Duolingo English Test: Subscores

Duolingo Research Report DRR-20-03
October 22, 2020 (16 pages)
englishtest.duolingo.com/research

Geoffrey T. LaFlair*

Abstract

The Duolingo English Test is a computer adaptive test that provides an overall score that represents test taker English language proficiency. However, stakeholders, such as university admissions officers, often want to make decisions based on test taker ability in one or more components of language ability, such as speaking. Similar to overall scores, subscores should meet standards of reliability when used for decision making. In addition, subscores should provide distinct information about the test takers' abilities above and beyond the overall score. In this paper, we report on the research behind four subscores reported by the Duolingo English Test (Literacy, Conversation, Comprehension, and Production) that can be used by stakeholders to make decisions about test takers.

*Duolingo, Inc.

Corresponding author:
Geoffrey T. LaFlair, PhD
Duolingo, Inc. 5900 Penn Ave, Pittsburgh, PA 15206, USA
Email: englishtest-research@duolingo.com

© 2020 Duolingo, Inc. All rights Reserved.

1 Introduction

The Duolingo English Test is a large-scale, high-stakes, computer-adaptive, online test of English language proficiency that can be delivered anywhere and at any time. The test is used by over 2,000 institutions to make admissions decisions at the undergraduate or graduate level, placement decisions into university English language programs, or exit decisions from English language programs. Since its creation, a total score, representing English language ability, has been reported to test users. As adoption of the test for admissions purposes by English-medium universities around the world has grown, so has interest from our stakeholders in additional information about language ability: subscores.

The purpose of subscores in a language test is to provide additional information about test taker abilities to use language at a finer grain size. For example, many large-scale English language assessments report subscores as test-taker ability to speak, write, read, and listen, all of which are skills that are important in a number of situations, including university study. Recent research in the field of language testing has underscored the importance of including subscores when making admissions decisions. Ginther & Yan (2018) and Bridgeman, Cho, & DiPietro (2016) both used TOEFL iBT subscores to evaluate the relationship between test taker ability in sub-domains of language and performance (as measured by GPA) in university study. Ginther & Yan (2018) found that a jagged profile (strong in one or two components of language and weak in other components) can have an effect on international students' university success. Additionally, Bridgeman et al. (2016) showed that the relative importance of different components of language ability may vary by field of study, which further supports the use of subscores when admitting people into degree programs. The results of these studies highlight the added value that subscores can provide about test-taker language ability beyond a total score.

At its inception, the Duolingo English Test was designed to be an overall measure of general English language proficiency (Settles, LaFlair, & Hagiwara, 2020). The initial items included on the test (i.e., the computer-adaptive items) reflect this decision (see Table 1). The items (i.e., Textvocab, Audiovocab, Ctest, Dictation, and Elicited speech) measure different aspects of language ability. For example, the Ctest items measure reading ability, and the yes/no vocabulary items (Textvocab and Audiovocab) measure vocabulary size. Additionally, they are highly predictive of other aspects of language ability. For example, in addition to being a known measure of vocabulary size, yes/no vocabulary items are predictive of reading ability (McLean, Stewart, & Batty, 2020; Milton et al., 2010; Staehr, 2008). In July 2019, measures of open-ended speaking and writing tasks were added to the scored portion of the test in order to increase the construct coverage of the assessment and support research into providing subscores. Between July 2019 and July 2020, scores on these items have been used to report a total score representing English language proficiency.

© 2020 Duolingo, Inc

Table 1. Duolingo English Test Item Types

Item format	Skills	Reference
T.vocab	L, R, W	Staehr (2008); Milton (2010); Zimmerman, Broder, Shaughnessy, & Underwood (1977)
A.vocab	L, S	Milton et al. (2010); Milton (2010)
Ctest	R, W	Klein-Braley (1997); Khodadady (2014); Reichert, Keller, & Martin (2010)
Dictation	L, W	Bradlow & Bent (2002); Bradlow & Bent (2008)
E.speech	R, S	Vinther (2002); Jessop, Suzuki, & Tomita (2007)
Speaking	S	Luoma (2004)
Writing	W	Cushing-Weigle (2002)

Similar to providing a total score, subscores should meet a set of standards before being reported. They should represent the underlying structure of the test, be reliable, and have added measurement value beyond the total test score (Haberman, 2008; Sawaki & Sinharay, 2013; Thissen & Wainer, 2001). Common techniques for evaluating the internal structure of a test include exploratory and confirmatory factor analysis, and multidimensional scaling (MDS). For example, confirmatory factor analysis has been used to show that the structure of the TOEFL iBT can be represented by a general factor of language ability and four group factors that represent speaking, writing, reading, and listening abilities (see Sawaki & Sinharay, 2013; Sawaki, Stricker, & Oranje, 2008). The dimensionality of other language assessments has also been investigated using MDS techniques. Chalhoub-Deville (1995) employed MDS to uncover the underlying dimensions in a test of L2 speaking ability that included measures of interaction (interview), narration, and reading aloud. MDS has also been used to investigate the underlying dimensions of an English language test that included measures of grammatical and vocabulary knowledge and reading ability (Hoyazin, 1986).

While MDS and factor analysis share a similar goal of investigating dimensionality, Ding (2018) discusses two advantages of MDS. First, it can represent linear and non-linear relationships. Second, it can accomodate a wider variety of data types. After establishing the underlying structure of a test and creating subscores based on that structure, it is necessary to investigate their reliability and added measurement value beyond the total score. The reliability and added value of subscores can be evaluated using approaches from classical test theory (Choi & Papageorgiou, 2020; Haberman, 2008; Sawaki & Sinharay, 2013; Sinharay, Puhan, & Haberman, 2011). For reliability, this means evaluating the internal consistency and test-retest reliability of the subscores. The most common CTT approach to measuring the added value of subscores compares proportional reduction in mean squared error (PRMSE) of the subscore with the PRMSE of the total score—the amount of variance in a true subscore that is accounted for by the observed subscore with the amount of variance that is accounted for by the observed total score (Feinberg & Jurich, 2017; Sinharay et al., 2011).

© 2020 Duolingo, Inc

In conducting research into subscores, three research questions were answered:

1. What is the underlying structure of the Duolingo English Test?
2. What is the reliability of the subscores?
3. Do the subscores have added value?

2 Method

2.1 Data

The data come from 101,604 tests that were administered between May 1, 2019 and June 30, 2020. The dimensionality analyses were conducted on the full set of data. However, the reliability analyses were conducted on subsets of the larger data set. To create the data set for the internal consistency analysis, the speaking and writing tasks, which are scored at the portfolio level operationally, were split apart and scored independently. This produced scores for four writing tasks and four speaking tasks that were then randomly assigned to either side of the split tests. The data for the internal consistency analysis contains 10,218 observations. The test-retest reliability coefficient was calculated on 10,306 pairs of tests in which the same test taker took the Duolingo English Test twice in a 30-day time period.

2.2 Test Administration

The majority of the test administration is fully adaptive, which means that the selection of each subsequent item depends on the test taker's performance on previous items*. The first four items administered during every test comprise the "burn-in" phase. During this phase, every test taker receives randomly sampled items of increasing difficulty. After the first four items the adaptive algorithm creates a provisional estimate of ability upon which the next item is chosen. Further into the test administration, subsequent items are selected based on item type. This constraint in the test administration ensures that the representation of the seven different item types is balanced across all administrations.

2.3 Analyses

To evaluate the internal structure, we used ordinal, non-metric multi-dimensional scaling (MDS). Multi-dimensional scaling is a data reduction technique that can be used to

*The portion of the test that is semi-adaptive is the writing and speaking prompts, the difficulty of which is based on the provisional estimate of ability of the test taker. However, the subsequent selection of the next item is not based on the test taker's performance on these items.

© 2020 Duolingo, Inc

explore the underlying structure, or latent structure, of assessments (Davison & Skay, 1991; Ding, 2018). The `smacof` R package was used to carry out the analysis (de Leeuw & Mair, 2009) and reduce the seven different item types to interpretable dimensions that also felicitously represented the relationship in the dissimilarity matrix. The dissimilarity measure that was employed in the analysis was the correlation among the item types transformed to Euclidean distances (see Equation (1), where *r* is the correlation coefficient between two item types and *d(x, y)* is the dissimilary betweeen two item types) (Ding, 2018). Davison & Sireci (2000) recommend that the selection of a solution for MDS analyses be based on two criteria: 1) that the stress value not exceed 0.10 and 2) that the results are interpretable. The stress value is an indicator of how well the distances among the items in the solution represent their actual (dis)similarity that the analysis is based on, with zero being perfect. For interpretability, it is recommended that the solution is parsimonious (i.e., fewer dimensions are preferred) and that the relationships among the items can be understood in the context of the intended construct being measured.

$$d(x, y) = \sqrt{1 - r} \tag{1}$$

Additionally, an exploratory factor analysis (EFA) was conducted on the same data that was used for the MDS analysis. This was done to corroborate the MDS analysis because Davison & Sireci (2000) argue that the MDS results reflect profile patterns in a population, whereas factor analytic results reflect latent traits. Parallel analysis was used to determine the number of factors to extract (Zwick & Velicer, 1986). However, because its accuracy is affected by sample sizes greater than 1,000 (Revelle, 2015), we conducted 1,000 parallel analyses on randomly sampled data sets with 600 observations, which is a large sample that still maintains accuracy (Green, Redell, Thompson, & Levy, 2016), and then used the median number of recommended factors across the 1,000 analyses to select the number of factors to extract. Then the factor analysis was conducted on the full data set.

The Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) was 0.90, which was acceptable for continuing with the factor analysis. The median number of factors extracted from the parallel analyses was three. We used maximum likelihood estimation and no rotation in the factor analysis. The decision not to rotate was based on examples (Davison, 1981; Davison & Skay, 1991) that illustrate that there is a relationship between the dimensions of MDS analyses and the specific factors (i.e., factors that are uncovered after a the first general factor). The MDS solution tends to ignore the general factor and focus on the secondary factors. Traditional rules of thumb for retaining variables on factors recommend that loadings be greater than 0.30–0.40 when assigning variables to a single factor (Loewen & Gonulal, 2015). However, it is expected that the Duolingo English Test items will cross-load onto all three factors. As a result we follow the 0.40-0.30-0.20 recommendation in Howard (2016). This split recommends that the test items load on the primary factor (general language ability) above 0.40, on alternative factors below 0.30, and show a minimum difference of 0.20 between their primary and alternative loadings.

© 2020 Duolingo, Inc

The reliability of the subscores was evaluated using classical test theory (CTT) techniques. First, split-half reliability was used to estimate the internal consistency of the Duolingo English Test. The random assignment of items was conducted such that each side of the split-half tests was representative of the full test. In other words, each side had equal representation of the different item types that are on the full test. We used Pearson's r to estimate the correlation between the two halves and the Spearman-Brown prophecy method to adjust the correlation coefficient so that it represented the internal consistency of the full test. Pearson's r[†] was also used to estimate the correlation between time one tests and time two tests in the test-retest reliability analysis (Bachman, 2004).

To examine the extent to which the subscores add distinct, interpretable information above and beyond the total score, we followed the proportional reduction in mean squared error (PRMSE) procedures recommended by Haberman (2008) and Sinharay et al. (2011). This approach compares the amount of variation in the subscore that is attributable to the items that are a part of that subscore (essentially the subscore's reliability) with the amount of variation in the subscore that is attributable to the total score. Here these are denoted $\text{PRMSE}_{\text{SUB}}$ (subscore PRMSE) and $\text{PRMSE}_{\text{TOT}}$ (total PRMSE). Additionally, we follow recommendations from Feinberg & Jurich (2017) that the value-added ratio (VAR) of the subscores meet a minimum threshold of 1.1. This ratio is defined as $\text{PRMSE}_{\text{SUB}}$ over $\text{PRMSE}_{\text{TOT}}$.

Post-hoc analysis

The relationship between the new Duolingo English Test subscores and subscores from other language tests (TOEFL iBT and IELTS) were examined in a set of post-hoc analyses. This relationship is estimated using Pearson's r on self-reported subscore data from our test takers. This data is gathered at the end of each test session. Prior to estimating the relationship, the data is cleaned by transforming the subscores under comparison to z-scores and removing any pairwise comparisons that are greater than three standard deviation units apart.

3 Results

3.1 Internal Structure

The results of the multidimensional scaling are shown in Figure 1 (the scale scores are shown in Table 2). A two dimension solution was selected because of its acceptable stress value (0.02) and its interpretability. By examining Figure 1, it can be seen that the items group together on either half of both dimensions. On Dimension 1, there is

[†]This coefficient is bound between -1 and 1 with values closer to one being expected and desirable.

© 2020 Duolingo, Inc

a group of items that measure aspects of language Comprehension. For example, the Ctest and Espeech items measure reading ability and the Dictation items measure listening ability. Additionally, the two sets of vocabulary items provide a measure of vocabulary size, which is predictive of ability to read and listen. On the other end of this dimension is Production. The open-ended Writing and Speaking tasks are measures of the test taker abilities to produce language. The positive end of the second dimension contains items that measure test-taker Literacy skills. This subscore comprises the Ctest, Textvocab, and Writing items. On the other side of this dimension, items that measure skills related to Conversation are grouped together. This subscore is composed of the Speaking, Dictation, Espeech, and Audiovocab items.

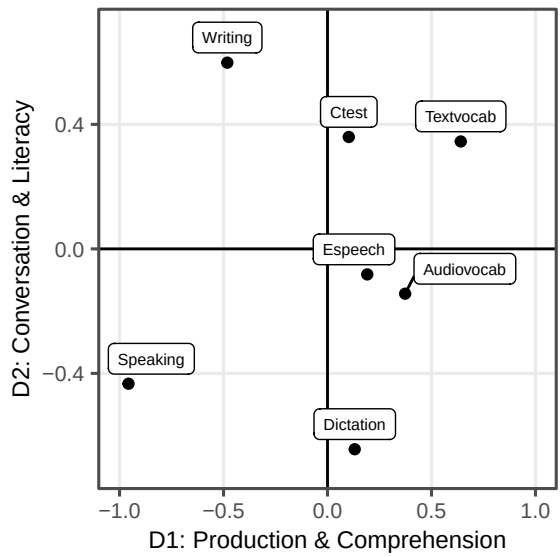

Figure 1. Duolingo English Test questions in two dimensions

Table 2. MDS Scale Scores in Two Dimensions (n = 101,604)

Item format	D1	D2
Audiovocab	0.37	-0.14
Ctest	0.10	0.36
Dictation	0.13	-0.64
Espeech	0.19	-0.08
Textvocab	0.64	0.35
Speaking	-0.96	-0.43
Writing	-0.48	0.60

© 2020 Duolingo, Inc

The three-factor solution supports the results of the MDS analysis (see Table 3). The majority of the loadings met the 0.40-0.30-0.20 criteria. As expected, there is a strong general factor representing English language proficiency, and all items load on the general factor above 0.40. There are two alternative factors. Four test items (Audiovocab, Dictation, Espeech, and Textvocab) load higher than 0.30 on one alternative factor. All of the differences between the primary loading and the alternative loadings are greater than 0.20, with the exception of Dictation on the *Lit/Conv* dimension. The Ctest items switched signs on the third factor (the analogue to the second dimension in the MDS). While there is one discrepancy between the MDS and the EFA solution, the EFA solution sufficiently corroborates the MDS analysis.

Table 3. Three Factor Solution (n = 101,604)

Item format	General factor	Comp/Prod	Lit/Conv
Audiovocab	0.7	0.22	0.34
Ctest	0.75	0.09	0.28
Dictation	0.61	0.09	0.46
Espeech	0.8	0.2	0.54
Textvocab	0.89	0.45	-0.07
Speaking	0.56	-0.04	0.25
Writing	0.87	-0.49	-0.02

3.2 Subscore Summary

The archived tests were "re-scored" to create the subscores (see Table 4). The Production score tends to be the lowest with a median score of 80, and Comprehension tends to be the easiest with a median score of 115. Conversation and Literacy fall in the middle as they require test takers to demonstrate both the ability to understand and produce language when responding to the items that comprise those subscores.

Table 4. Subscore Summary Statistics (n = 101,604)

Subscores	Mean	SD	Min	P25	Median	P75	Max
Comprehension	112.74	19.42	20	100	115	125	150
Conversation	95.05	21.37	15	80	95	110	155
Literacy	104.95	19.18	15	95	105	120	155
Production	82.43	21.72	10	70	80	95	155

© 2020 Duolingo, Inc

3.3 Subscore Reliability

Two indices of reliability were considered: internal consistency and test-retest reliability. The results in Table 5 show that the subscores are reliable, with the least reliable subscore being Production. This lower estimate is likely due to the application of the ML scoring model on the individual prompts rather than the portfolio—the algorithm was trained on portfolios. A corroborating piece of evidence for this is a coefficient α analysis of the speaking and writing scores, which yields a coefficient of 0.86.

Table 5. Subscore Internal Consistency (n = 10,218) and Test-retest Reliability (n = 10,306)

Subscore	Internal consistency	Test-retest
Literacy	0.89	0.80
Conversation	0.93	0.77
Comprehension	0.95	0.76
Production	0.76	0.81

3.4 Subscore Added Value

The results of the PRMSE show that all of the $\text{PRMSE}_{\text{SUB}}$ values are larger than the $\text{PRMSE}_{\text{TOT}}$ values (see Table 6). Additionally, the value-added ratio (VAR; the ratio of $\text{PRMSE}_{\text{SUB}}$ over $\text{PRMSE}_{\text{TOT}}$) for each subscore meets the Feinberg & Jurich (2017) recommendation that it be at minimum 1.1 (Comprehension is rounded up). As a result, we conclude that the subscores add distinct interpretable information above and beyond the total score.

Table 6. Subscore Added Measurement Value

Subscore	$\text{PRMSE}_{\text{SUB}}$	$\text{PRMSE}_{\text{TOT}}$	VAR
Literacy	0.89	0.77	1.16
Conversation	0.93	0.75	1.24
Comprehension	0.95	0.89	1.07
Production	0.76	0.43	1.77

© 2020 Duolingo, Inc

3.5 Relationship to Other Tests

As a post-hoc analysis of the subscores, the relationship between the Duolingo English Test subscores and the subscores from TOEFL (see Figure 2) and IELTS (see Figure 3) was examined. With the differences in configuration between integrated modalities (i.e., Literacy, Conversation, Comprehension, and Production) and four skills (i.e., Speaking, Writing, Reading, Listening), we expect this relationship to be moderate to strong (e.g., 0.40-0.70) and positive. All of the comparisons achieve this range with the lowest being the relationship between the Duolingo English Test Production scores and IELTS Writing and Speaking scores and the Literacy score and IELTS Reading scores. The correlation between Duolingo English Test subscores and TOEFL subscores all range between 0.59 and 0.69.

4 Discussion

Three primary analyses and one post-hoc analysis were conducted to investigate the utility of subscores for the Duolingo English Test. The answer to the first question, *What is the underlying structure of the Duolingo English Test?*, is that the underlying structure, as illustrated by the MDS analysis, is best represented from the perspective of integrated modalities. This result is corroborated by the EFA, which also showed a strong general factor in addition to the two secondary factors. The items tend to place themselves along dimensions that can be interpreted as measuring skills that are important to Literacy, Conversation, Comprehension, and Production. This underlying structure is congruent with the integrated skills perspective on language teaching and learning. In this approach, courses are designed in a manner that reflects the importance of the relationship between language skills at various "grain sizes", with speaking, writing, reading, and listening being "coarser grain" and vocabulary and grammatical skills/knowledge being "finer grain" (Hinkel, 2006, 2010; Widdowson, 1978). The second question was *What is the reliability of the subscores?*. The results of the analyses for the second question showed that the reliability of the four subscores is acceptable; they provide sufficiently consistent scores for decision making. The answer to the third question, *Do the subscores have added value?*, is "yes". The values of $PRMSE_{SUB}$ for each subscore were larger than the corresponding value of $PRMSE_{TOT}$, and the value-added ratio (VAR) threshold of 1.1 was met for each subscore as well.

This study provides quantitative evidence that supports the Duolingo English Test subscores. However, additional research into these subscores and their uses will strengthen the validity evidence for their interpretations and uses. Future directions for this research include investigating the three same criteria (internal structure, reliability, and added value) for different strata of test taker proficiency (e.g., beginner, intermediate, advanced) as well as by L1 (to ensure that the internal structure is invariant across test-taker L1s). Additionally, now that the internal structure has been established through exploratory dimensionality analyses, confirmatory methods, such as bifactor modeling,

© 2020 Duolingo, Inc

could provide additional support. Finally, research on the extent to which these subscores could help stakeholders understand how to set admissions thresholds for the subscores could provide further evidence that supports their uses in admissions decisions.

© 2020 Duolingo, Inc

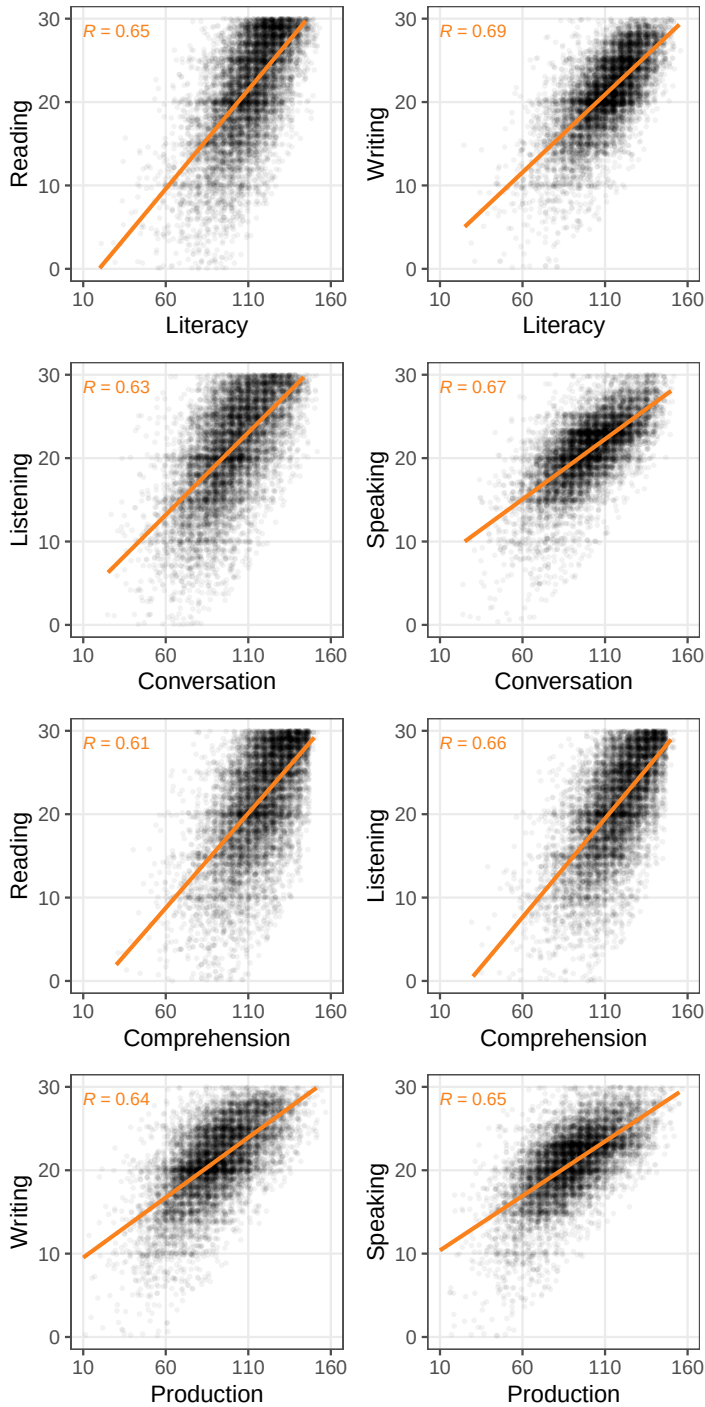

Figure 2. Relationship to TOEFL subscores

© 2020 Duolingo, Inc

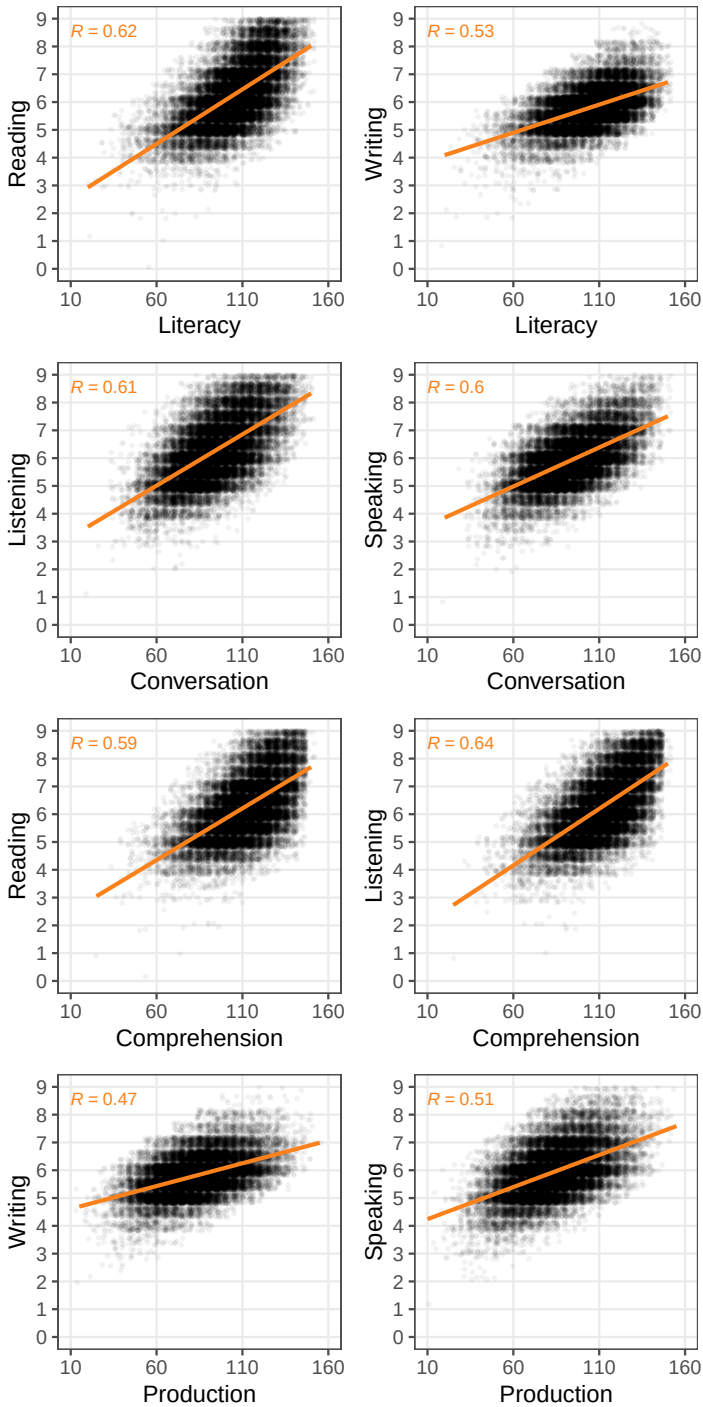

Figure 3. Relationship to IELTS subscores

© 2020 Duolingo, Inc

References

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America*, *112*, 272–284.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*, 707–729.

Bridgeman, B., Cho, Y., & DiPietro, S. (2016). Predicting grades from an English language assessment: The importance of peeling the onion. *Language Testing*, *33*(3), 307–318. https://doi.org/10.1177/0265532215583066

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing*, *12*(1), 16–33. https://doi.org/10.1177/026553229501200102

Choi, I., & Papageorgiou, S. (2020). Evaluating subscore uses across multiple levels: A case of reading and listening subscores for young efl learners. *Language Testing*, *37*(2), 254–279. https://doi.org/10.1177/0265532219879654

Cushing-Weigle, S. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Davison, M. L. (1981). Multidimensional scaling vs. Factor analysis of tests and items. *Annual meeting of the american psychological association*, 1–23. Retrieved from https://files.eric.ed.gov/fulltext/ED211580.pdf

Davison, M. L., & Sireci, S. G. (2000). Multidimensional scaling. In *Handbook of applied multivariate statistics and mathematical modeling* (pp. 323–352). Elsevier.

Davison, M. L., & Skay, C. L. (1991). Multidimensional scaling and factor models of test and item responses. *Psychological Bulletin*, *3*(110), 551–556. https://doi.org/https://doi.org/10.1037/0033-2909.110.3.551

de Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software*, *31*(3), 1–30. Retrieved from http://www.jstatsoft.org/v31/i03/

Ding, C. S. (2018). *Fundamentals of applied multidimensional scaling for educational and psychological research*. Springer.

Feinberg, R. A., & Jurich, D. P. (2017). Guidelines for interpreting and reporting subscores. *Educational Measurement: Issues and Practice*, *36*(1), 5–13. https://doi.org/10.1111/emip.12142

Ginther, A., & Yan, X. (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing*, *35*(2), 271–295. https://doi.org/10.1177/0265532217704010

© 2020 Duolingo, Inc

Green, S. B., Redell, N., Thompson, M. S., & Levy, R. (2016). Accuracy of revised and traditional parallel analyses for assessing dimensionality with binary data. *Educational and Psychological Measurement*, *76*(1), 5–21.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204–229. https://doi.org/10.3102/1076998607302636

Hinkel, E. (2006). Current perspectives on teaching the four skills. *TESOL Quarterly*, *40*(1), 109–131.

Hinkel, E. (2010). Integrating the four skills: Current and historical perspectives. In R. B. Kaplan (Ed.), *Oxford handbook in applied linguistics* (pp. 323–352). Oxford: Oxford University Press.

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, *32*(1), 51–62.

Hoyazin, R. (1986). The graphic representation of language competence: Mapping EFL proficiency using a multidimensional scaling technique. *Language testing: Selected papers from the colloquium*, 39–59. Retrieved from https://files.eric.ed.gov/fulltext/ED287286.pdf

Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review*, *64*(1), 215–238.

Khodadady, E. (2014). Construct validity of C-tests: A factorial approach. *Journal of Language Teaching and Research*, *5*(6), 1353–1362.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*(1), 47–84.

Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 182–212). Routledge New York, NY.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting l2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *OnlineFirst*. https://doi.org/10.1177/0265532219898380

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 211–232). EuroSLA.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (Vol. 52, pp. 83–98). Multilingual Matters.

© 2020 Duolingo, Inc

Reichert, M., Keller, U., & Martin, R. (2010). The C-test, the TCF and the CEFR: A validation study. In R. Grotjahn (Ed.), *The c-test: Contributions from current research* (pp. 205–231). Peter Lang.

Revelle, W. (2015). *Psych: Procedures for psychological, psychometric, and personality research (Version 1.5.1) [Computer software]*. Retrieved from http://CRAN.R-project.org/package=psych

Sawaki, Y., & Sinharay, S. (2013). *Investigating the value of section scores for the TOEFL iBT test* (ETS Research Report No. 13-35).

Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL internet-based test iBT: Exploration in a field trial sample* (ETS Research Report No. 08-09).

Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, *8*, 247–263. https://doi.org/10.1162/tacl/_a/_00310

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An ncme instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*(3), 29–40. https://doi.org/10.1111/j.1745-3992.2011.00208.x

Staehr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*, 139–152.

Thissen, D. E., & Wainer, H. E. (2001). *Test scoring.* Lawrence Erlbaum Associates Publishers.

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*(1), 54–73.

Widdowson, H. G. (1978). *FTeaching language as communication*. Oxford: Oxford University Press.

Zimmerman, J., Broder, P. K., Shaughnessy, J. J., & Underwood, B. J. (1977). A recognition test of vocabulary using signal-detection measures, and some correlates of word and nonword recognition. *Intelligence*, *1*(1), 5–31.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432–442.

© 2020 Duolingo, Inc