering question i, as some function of their urnings, and an extra Metropolis-Hastings step needed to ensure we end up with the right invariant distribution.

The differences in discrimination that derive from the dyadic expansion of the continuous response variables in the SRT model translate into differences in the stakes of the game. Literally, when the stakes equal 4*, we continue drawing four balls from both urns until we get four green ones from the one urn and four red ones from the other. We once again just replace these four balls by four of the color consistent with the real item response. That is, a learner stands to loose or gain four balls based on their response to this particular item, which is why we refer to the discriminations as stakes in this context. Put differently, the higher the stakes, the larger the impact the item response will have on estimated ability.

Figure 4 highlights some of the key features of the urnings rating system, based on simulated data. The left panel of the figure demonstrates that the urnings rating system does what it is supposed to do. For every combination of urnings the fitted and observed proportions of correct responses are the same. The right side panel demonstrates that for a given urn size, we can set up a 95% coverage ellipse and get the theoretical guarantee that 95% of our data (combinations of true ability and urnings) are inside of this ellipse. This gives guarantees on overall reliability and local measurement precision.

Results

The Duolingo English Test comprises three parts, a regular CAT, a writing section, and a speaking section. The data are highly skewed in terms of the number of responses for unique items in the CAT section, with the vast majority of the items having only a handful of observations (< 10).

As the data are very skewed we cannot just estimate item difficulty parameters for the CAT items. Hence, for the purpose of analyzing the data new synthetic items were constructed by combining the sub-skill an item relates to with the (rounded) item level as provided by Duolingo. This gives rise to 55 "items". For the items in the CAT section we used Y_{i1} for the analyses.

The items in the writing and speaking section are automatically scored on a scale from 0 to 10. These are divided by 10, and we use the first three steps in the dyadic expansion for the analyses. The resulting stakes for these items are 4, 2, and 1, respectively. Hence, every response in these sections contributes 7 times as much as a response in the CAT section.

Figure 5 demonstrates that the Duolingo English Test responses fit the Rasch model very well. As the items are overall relatively easy, most of the data sits in the lower right quadrant, which explains why some of the empirical contours are smoother than others. Across the board, for all combinations of skill and difficulty, the fit is excellent. Since the expected proportions for items with different stakes are not easily displayed in one contour plot, the right side just plots the observed versus expected proportions of correct responses. The closer these are to the straight line, the better the model fits the data.

To evaluate whether the results depend on how the synthetic items were constructed, we repeated the analyses with double the amount of synthetic items. The basic outcome of the analyses is that for every observation the probability that it is correct is based on the current urnings of both the person and the item. Figure 6 replicates the main findings, based on these more fine grained synthetic items. As almost every point in the top panel of Figure 6 relates to a particular combination of person and item urnings, and not all of these combinations occur with equal frequency, we also look at rounded fitted values. The middle and bottom panel of Figure 6 provide the same information as in the top panel when fitted values are rounded to single or double digits. The middle panel of Figure 6 shows that whenever, based on urnings, the rounded probability of a correct response equals 0.7, say, the observed proportion of correct responses is in perfect agreement. Turning to the double digit rounding (bottom panel of Figure 6), we see some more scatter which is due to the smaller numbers of observations for double digit rounded values. Similarly, in the top panel of Figure 6 where almost every observation relates to a particular combination of person and item urnings, the scatter increases more due to smaller sample sizes.

*Remember that we can always multiply discriminations by some number, as long as we also divide the ability and difficulty parameters by the same number.

Match making: pair person p with item i with probability $M_{pi}(\mathbf{u})$

Reality:

repeat
$y_p^* \sim Bernoulli(\pi_p)$
$y_i^* \sim Bernoulli(\pi_i)$
until $y_p^* \neq y_i^*$
return (y_p^*, y_i^*)

Urnings:

repeat
$y_p \sim Bernoulli(U_p/n)$
$y_i \sim Bernoulli(U_i/n)$
until $y_p \neq y_i$
return (y_p, y_i)

Update:

$$u_p^* = u_p + y_p^* - y_p$$
$$u_i^* = u_i + y_i^* - y_i$$

Metropolis-Hastings: accept new urnings with probability:

$$\min\left(1, \frac{u_p(n-u_i)+(n-u_p)u_i}{u_p^*(n-u_i^*)+(n-u_p^*)u_i^*}\,\frac{M_{pi}(\mathbf{u}^*)}{M_{pi}(\mathbf{u})}\right)$$

Figure 3. Urnings rating system

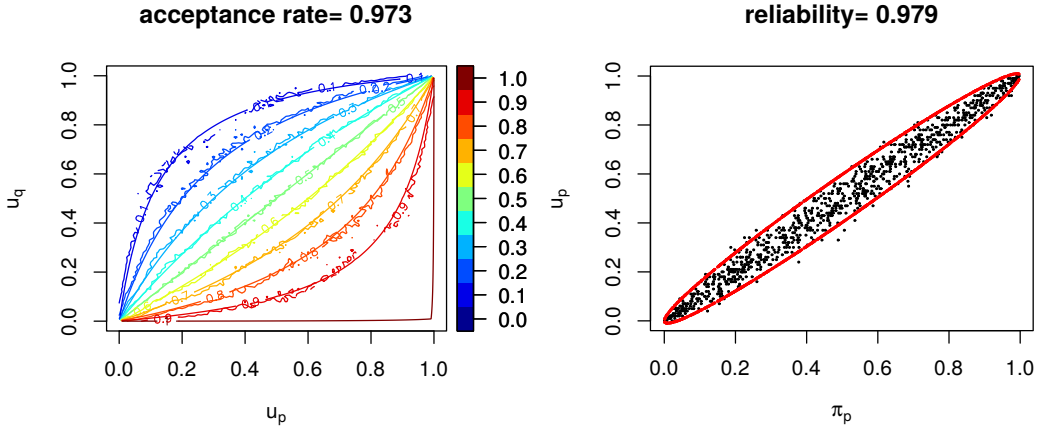


Figure 4. The left figure gives contours for the predicted (straight) and observed (squiggly) proportion of correct responses for every combination of urnings, with person urnings on the horizontal and item urnings on the vertical axis. The right figure gives the relation between true values and urnings. The red ellipse gives the 95% coverage ellipse (i.e., 95% of the combinations of true ability and urnings are inside of the ellipse). Both figures are based on a simulated example.

Logistic regression: Main effects only

The close fit between the data and the model already provides key evidence of there not being large violations of model fit. To evaluate whether there is substantial differential model fit with respect to the various item types and person characteristics, we ran a logistic regression with the binary response variables as the dependent variable and the logit of the fitted value[†] (logit(urn.fit) in the Tables), item type, gender, operating system, and screen size as independent variables. For items in the CAT part of the Duolingo English Test the first binary variable from the dyadic expansion was used, whereas for the speaking and writing section the first three were used. Table 1 gives the results of the logistic regression.

To put the results in Table 1 into perspective, the standard deviation of the logit fitted values (excluding those that are infinite) is 2.22. Maybe the most surprising result is that given the sample size, many effects are not significantly different from

[†] For every observation we compute the probability of it being correct using the current value for person and item urnings and the stakes.

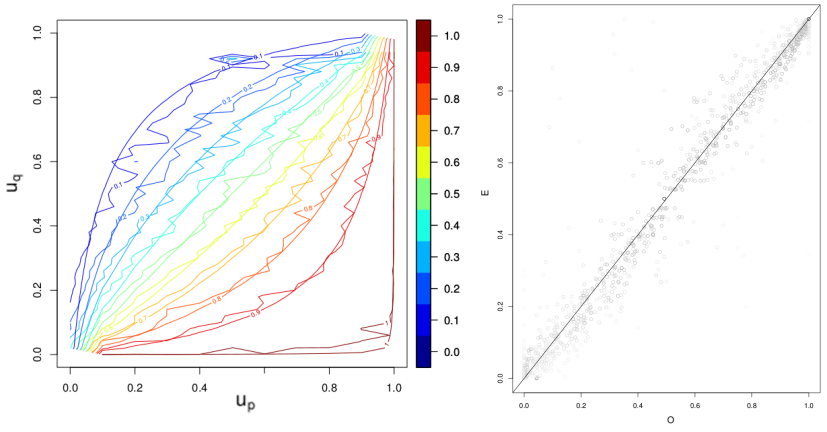


Figure 5. Left panel: contours for the fitted and observed proportion of correct responses for every combination of urnings for the CAT responses. Right panel: Observed versus fitted proportions of correct responses for every combination of urnings, for all responses.

zero. Even those that are are very small (in terms of logits). The only effect that stands out is for the writing-text item type, which with an effect of 0.335 is still not very large.

With the exception of a small effect for the writing-text item type, neither item type, nor operating system, nor screen size, nor geographic location add anything to urnings when it comes to explaining the observed responses. Two test takers with the same urnings, but one known to be a male test taker from Africa, on a windows machine with a small screen, the other a female, from Asia, on a macintel with a large screen, has no impact on how well we can explain their observed responses.

Logistic regression: Main effects and interactions

A final step in the analyses is looking at interactions between item type and various person covariates. The output of this analysis is split over a number of tables. As is typical for logistic regressions with interactions between nominal independent variables, interpreting the results is a bit tricky. Moreover, the independent variables are not truly independent, and some collinearity is likely to be present.

Fortunately, not a lot of interpretation is needed. For every observation, we get a fitted value from the logistic regression, and another one based on urnings alone. These correlate 0.999, which signifies that not a lot is happening, that is not accounted for by urnings alone. Put differently, adding in all of the main effects and interactions does not lead to a noticeable increase in model fit.

Discussion

The take away messages from these analyses are that a) the SRT model seems to fit the Duolingo English Test data well, and b) there is no compelling evidence of there being differential model fit for different groups of test takers, item types, or computers (screen size and operating system). From a construct validity point of view, the Duolingo English Test seems to be in good shape.

These findings do not preclude the existence of differential item functioning for a particular item. However, as every automatically generated item is only administered to a small number of test takers detecting, it would be difficult, and its impact would also be small.

The cross entropy based approach currently in use, at least for an adaptive test, seems to function quite well. The added value of being able to evaluate model fit, reliability, and standard errors that come with a model based approach seem worthwhile to pursue.

As the Duolingo English Test is intended to measure the various skills that together comprise language proficiency, it is comforting that the data support giving test takers a single composite score. However, for individual test takers there can be great value in getting information on their standing with respect to the various finer grained skills. Such information helps determining which skills they would need to improve on to increase their overall score.

Such information is worthwhile, and can be delivered by (an extended version of) the urnings rating system. As the Duolingo English Test is designed to measure English language proficiency, using a more fine grained diagnostic rating system could help learners in their preparation for the Duolingo English Test, and their becoming proficient in English.

The results reported here are a snapshot of the current state of affairs. With the population of test takers and the world around us changing, continuously monitoring (differential) model fit over time and across populations is important to ensure that scores remain comparable.

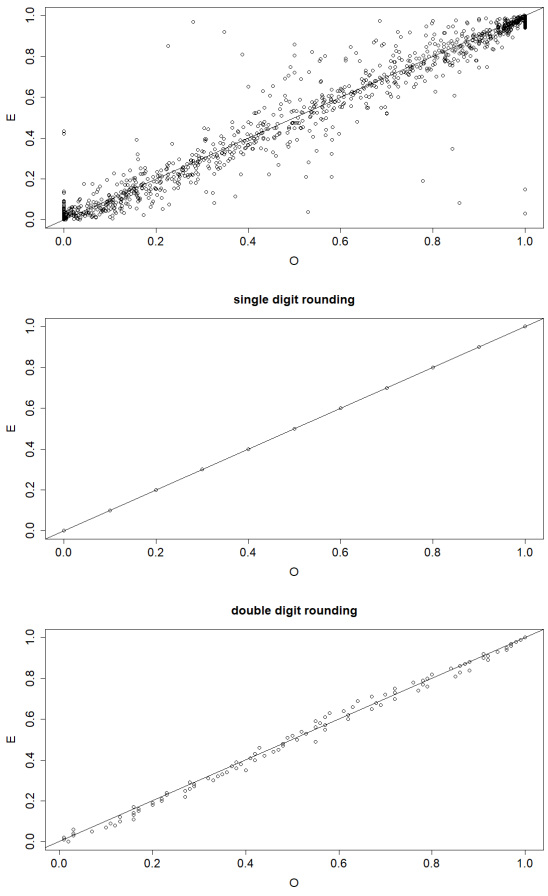


Figure 6. Top panel: Observed versus fitted proportions of correct responses for every combination of urnings for the more fine grained synthetic items. Middle panel: Observed versus fitted proportions of correct responses, rounded to a single digit for the more fine grained synthetic items. Bottom panel: Observed versus fitted proportions of correct responses, rounded to double digits for the more fine grained synthetic items

Table 1. Logistic Regression

Predictor	Effect	SE
Intercept	0.076^{***}	(0.022)
logit(urn.fit)	0.981^{***}	(0.004)
audiovocab		
ctest	0.034^{*}	(0.019)
speaking-image	-0.058^{***}	(0.019)
speaking-audio	-0.096^{***}	(0.017)
speaking-text	-0.063^{***}	(0.019)
dictation	-0.099^{***}	(0.019)
elicited speech	0.062^{*}	(0.036)
vocab	0.020	(0.023)
writing-image	-0.059^{***}	(0.019)
writing-text	0.335^{***}	(0.028)
windows		
android	0.017	(0.371)
linux	0.065	(0.053)
macintel	0.026^{**}	(0.011)
male		
gender unknown	0.018^{*}	(0.011)
female	0.007	(0.012)
other	0.140	(0.216)
Europe		
unknown	-0.035	(0.051)
Africa	-0.031	(0.027)
Americas	0.004	(0.017)
Asia	-0.038^{**}	(0.016)
Oceania	-0.055	(0.104)
< 800		
800-899	0.002	(0.013)
900-999	0.004	(0.013)
≥ 1000	0.020	(0.014)
Observations	$363,689$	
Log Likelihood	$-164,865.700$	
Akaike Inf. Crit.	$329,781.300$	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 2. Main Effects

Predictor	Effect	SE
Intercept	-0.094^{**}	(0.046)
logit(urn.fit)	0.987^{***}	(0.004)
< 17		
$17 - 25$	0.088^{**}	(0.043)
$26 - 45$	0.216^{***}	(0.047)
> 45	0.333^{***}	(0.086)
audiovocab		
ctest	-0.045	(0.060)
speaking-image	0.214^{***}	(0.061)
speaking-audio	0.119^{**}	(0.053)
speaking-text	0.212^{***}	(0.061)
dictation	0.017	(0.058)
elicited speech	0.546^{***}	(0.120)
vocab	-0.305^{***}	(0.067)
writing-image	0.339^{***}	(0.056)
writing-text	0.593^{***}	(0.086)
windows		
android	0.310	(1.096)
linux	0.047	(0.167)
macintel	0.108^{***}	(0.036)
male		
unknown	0.024^{**}	(0.011)
female	0.036^{***}	(0.012)
other	0.148	(0.225)
Europe		
region unknown	0.003	(0.051)
Africa	-0.078^{***}	(0.027)
Americas	-0.028	(0.017)
Asia	-0.062^{***}	(0.016)
Oceania	0.052	(0.105)
< 800		
$800 - 899$	0.101^{**}	(0.041)
$900 - 999$	0.017	(0.041)
≥ 1000	0.160^{***}	(0.045)

Table 3. Age by Item Type interactions

Predictor	Effect	SE
$17 - 25$: ctest	0.246^{***}	(0.062)
$26 - 45$: ctest	-0.155^{**}	(0.067)
> 45: ctest	-0.311^{***}	(0.120)
$17 - 25$: speaking-image	-0.147^{**}	(0.062)
$26 - 45$: speaking-image	-0.251^{***}	(0.067)
> 45: speaking-image	-0.182	(0.121)
$17 - 25$: speaking-audio	-0.117^{**}	(0.054)
$26 - 45$: speaking-audio	-0.186^{***}	(0.058)
> 45: speaking-audio	-0.266^{**}	(0.106)
$17 - 25$: speaking-text	-0.188^{***}	(0.062)
$26 - 45$: speaking-text	-0.263^{***}	(0.067)
> 45: speaking-text	-0.318^{***}	(0.120)
$17 - 25$: dictation	0.004	(0.059)
$26 - 45$: dictation	-0.220^{***}	(0.064)
> 45: dictation	-0.608^{***}	(0.116)
$17 - 25$: elicited speech	-0.245^{**}	(0.122)
$26 - 45$: elicited speech	-0.290^{**}	(0.130)
> 45: elicited speech	-0.531^{**}	(0.217)
$17 - 25$: vocab	0.502^{***}	(0.069)
$26 - 45$: vocab	0.621^{***}	(0.077)
> 45: vocab	0.733^{***}	(0.154)
$17 - 25$: writing-image	-0.240^{***}	(0.056)
$26 - 45$: writing-image	-0.532^{***}	(0.060)
> 45: writing-image	-0.549^{***}	(0.114)
$17 - 25$: writing-text	-0.148^{*}	(0.088)
$26 - 45$: writing-text	-0.365^{***}	(0.093)
> 45: writing-text	-0.641^{***}	(0.173)

Table 4. Platform by Item Type interactions

Predictor	Effect	SE
ctest: android	7.792	(52.138)
speaking-image: android	1.054	(2.396)
speaking-audio: android	−0.299	(1.426)
speaking-text: android	0.332	(1.688)
dictation: android	−1.364	(1.445)
elicited speech: android	7.398	(51.488)
vocab: android	7.687	(44.862)
writing-image: android	−0.661	(1.377)
writing-text: android		
ctest: linux	0.172	(0.248)
speaking-image: linux	−0.158	(0.234)
speaking-audio: linux	−0.212	(0.206)
speaking-text: linux	0.159	(0.237)
dictation: linux	0.101	(0.222)
elicited speech: linux	−0.248	(0.432)
vocab: linux	0.316	(0.316)
writing-image: linux	0.072	(0.213)
writing-text: linux	0.075	(0.344)
ctest: macintel	0.013	(0.052)
speaking-image: macintel	-0.130^{***}	(0.049)
speaking-audio: macintel	-0.128^{***}	(0.043)
speaking-text: macintel	-0.154^{***}	(0.049)
dictation: macintel	0.014	(0.047)
elicited speech: macintel	-0.714^{***}	(0.093)
vocab: macintel	0.006	(0.062)
writing-image: macintel	-0.201^{***}	(0.047)
writing-text: macintel	-0.157^{**}	(0.073)

Table 5. Screen Size by Item Type interactions

Predictor	Effect	SE
ctest: 800 − 899	−0.074	(0.059)
speaking-image: 800 − 899	−0.109*	(0.057)
speaking-audio: 800 − 899	−0.101**	(0.050)
speaking-text: 800 − 899	−0.093	(0.057)
dictation: 800 − 899	−0.047	(0.054)
elicited speech: 800 − 899	−0.079	(0.112)
vocab: 800 − 899	−0.151**	(0.071)
writing-image: 800 − 899	−0.022	(0.054)
writing-text: 800 − 899	−0.091	(0.083)
ctest: 900 − 999	−0.036	(0.059)
speaking-image: 900 − 999	−0.055	(0.057)
speaking-audio: 900 − 999	−0.029	(0.050)
speaking-text: 900 − 999	−0.067	(0.057)
dictation: 900 − 999	−0.039	(0.055)
elicited speech: 900 − 999	0.195*	(0.111)
vocab: 900 − 999	−0.083	(0.070)
writing-image: 900 − 999	0.050	(0.055)
writing-text: 900 − 999	0.160*	(0.085)
ctest: ≥ 1000	0.057	(0.065)
speaking-image: ≥ 1000	−0.193***	(0.062)
speaking-audio: ≥ 1000	−0.153***	(0.054)
speaking-text: ≥ 1000	−0.173***	(0.062)
dictation: ≥ 1000	−0.111*	(0.059)
elicited speech: ≥ 1000	−0.268**	(0.117)
vocab: ≥ 1000	−0.182**	(0.078)
writing-image: ≥ 1000	−0.083	(0.059)
writing-text: ≥ 1000	−0.090	(0.092)
Observations	357,384	
Log Likelihood	−161,330.900	
Akaike Inf. Crit.	322,877.800	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Author biography

Gunter Maris is ACTNext's director of advanced psychometrics. Prior to joining the ACTNext team, Gunter was a full professor of psychological methods at the University of Amsterdam (for 10 years), and a principal research scientist with CITO (for 16 years).

References

Brinkhuis, M. J., & Maris, G. (2019). Tracking ability: Defining trackers for measuring educational progress. In *Theoretical and practical advances in computer-based educational measurement* (pp. 161–173). Springer.

Council of Europe. (2001). *Common european framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Elo, A. E. (1978). *The rating of chess players, past and present*. London: B.T. Batsford, Ltd.

LaFlair, G. T., & Settles, B. (2019). *Duolingo English Test: Technical manual*. Duolingo. Pittsburgh, PA.

Maris, G., Bolsinova, M., Hofman, A., van der Maas, H., & Brinkhuis, M. (2019). Urnings: A rating system. *in preparation*.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*(4), 615–633.

Settles, B., LaFlair, G., & Hagiwara, M. (in press). Machine learning driven language assessment. *Transactions of the Association for Computational Linguistics*.