
Article:
Isaacs, Talia, Hu, Ruolin, Trenkic, Danijela orcid.org/0000-0001-6340-6030 et al. (1 more author) (2023) Examining the predictive validity of the Duolingo English Test:Evidence from a major UK university. Language Testing. pp. 748-770. ISSN 0265-5322

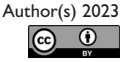
https://doi.org/10.1177/02655322231158550

Article

LANGUAGE TESTING

Examining the predictive validity of the Duolingo English Test: Evidence from a major UK university

Language Testing
2023, Vol. 40(3) 748–770



DOI: 10.1177/02655322231158550
journals.sagepub.com/home/ltj


Talia Isaacs
University College London, UK

Ruolin Hu
University College London, UK

Danijela Trenkic
University of York, UK

Julia Varga
University College London, UK

Abstract

The COVID-19 pandemic has changed the university admissions and proficiency testing landscape. One change has been the meteoric rise in use of the fully automated Duolingo English Test (DET) for university entrance purposes, offering test-takers a cheaper, shorter, accessible alternative. This rapid response study is the first to investigate the predictive value of DET test scores in relation to university students' academic attainment, taking into account students' degree level, academic discipline, and nationality. We also compared DET test-takers' academic performance with that of students admitted using traditional proficiency tests. Credit-weighted first-year academic grades of 1881 DET test-takers (1389 postgraduate, 492 undergraduate) enrolled at a large, research-intensive London university in Autumn 2020 were positively associated with DET Overall scores for postgraduate students (adj. $r = .195$) but not undergraduate students (adj. $r = -.112$). This result was mirrored in correlational patterns for students admitted through IELTS ($n = 2651$) and TOEFL iBT ($n = 436$), contributing to criterion-related validity evidence. Students admitted with DET enjoyed lower academic success than the IELTS and TOEFL iBT test-takers, although sample characteristics may have shaped this finding. We discuss implications

Corresponding author:
Talia Isaacs, UCL Centre for Applied Linguistics, IOE–UCL's Faculty of Education and Society, University College London, 20 Bedford Way, London WC1H 0AL, UK.
Email: talia.isaacs@ucl.ac.uk

for establishing cut scores and harnessing test-takers' academic language development through pre-sessional and in-sessional support.

Keywords
Academic attainment, academic readiness, English for Academic Purposes, higher education, language proficiency, predictive validity, second language, standardised tests, university admissions, validation

Introduction

The COVID-19 pandemic has changed the university admissions and English language proficiency (ELP) testing landscape (Ockey, 2021). One tangible change has been the meteoric rise in use of Duolingo English Test (DET) for high-stakes university entrance purposes, with rapid and widespread uptake of the test following test centre closures in 2020 (Isbell & Kremmel, 2020). During that year, popular international student destinations that rely heavily on student tuition fees to offset budget shortfalls (e.g., United Kingdom—the setting for the current study; Bolton & Hubble, 2021) were under pressure to offer some form of alternative to in-person testing as a mechanism for allowing students to continue to provide proof of ELP during lockdown or restricted movement conditions. Traditional ELP testing companies with longstanding inroads in the higher education market moved quickly to adapt to the changing circumstances through launching home versions of their established tests (Clark et al., 2021; Papageorgiou & Manna, 2021). In contrast, DET, a relative newcomer in the ELP testing market, was, from its inception, designed with the flexibility to be administered at any time and location, given access to a computer and Internet connection (LaFlair et al., 2022). It, therefore, had the necessary online infrastructure and test security measures already in place for home administration (i.e., remote proctoring), helping it fill a vacuum, particularly before industry competitors' at-home versions of existing tests had been brought to market. DET also offered university applicants a cheaper, shorter, more accessible option than pre-COVID-19 standardised ELP tests (e.g., IELTS, TOEFL iBT), making it an attractive alternative.

As a computer-adaptive test, DET offers fast, reliable results using proprietary machine scoring algorithms (LaFlair et al., 2022). Since its rapid and widespread uptake early in the pandemic, competing testing organisations, in a possible effort to retain market competitivity, have since shortened their existing tests (Pearson PTE, n.d.) or introduced a new, cheaper test with more varied item types, some of which parallel DET's (Papageorgiou et al., 2021). In sum, DET's prominence during the pandemic has changed the language testing world irretractably. It is likely to remain an attractive option for test-takers and their target higher education destinations in the years ahead.

DET is a fully automated (i.e., machine administered and scored) Internet-based test. It follows that the way that DET operationalises the ELP construct (see Cardwell et al., 2022) is markedly different than the approach of traditional, human-scored ELP tests. This includes using mostly controlled item types to elicit predictable test-taker output that the machine can easily recognise and automatically score. To elaborate, for fully

automated high-stakes tests, including DET, test-takers' outputs tend to be more predictable than in integrated tasks requiring information synthesis from multiple sources (e.g., Frost et al., 2012) or dialogic tasks with an interlocutor, including those that simulate non-test language use situations. As Khabbazbashi et al. (2021) note, concerns about a limited communicative orientation and narrowing of the ELP construct are common to contemporary high-stakes fully automated tests due to the state of the technology. For example, when considering a broad construct of L2-speaking (Lim, 2018), linguistic factors such as pragmatic appropriateness, pitch-related aspects of intonation, and interactional features tend to be difficult for machine algorithms to capture (Isaacs, 2018). Although there will always be limitations in what machine scoring alone can do (e.g., Nakatsuhara & Berry, 2021), technological advances unveil new possibilities, making it possible for fully automated tests to evolve. For example, DET introduced constructed-response item types for speaking and writing in its current version (Isbell & Kremmel, 2020), representing a stronger emphasis on complementing receptive skills assessment with more spontaneous but contained spoken and written productions. Interactive reading, which involves test-taker engagement with written text, is the most recent item type to have been incorporated into the test at the time of writing this manuscript (Attali et al., 2022). The introduction of these new item types has signalled movement towards a more communicative orientation of the test within the limits of technology. It should be noted that existing reviews of the test (Wagner, 2020; Wagner & Kunnan, 2015) were published before these new DET item types had been introduced, meaning that some points of critique do not reflect the latest version of the test. For example, discourse-level productive tasks were not available at the time that these critiques were published but are now. These item types were also not operational for test-takers in the current study.

The DET technical manual describes DET "as a measure of English language proficiency for communication and use in English-medium settings" while claiming to assess both general and academic English, because both are crucial for academic and professional success (Cardwell et al., 2022, p. 3). The manual also designates "post-secondary admission decisions" as one of several test uses in the description of test purpose (p. 3). Wagner's (2020) test review underscored the pressing need for external validation evidence to support test score use and interpretation given the widespread uptake of DET as a gatekeeping measure at English-medium universities. Put differently, perhaps the most prominent source of critique for the test has been the dearth of robust external validation research needed to support test score interpretation, which can, in turn, be used to inform university admissions decisions. To our knowledge, the only DET predictive validity study that has, as yet, been published examined correlations between DET Overall scores and English for Academic Purposes teachers' ratings of student essays and spoken comprehensibility (Ishikawa et al., 2016). However, this study did not provide measures of students' actual academic success through grades. In addition, because the study involved the retired version of DET and substantial changes to the test have been made since then, an updated study is necessary.

The DET has an established internal research programme and has introduced external validation funding streams to build the evidence base to support test score use for its intended purposes. In 2022, Burstein and her DET colleagues published a "theoretical assessment ecosystem" comprising a set of frameworks that underlie test design and

development to guide future validation efforts. Using an argument-based approach (e.g., Chapelle, 2021), they presented a digital chain of inferences that considers how DET's artificial intelligence capabilities contribute to test scores, following the validation tradition notably for TOEFL iBT. In terms of concurrent validity, in-house research reports moderate correlations between DET Overall scores and IELTS ($r = .65$; $n = 1643$) and strong correlations with test centre–administered TOEFL iBT scores ($r = .82$; $n = 183$; Cardwell et al., 2022).

Given that DET is a relatively new high-stakes test and that widespread uptake has been recent, large-scale data sets of test performance have only recently become available. The current study contributes to empirical validation of DET through secondary analysis of international students' DET Overall scores and academic grades shortly after DET was accepted for admissions at a major UK university, taking into account degree level, subject area, and nationality. It also examines how well students perform academically depending on which ELP test they used for admissions. In the next section, we overview main findings from previous predictive validity research on other ELP tests to foreground the current study.

Predictive validity: Some considerations from previous research

Proficiency in the language of instruction is critical for success in every academic subject and at all levels of education (e.g., Prevoo et al., 2016; Trenkic & Warmington, 2019; Whiteside et al., 2017). Limited proficiency constrains not only how knowledge can be demonstrated in assessment, but crucially what can be learnt. Despite this, ELP test scores used for university entrance purposes are often only weakly correlated with academic outcomes (Ihlenfeldt & Rios, 2022), and sometimes no relationship is detected at all (Arcuino, 2013; Krausz et al., 2005). Some studies with null results might also not have been disseminated due to publication bias (Franco et al., 2014).

One reason for these inconsistent findings is the plethora of factors that can play a part in academic progress and success, with variables possibly interacting with or cancelling each other out (e.g., area of study, educational background, test-taker attitudes, study skills, support networks; Oliver et al., 2012; see Pearson, 2020, for a discussion of methodological considerations). That is, heterogeneous participant samples can obscure the predictive validity of ELP tests, thereby attenuating the relationship with ELP test scores. Bridgeman et al. (2016) illustrated the importance of conducting subgroup analyses to investigate whether a correlation involving all participants conceals patterns that could differ across meaningful subgroups. Variables such as degree level, nationality, and subject area, among others, have been shown to be important to disaggregate in data analysis (Harsch et al., 2017). In a pandemic, with greater levels of unpredictability than in normal times, there are yet other variables that could differentially affect test-takers and how they perform for reasons unrelated to ELP or academic ability (e.g., illness, travel restrictions, resource access, programme mitigations), compounding the uncontrolled factors that could come into play in shaping students' academic outcomes.

There is also the issue of range restriction, a statistical artefact in predictive validity studies that occurs when the performance range of the sample is constrained compared

to that of the target population, resulting in a truncated distribution (Schneider, 2014). As in other predictive validity studies, we could only examine the academic performance of test-takers who had enrolled at the university, not test-takers who chose not to apply or were not admitted due to test scores below the admissions thresholds. Therefore, only a restricted sample of prospective applicants who took the test (i.e., high scorers) were admitted and received academic grades and, hence, are reported on, reflecting a narrower range of performance than the population of interest (i.e., all prospective applicants who took DET). To summarise, although well-developed proficiency in the language of instruction is crucial for success in every academic subject, proficiency scores are often only weakly correlated with academic outcomes, if any relationship is detected at all. This is because (a) many factors are predictive of academic outcomes; (b) in underpowered heterogenous samples, these variables can statistically cancel each other out; and (c) the sample likely has a restricted proficiency range. That is why even small correlational values between ELP test scores and academic outcomes can be considered meaningful (Bridgeman et al., 2016).

Finally, test preparation, also known as coaching, a well-researched washback artefact that is common in societies where learning is largely driven by exam-oriented pedagogy (e.g., Clark & Yu, 2021; Trenkic & Hu, 2021), can undermine the validity of the test. Research has demonstrated that the higher the stakes, the more likely test-takers are to engage with intensive test preparation practices in a bid to boost test scores in a short time period (Cheng et al., 2015). Such practices may be captured in test performance and scores, potentially jeopardising construct and criterion-related validity. Because DET was not predominantly used for university admission purposes prior to the pandemic, it is unlikely that intensive test preparation targeted DET at the time of the current study because the test was still relatively unknown (Cushing & Ren, 2022). We thus hypothesise that DET scores obtained at this timepoint are likely unadulterated by test preparation and, therefore, could be more valid measures of students' ELP and potentially be more predictive of students' subsequent academic attainment.

The current study

This rapid response research study is the first, to our knowledge, to investigate the predictive value of DET in relation to university students' academic attainment for the version of the test in use during the first few months of the pandemic. It is also the first comparative study of DET with well-established ELP tests. Our first research question investigates the relationship between DET Overall scores and full-time university students' first-year academic grades by degree level, academic subject classification, and nationality. The absence of a relationship between test scores and academic outcomes could suggest that the test is not a valid measure of ELP. However, such a result could also be an artefact caused by properties of the test-taker sample. One way to rule out the first option is to examine how other, more established tests behave in the same context. This led to our second research question, which examined the relationship between Overall scores on three ELP tests (DET compared to IELTS and TOEFL iBT) and students' first-year grades by degree level.

DET in-house research has recently established points of correspondence with IELTS and TOEFL iBT and benchmarked DET performance against Common European Framework of Reference for Languages (CEFR) levels, although they note that correspondence values may change to take account of findings from future alignment studies (Duolingo English Test, n.d.). Nonetheless, with correspondence values published, it is now possible to investigate the practical question of whether students who met the ELP entry requirements for DET achieve a similar level of academic success as those who had their ELP certified using more established ELP measures. This motivated our third research question, which investigated whether students who took DET to meet ELP admissions requirements enjoy the same level of academic success as those admitted with IELTS or TOEFL iBT scores.

Methods

Research context and data

The setting for this study is University College London (UCL), a large, research-intensive (Russell Group) university in central London with 40% international students. The study, which uses university admissions and attainment data for the 2020–2021 student intake, captures the unique circumstances that arose during the COVID-19 pandemic months after the university and many other English-medium settings in the United Kingdom and elsewhere adopted DET. The university Admissions Requirements Panel accepted DET as proof of ELP for university entrance purposes in early March 2020. The decision was made before the start of the first UK lockdown but when virus containment efforts were already affecting China and other parts of the world. This was before IELTS and TOEFL iBT had launched home editions of their tests. Against this backdrop, nearly all teaching was conducted remotely for this cohort due to UK government restrictions, including, at some junctures, stay-at-home orders and social distancing guidelines that precluded face-to-face meeting. This led to adapted instruction and, in some cases, assessments. Due to the substantial disruption and unpredictability of the situation, the university implemented a "no-detriment policy" to mitigate students being unduly penalised. Compared to a non-pandemic year, there were higher numbers of extensions, interruptions of study, marking delays, and delayed exam boards. It is important to acknowledge these and other pandemic-related irregularities inherent in the data set.

We conducted secondary analysis on first-year academic grades of 1881 DET test-takers (1389 postgraduate, 492 undergraduate [UG]) enrolled full-time in a new programme at the university in Autumn 2020. This included 1389 postgraduate taught (PGT) students, (429 male, 959 female, 1 Other; $M_{age} = 24.67$ years, $SD = 3.59$) and 492 UG students (239 male, 252 female, 1 other; $M_{age} = 19.03$, $SD = 1.05$). The PGT students were from 243 distinct degree programmes and represented 74 different nationalities, with Chinese (1073), the largest subgroup (nearly 80%), followed by Italian (31), German, and Indian (24 each). For UG students representing 48 nationalities, Chinese students again constituted by far the largest group (305; over 60%), followed by Malaysian (21), Polish (20), and Spanish (16).

At the time of the university adopting DET, Duolingo only routinely reported overall test scores. Thus, only DET Overall scores were taken into account in admissions decisions,

with cut scores applied depending on the linguistic demands of the programme, as specified at programme level (Standard: 115; Good: 125; Advanced: 135). DET subscores were only routinely available for test-takers who had applied from 7 July 2020, onwards (LaFlair & Tousignant, 2020), limiting the data set that we could explore in relation to subscore performance. Therefore, in this paper, we solely report DET Overall scores.

To investigate DET test-takers' academic grades compared to those of students admitted using established ELP tests, we analysed the data of an additional 3087 full-time students who had commenced their studies at the same (Autumn 2020) entry point with scores for either IELTS (2650) or TOEFL iBT (430). The admissions team does not routinely enter scores from more than one ELP test for applicants into the admissions portal, precluding analysis of individuals' performances across multiple ELP tests if there were cases of this. Admissions staff also did not log whether students had taken traditional test centre versions of these tests, or remote versions introduced during the pandemic. Proportionally, 54.6% of IELTS and 52.6% TOEFL iBT test-takers were Chinese compared to 73.3% of DET test-takers. This discrepancy could be accounted for by test centres closures in China and at-home versions of the competitor tests not being available there (Isbell & Kremmel, 2020).

Data preparation and coding

We organised and coded the DET data set using the following variables pertaining to test-takers:

1. Degree level: PGT or UG.
2. Nationality: With Chinese students constituting the largest nationality and due to small sample sizes of other nationalities, we compared Chinese students, including those from Hong Kong, Macau, and Taiwan, with students from all other nationalities (hereafter Chinese or non-Chinese).
3. Subject classification for programme of study: To categorise 341 distinct programmes represented in the DET data set into broader subject areas, we adopted the European Research Council's (ERC) disciplinary typology of (a) Life Sciences (LS), (b) Physical Sciences and Engineering (PSE), and (c) Social Sciences and Humanities (SSH; ERC, 2021). These broad categories also accord with three of Durrant's (2017) corpus-informed disciplinary groupings. Detailed discipline codes under each ERC category facilitated coding decisions.
4. DET scores: DET Overall score and, where available, subscores (not reported on in this paper).
5. First-year credit-weighted academic grades at first attempt: These grades included coursework and exams for taught subjects but excluded theses or dissertations. We focused on grades of enrolled students' first year rather than termly grades because course structures differ across subject areas, with some courses spanning multiple terms, thereby conflating termly performance distinctions. In addition, at UK universities, yearly (not termly) grades are consequential for progression and degree classification decisions. We analysed initial (not final) course grades. All faculties use percent grading except for the Education faculty, which uses letter grades (A–F) for PGT students. We therefore aligned letter grades to the percent scale using the median of each band for these students (e.g., A = 85, B = 65, C = 55).

Data analysis

We used the open-source statistical tool, R, to compute statistical analysis and the tidyverse package for data wrangling. To explore the relationship between DET Overall scores and academic grades (Research Question 1), we ran Pearson correlations (two-tailed) for the full DET test-taker sample and for PGT and UG students, then subgroup analysis by subject classification and nationality. We computed the Thorndike Case 2 formula (Sackett & Yang, 2000) to correct for range restriction using unpublished auxiliary data supplied by Duolingo for tests administered between 31 July 2020 and 13 July 2021, with indices available by degree level only (not by subject area or nationality). This statistical adjustment accounts for enrolled students, who are at the higher end of the DET performance distribution, not being representative of all prospective applicants who take DET.

We then computed correlations between distinct groups of PGT and UG students admitted to the university using IELTS or TOEFL iBT scores at the same entry point as the DET test-takers. We were unable to correct for range restriction due to the unavailability of comparable auxiliary data across tests. To compare academic grades by students admitted using DET versus IELTS and TOEFL iBT (Research Question 3), we performed a one-way analysis of variance (ANOVA) to examine whether there were differences in PGT and UG students' academic grades depending on whether they were admitted with DET or one of the other tests. Finally, we performed ANOVAs or *t*-tests depending on sample size to examine whether test-takers, who were considered by their test providers to have achieved a particular CEFR level (B2–C2), performed differently academically as a function of ELP test.

Results

The first research question examined correlations between DET Overall scores and first-year academic attainment. Table 1 shows descriptive statistics and both Pearson correlations and range restriction–adjusted correlations between DET Overall scores and weighted average first-year academic grades. DET scores positively correlated with grades for the full sample DET test-takers ($r = .046$). Breaking down the result by degree level shows that the strength of the relationship is driven by the PGT cohort. Some caution should be applied in interpreting the UG result, however, with negative correlations suggesting that dissimilar groups were likely aggregated. The relationships for all coefficients, whether positive or negative, were strengthened after correcting for range restriction.

Table 2 shows descriptive statistics and correlations between DET scores and academic grades for PGT and UG students disaggregated by subject classification: LS, PSE, and SSH. For PGT students, positive correlations were revealed between DET scores and academic attainment for both PSE students ($r = .255$) and SSH students ($r = .085$); however, we could not adjust for range restriction due to unavailable parameters. For the UG cohort with a smaller overall sample size, relationships between DET scores and academic grades were strongest for SSH students (.120) than PSE students (.095).

We can observe that PSE students arrived with relatively lower DET scores but finished the year with higher average grades than the other two groups. Hu and Trenkic

Table 1. Descriptive statistics, correlations, and range restriction–adjusted correlations for DET Overall scores and first-year credit-weighted academic grades for the full sample of DET test-takers then grouped by degree level.

Cohort	Measures	*n*	*M*	*SD*	*r*	Adj. *r*
All DET	DET scores[a]	1881	126.54	7.72	.046*	.110
Test-takers	Academic grades		66.13	7.96		
PGT	DET scores	1389	126.20	7.49	.089**	.195
	Academic grades		65.76	7.08		
UG	DET scores	492	127.51	8.25	−.047	−.112
	Academic grades		67.19	9.95		

Note: DET: Duolingo English Test; PGT: postgraduate taught; UG: undergraduate.
[a]Scores reported using DET's 0–160 scale.
*$p < .05$. **$p < .01$.

Table 2. Descriptive statistics and correlations between DET scores and PGT and UG students' first-year academic grades by subject classification.

Degree level	Subject (*n* programmes[a])	Measures	*n*[b]	*M*	*SD*	*r*
PGT	LS (55)	DET scores	135	127.26	7.35	.065
		Academic grades		65.56	6.71	
	PSE (52)	DET scores	198	123.61	7.82	.255**
		Academic grades		68.22	7.57	
	SSH (146)	DET scores	1056	126.55	7.35	.085**
		Academic grades		65.32	6.95	
UG	LS (21)	DET scores	51	129.71	7.58	.067
		Academic grades		68.01	13.18	
	PSE (35)	DET scores	187	121.34	6.77	.095
		Academic grades		69.80	10.67	
	SSH (46)	DET scores	254	131.61	6.41	.120
		Academic grades		65.11	8.04	

Note: DET: Duolingo English Test; PGT: postgraduate taught; UG: undergraduate; LS: Life Sciences; PSE: Physical Sciences and Engineering; SSH: Social Sciences and Humanities.
[a]*n* programmes included under each subject category as represented in the data set.
[b]*n* students.
*$p < .05$. **$p < .01$.

(2021) observed a similar trend, indexing by IELTS scores. A one-way ANOVA with academic discipline as the independent variable and DET Overall score as the dependent variable confirmed a statistically significant between-group difference in the PGT cohort, $F(2, 1386) = 14.618$, $p < .001$, $\eta^2 = .21$. Pairwise comparisons using Tukey's Honest Significant Difference (HSD) test ($\alpha = .05$) revealed that PSE students were admitted with significantly lower Overall DET scores than both LS students, $p < .001$, 95% confidence interval (CI) = [1.704, 5.591], and SSH students, $p < .001$, 95% CI = [−4.290, −1.594]. The groups also differed significantly in achieved academic grades, $F(2, 1386) = 14.265$,

Table 3. Descriptive statistics and correlations between DET scores and academic grades for Chinese and non-Chinese students grouped by degree level.

Degree level	Nationality	Measures	*n*	*M*	*SD*	*r*
PGT	Chinese	DET Overall	1089	125.08	7.03	−.044
		Academic grades		64.90	6.52	
	Non-Chinese	DET overall	300	130.28	7.72	.206**
		Academic grades		68.86	8.13	
UG	Chinese	DET Overall	313	124.79	7.36	.070
		Academic grades		66.78	9.54	
	Non-Chinese	DET Overall	179	132.26	7.55	−.094
		Academic grades		67.91	10.62	

Note: DET: Duolingo English Test; PGT: postgraduate taught; UG: undergraduate.
*$p < .05$. **$p < .01$.

$p < .001$, $\eta^2 = .020$. Pairwise differences confirmed that PSE students, despite arriving with lower DET scores, finished the year with higher academic grades on average compared to both LS students, $p < .001$, 95% CI = [−4.491, −0.816], and SSH peers, $p < .001$, 95% CI = [1.621, 4.171]. Overall academic attainment data for all university students for the 2020–2021 academic year showed that students in PSE subjects obtain higher grades on average than do SSH students. For example, 57.7% of PGT students from the two largest PSE faculties were awarded distinctions for their degrees compared to just 18.5% of PGT students from the two largest SSH faculties.

In the UG cohort, there were also significant main effects for DET scores, $F(2, 489) = 130.819$, $p < .001$, $\eta^2 = .349$, and academic grades, $F(2, 489) = 12.745$, $p < .001$, $\eta^2 = .050$. Pairwise comparisons confirmed that UG students studying PSE were admitted with lower overall DET scores compared to LS students, $p < .001$, 95% CI = [5.891, 10.847], and SSH students, $p < .001$, 95% CI = [−11.789, −8.766], but finished the year with higher grades than SSH students, $p < .001$, 95% CI = [−6.895, −2.450]. Such discrepancies suppress the strength of the correlation between the observed ELP test score and academic attainment in the aggregated samples.

Table 3 shows descriptive statistics and correlations between DET scores and academic grades for PGT and UG students disaggregated by nationality. The strongest positive association was detected between DET scores and academic grades for non-Chinese PGT students, $r = .206$. In all other groups, the observed association was weak (positive for UG Chinese, but negative for PGT Chinese and UG non-Chinese students).

Independent samples *t* tests showed that at both degree levels, Chinese students achieved significantly lower DET scores than non-Chinese students, $t_{PGT}(444.989) = 10.541$, $p < .001$; $t_{UG}(490) = 10.725$, $p < .001$. Chinese students also received lower academic grades than their non-Chinese peers; however, this difference was only significant for PGT students, $t = 7.779$, $p < .001$. No positive relationship was detected between DET scores and grades for Chinese UG students, $r = -.094$.

To probe whether different patterns might emerge for Chinese and non-Chinese students depending on their academic field, we further disaggregated PGT student data by

Table 4. Correlations between DET scores and academic grades for Chinese and non-Chinese PGT students by subject classification.

Subject	PGT Chinese		PGT Non-Chinese	
	n	*r*	*n*	*r*
LS	86	−.023	49	.055
PSE	145	−.096	53	.365**
SSH	858	−.022	198	.248***

Note: DET: Duolingo English Test; PGT: postgraduate taught; LS: Life Sciences; PSE: Physical Sciences and Engineering; SSH: Social Sciences and Humanities.
*$p < .05$. ** $p < .01$. **$p < .001$.

Table 5. Descriptive statistics and correlations between IELTS and TOEFL iBT scores and students' academic grades by degree level.

ELP tests	Degree level	*n*	*M*	*SD*	*r*
IELTS	PGT	1617			
	IELTS scores		7.32	0.54	
	Academic grades		67.92	6.87	.175**
	UG	1033			
	IELTS scores		7.54	0.58	
	Academic grades		70.01	9.25	−.078**
TOEFL iBT	PGT	248			
	TOEFL scores		107.29	5.39	
	Academic grades		69.98	6.98	.149*
	UG	182			
	TOEFL scores		107.54	8.72	
	Academic grades		72.15	8.54	.047

Note: ELP: English language proficiency; PGT: postgraduate taught; UG: undergraduate.
*$p < .05$. **$p < .01$.

nationality and subject classification. Table 4 shows that the most robust associations for subject classification were detected for PSE and SSH PGT Non-Chinese students ($r_{PSE} = .365$, $r_{SSH} = .248$), despite smaller sample sizes than for their Chinese counterparts. The main insight is the confirmation that the correlation strength between DET and academic grades goes up in more homogenous samples, as can be seen in the increased *r* value for both PSE and SSH when Chinese students are excluded from the sample.

The second research question examined the association between DET scores and academic grades for enrolled DET test-takers compared to those admitted using another high-stakes ELP test. Table 5 shows descriptive statistics and correlations for PGT and UG students' overall scores on IELTS and TOEFL iBT and first-year credit-weighted academic grades.

At PGT level, following the trend for DET (Table 1), IELTS and TOEFL iBT scores positively correlated with academic grades. Thus, PGT students arriving with higher

scores on all three ELP tests performed better academically than peers who received lower scores on each of these three tests. The strength of correlation was higher for both IELTS ($r = .175$) and TOEFL iBT ($r = .149$) than for DET ($r = .089$). However, sample size differences and different test-taker background characteristics across cohorts (e.g., higher proportion of Chinese students for DET compared to the other two tests) make it important to only interpret overall trends. At UG level, a weak positive correlation for TOEFL iBT ($r = .047$) and weak negative correlation for IELTS ($r = –.078$) and DET ($r = –.047$) suggest high heterogeneity in the UG sample, which can mask meaningful patterns that exist within subgroups. Indeed, we observed disciplinary differences in the UG cohort in Table 2 for DET test-takers but did not have discipline-related information for IELTS and TOEFL iBT test-takers. Therefore, we could not examine group differences by subject classification for these tests.

To explore whether students who demonstrated readiness to study in English via DET experienced similar levels of academic success as those admitted using IELTS or TOEFL iBT (Research Question 3), we first performed a one-way ANOVA, comparing academic grades for the test-taker groups at each degree level. For PGT students, there were significant between-group differences for academic grades $F(2, 3251) = 58.080$, $p < .001$, $\eta^2 = .034$. Tukey's HSD post hoc test confirmed significant differences between PGT students accepted with IELTS versus DET, $p < .001$, 95% CI = [1.564, 2.759], and between TOEFL iBT and DET students, $p < .001$, 95% CI = [3.100, 5.353], with DET students performing less well academically in both cases. Similar patterns were observed in the UG subsample, with a one-way ANOVA confirming a significant between-group difference $F(2, 1704) = 23.561$, $p < .001$, $\eta^2 = .027$. Tukey's HSD pairwise comparisons confirmed that UG students who took IELTS received higher academic grades than those who took DET, $p < .001$, 95% CI = [1.608, 4.020]. TOEFL iBT students also performed more strongly academically than DET students, $p < .001$, 95% [CI = 3.045, 6.865].

As the average level of ELP with which different test-taker groups were admitted might have differed and so contributed to the observed differences in grades, we mapped each student's English proficiency score onto a CEFR level based on the test providers' own published correspondences between the scores on their test and CEFR equivalents (Duolingo English Test, n.d.; ETS, 2022; IELTS, 2022). Notably, different scales are used for each test and test providers may have used different equating and standard-setting methods or other procedures for establishing equivalences.

Table 6 summarises the data by degree level grouped by the CEFR levels that are generally relevant for university admissions (B2, C1, and C2) for PGT and UG students, respectively. A one-way ANOVA for PGT students at C1 level (the largest subgroup) revealed a significant main effect for ELP test, $F(2, 2613) = 45.264$, $p < .001$, $\eta^2 = .033$. Tukey's HSD pairwise comparisons confirmed that PGT students accepted with IELTS, $p < .001$, 95% [CI = 1.625, 2.982], and TOEFL iBT, $p < .001$, 95% [CI = 2.637, 5.151], obtained significantly higher grades than DET test-takers. A significant result for UG students was also found at C1 level, $F(2, 1362) = 20.869$, $p < .001$, $\eta^2 = .030$, with pairwise comparisons revealing that students accepted with DET received lower grades on average than students accepted with both IELTS, $p < .001$, 95% [CI = 1.775, 4.438], and TOEFL iBT, $p < .001$, 95% CI = 2.746, 6.874]. At B2 level, Welch's *t*-tests (conducted due to sample size restrictions and imbalance)

Table 6. First-year academic grades for PGT and UG students admitted using DET, IELTS, and TOEFL iBT by CEFR level.

	CEFR level	DET			IELTS			TOEFL iBT[a]		
		n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
PGT	B2	236	65.09	6.08	272	66.49	6.12	2	68.20	0.42
	C1	1136	65.76	7.22	1273	68.06	6.96	207	69.65	7.17
	C2	17	74.74	4.83	72	70.74	6.72	39	71.82	5.81
UG	B2	81	68.37	10.09	119	70.95	10.24	0	-	-
	C1	396	66.85	9.97	815	69.95	9.00	154	71.66	8.69
	C2	15	69.92	8.24	99	69.30	9.92	27	74.91	7.37

Note: PGT: postgraduate taught; UG: undergraduate; DET: Duolingo English Test.
[a]One student, whose TOEFL iBT score was recorded as 11 (a likely data entry error), was removed from the analysis.

Table 7. First-year academic grades for PGT Chinese students admitted using DET, IELTS, and TOEFL iBT by CEFR level.

Degree level	CEFR level	DET			IELTS			TOEFL iBT		
		n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
PGT	B2	215	65.03	6.16	206	66.88	6.08	2	68.20	0.42
	C1	873	64.87	6.61	729	67.58	6.85	100	68.01	7.11
	C2	1	67.51	–	3	68.69	2.42	6	65.60	4.30

Note: PGT: postgraduate taught; DET: Duolingo English Test.

revealed that PGT students accepted with DET received significantly lower grades on average than students accepted with IELTS, $t(497.03) = -2.572$, $p < .05$. This difference was nonsignificant at UG level. For PGT students, DET test-takers considered to be at C2 level performed more strongly overall than C2 level PGT students admitted using more established tests in terms of mean scores.

Finally, to rule out the possibility that the pattern was driven by a higher proportion of Chinese students among DET test-takers, we segregated the ELP data by CEFR level for PGT Chinese test-takers only (see Table 7). Significance testing mirroring that reported above revealed the same trend—PGT Chinese DET test-takers received significantly lower scores at B2 and C1 levels on average than their compatriots who had been admitted using the other ELP tests. A one-way ANOVA at C1 level revealed a significant main effect for ELP test, $F(2, 1699) = 36.003$, $p < .001$, $\eta^2 = .041$. Tukey's HSD confirmed that PGT Chinese students at C1 who were accepted with IELTS, $p < .001$, 95% [CI = 1.922, 3.509], and TOEFL iBT, $p < .001$, 95% [CI = 1.477, 4.816], obtained significantly higher grades than did DET test-takers. Welch's *t*-tests revealed that B2 level PGT students accepted with IELTS received significantly higher grades than did students admitted using DET scores, $t(418.63) = -3.115$, $p < .05$.

Discussion

This study is the first, to our knowledge, to examine the predictive validity of the revised (non-retired) version of DET in relation to academic attainment and to offer comparisons with established ELP tests. As a new test, it is only recently that large enough data sets of DET test-takers have become available with which to conduct such analyses. Validating the test through a programme of rigorous research is pressing, particularly due to its high-stakes and widespread use. Situated in the early stages of the pandemic shortly after DET had been adopted at a major UK university, this study contributes to criterion-oriented validity evidence. The headline predictive validity finding is that university students who arrive with higher DET Overall scores go on to achieve higher academic grades in their first year of study than do lower DET performers. This finding is in line with small but significant positive results for the established ELP tests in most previous predictive validity studies (Ihlenfeldt & Rios, 2022), notwithstanding publication bias.

After conducting subgroup analysis by degree level, we found that the correlation between DET scores and academic grades was positive for PGT but not UG students. We observed similar patterns for IELTS and TOEFL iBT as we did for DET, with a stronger result for PGT than UG for both competitor tests. It would be wrong to conclude from these results either that ELP is inconsequential in UG studies or that all three tests are poor measures of ELP. There are at least three plausible reasons why all three tests failed to detect a relationship between ELP and grades. The first is that in contrast to nearly all PGT programmes, first-year UG grades do not count for final degree classifications, potentially resulting in less effort on the part of UG students in summative assessments in their first year, which could weaken the predictive value of ELP test scores. Second, PGT courses tend to involve essays, which may be more linguistically demanding and involve more language production than exams, which tend to pervade UG assessment (Woodrow, 2006), potentially attenuating correlations between ELP scores and grades for UG students. Finally, as in all predictive validity studies, factors other than ELP come into play (Oliver et al., 2012). In highly heterogeneous samples, such factors can mask the predictive power of ELP scores, and this is likely to be the case here too.

To "peel . . . the onion" and examine whether different patterns exist across subgroups for the variables identified in the research questions (Bridgeman et al., 2016, p. 310), we disaggregated the PGT and UG results first by subject classification. For the PGT cohort, the overall association between DET scores and academic grades was driven by PSE and SSH but not LS students. For UG students, a very weak association by subject classification was found solely for SSH students, with no relationship detected for PSE and LS students.

Within the PGT cohort, the stronger positive correlation for PSE compared to SSH students could appear counterintuitive. That PSE students should arrive with lower mean DET scores is unsurprising, since cut scores on ELP tests for less linguistically demanding disciplines are typically lower. Why they go on to achieve higher academic scores is less clear. Previous predictive validity studies have shown that ELP tends to be more consequential for SSH subjects, which tend to be more English-intensive (e.g., more readings, extended writing), compared to PSE disciplines that focus on numeracy skills (e.g., Harsch et al., 2017). Although SSH students attain higher ELP test scores on

average, this is offset by higher linguistic demands. Therefore, SSH students' academic grades may be lower than ELP tests predict. In PSE disciplines, by comparison, math- or coding-oriented academic demands may mean less emphasis on ELP skills. It is also likely that the nature of assessment and/or grading practices played a role in different grade ranges for PSE compared to SSH students. University data on average degree classifications confirmed that students from PSE disciplines receive higher grades than do SSH students, which suggests that the nature of the assessment and marking practices are likely driving the effect. With the large number of programmes under the umbrella SSH classification, it is also possible that there is more diversity of practice and variability in linguistic demands within this cohort than in the other subject areas. Taken together, the results highlight the need to either control for or disaggregate analyses by subject classification when investigating the role of ELP in academic outcomes.

Chinese students, who constituted the largest subgroup at both degree levels in our sample, arrived with significantly lower overall DET sores compared to non-Chinese students. This finding is consistent with Harsch et al.'s (2017) and Bridgeman et al.'s (2016) predictive validity studies, which found that Chinese students received lower TOEFL iBT Overall scores and subscores than non-Chinese students. In our study, Chinese students also achieved lower academic grades than non-Chinese students, and this difference was significant in the PGT cohort. However, DET was predictive for the (smaller) non-Chinese student group but non-predictive for the (larger) Chinese student group. Because the Chinese group would seem to be more homogeneous than the group consisting of all other nationalities—notwithstanding the considerable linguistic and other diversity within China (Gong et al., 2011)—we might have expected a stronger correlation for Chinese students compared to their non-Chinese counterparts, which is consistent with what both Bridgeman et al. (2016) and Harsch et al. (2017) found for TOEFL iBT and IELTS, respectively. The fact that Chinese students are more numerous in our study compared to all other students means that they were likely more heterogeneous at the level of represented programmes and disciplines in our study compared to non-Chinese students.

When we conducted subgroup analysis by discipline, we again found weak negative correlations for PGT Chinese students for all three subject classifications compared to larger positive correlations for two of the disciplinary groups for non-Chinese students (PSE and SSH). One possible explanation is that Chinese test-takers who took DET had no alternative tests with which to demonstrate their ELP level, with IELTS Indicator and TOEFL iBT Home Edition not in operation in China after they were introduced (Isbell & Kremmel, 2020). By contrast, test-takers in almost all other countries could choose among these tests as alternatives to DET once the university had accepted them as proof of ELP. Thus, non-Chinese students who opted to take DET are more likely to have preferred and deliberately chosen it over the competitor tests. Consequently, they may have been more familiar with the test format than their Chinese counterparts, and some may have performed more poorly than they would have had they been more accustomed to the test. Whatever the reason, future research is necessary to confirm or disconfirm the patterns revealed in this study.

Students accepted with DET experienced lower academic success than those accepted with IELTS and TOEFL iBT. It is possible that had DET's latest concordance data

(Cardwell et al., 2022) been available at the time of the study, some lower-scoring students who had gained entry to the university through DET might not have been admitted. As described above, correlational patterns between DET scores and academic grades were mirrored in the results for IELTS and TOEFL iBT. Neither IELTS nor TOEFL iBT scores related to first-year academic outcomes in the UG cohort, confirming an underlying complexity and heterogeneity of this sample. However, all three tests had positive relationships with grades for PGT students, with the correlation for DET being somewhat lower than for IELTS and TOEFL iBT. This supports the predictive validity of DET (i.e., higher DET scores tends to mean higher grades) because of roughly similar patterns between DET Overall scores and scores for IELTS and TOEFL iBT. In particular, it confirms that the negative correlation between UG students' DET scores and grades most likely stems from the properties of the sample rather than being a property of the test itself.

This pattern remained when the students were grouped by CEFR level. At C1 level, both PGT and UG students who had been admitted with DET achieved lower grades than their peers who had test scores certified by DET's more established industry competitors. This was the most common CEFR level in our data set, thereby yielding the most statistically robust result. At B2 level for PGT students but not UG students, IELTS test-takers received significantly higher scores than DET test-takers. Conversely, at C2 level, DET test-takers received higher mean academic grades than the other test-takers. In sum, students arriving with DET (with the possible exception of a the few students at CEFR's top level) experienced lower academic success than students arriving with IELTS and TOEFL iBT scores. Finally, by detecting the same pattern on the subgroup of Chinese students, we ruled out the possibility that DET test-takers' lower academic grades were an artefact of the higher ratio of Chinese students in the DET sample compared to the IELTS and TOEFL iBT samples: Chinese students arriving with DET still went to achieve lower academic grades than Chinese students admitted with IELTS or TOEFL iBT.

This analysis relies on test providers' own published equivalences of test scores and CEFR levels. However, this does not imply uniformity in method, nor consensus on which CEFR level equates to each test score. Correspondence points might not be adequately set nor directly comparable. In terms of DET, Cushing and Ren (2022) highlighted that although there are now published concordances between DET and other ELP tests and between DET scores and CEFR levels (DET, n.d.), it is unclear how these concordances were established. This calls for a closer look at the correspondence points between tests and suggests that higher DET cut scores, potentially of Overall scores in conjunction with subscores, might be necessary to ensure a comparable academic performance with students certified through the traditional ELP tests. Most students in the current study were admitted through Overall DET scores alone, as subscores were not routinely reported to universities until the first week of July 2020 (LaFlair & Tousignant, 2020).

It is possible that DET test-takers' comparatively lower academic performance has to do with the test itself. Alternatively, their performance may have nothing to do with the test at all (e.g., could be attributed to sample characteristics and external factors). In relation to the first point, the fully automated DET is measuring a vastly different ELP

construct than human-scored traditional ELP tests (Khabbazbashi et al., 2021). Cushing and Ren's (2022) analysis of IELTS Academic and DET based on publicly available information highlighted differences between the two tests in terms of cognitive and context validity (e.g., cognitive skills elicited, academic orientation). Regarding the second point, it is possible that academically weaker students applied in the latter part of the university's admission cycle, when DET was being accepted, and that this explains the lower academic grades. One of the key benefits of DET is its greater accessibility and affordability compared to market competitors (e.g., Wagner, 2020). It could be that students with lower socioeconomic status, who also tend to have lower academic achievement (Muttaqin et al., 2022), opted to take DET compared to more expensive alternative tests. In sum, test-taker characteristics and other factors that are extraneous to the test could have shaped the findings.

Clearly, there is a need for further research to establish whether the patterns observed in this study persist in other contexts and, if so, why. But on a practical level, it suggests that, for whatever reason, students arriving with ELP scores via DET may need a higher cut-off point to perform at a comparable level to students arriving with other ELP tests or may need additional support to fulfil their academic potential. It should be noted that few DET test-takers benefitted from pre-sessional support in our study. Pre-sessional DET student performance compared to that of test-takers who were admitted outright could be an additional variable to examine in future research (e.g., Trenkic & Warmington, 2019). Most students in our DET cohort were only admitted on the basis of DET Overall scores, due to DET reporting practices that changed in the latter part of the admissions cycle for the 2020–2021 intake, when DET introduced routine subscore reporting. Future research could examine the effect of different DET score profiles on academic grades (Ginther & Yan, 2018), how this can be interpreted in light of DET's movement towards reporting composite skills through subscores rather than the traditional speaking/listening/reading/writing demarcation (Cardwell et al., 2022), and the appropriateness of cut scores for different disciplines.

Limitations

This study contributes to pressing criterion-related validity research on DET in the first full academic year after the onset of the pandemic at the United Kingdom's largest on-campus university. The results reflect this snapshot of time. In line with government measures to stop the virus spread, most programmes needed to move to fully online provision. Students had little or no access to on-campus resources and less exposure to university life than in a non-pandemic academic year, with some students conducting their studies from abroad in fully non-English medium environments. Changes in teaching mode and, in some cases, adjustments to assessments may have affected students' academic performance and grades. For example, to account for pandemic-related disruption, PGT students were allowed up to three resubmissions of failed assignments with no capped grades compared to only one capped resubmission pre-COVID. This policy, which inadvertently disadvantaged students who had submitted their assignments on time and passed but did not receive the opportunity to resubmit later and improve their grade, may have resulted in gaming the system. More students

were also granted deferrals and interruptions of study than in a normal year. Some markers may have scored differently and potentially more generously because of perceived challenges to do with remote delivery (Karadag, 2021). The academic attainment data in this study were, thus, coloured by these factors, limiting the generalisability of the findings.

Because widespread uptake of DET for admissions purposes was recent at the time of the study, DET test-takers may have been unfamiliar with DET item types (Cushing & Ren, 2022). Thus, the resulting scores may have underestimated their ELP. However, the fact that the DET test preparation industry was also not yet well established at the time also means that intensive coaching was limited, hence reducing tailored test-taking strategies as a potential source of construct-irrelevant variance. In this sense, the DET scores in this data set may be a purer measures of test-takers' ELP than in the future, when test-takers are more rehearsed, test-wiseness strategies are explicitly taught, and test-takers might have access to more resources to help them game the system and optimise their scores. This may especially be the case in exam-oriented societies that have established test preparation industries and cram school traditions (e.g., Clark & Yu, 2021; Tsang & Isaacs, 2022). However, Cushing and Ren's (2022) analysis of postings on Chinese online discussion fora (2020–2021) suggested that at least one online poster felt that rigorous drilling, question pattern analysis strategies, and the use of contextual cues, which are apparently coached for IELTS and TOEFL iBT, may not be useful for DET, which requires knowledge of English structures under response time pressures. Future research should investigate how resilient DET is to coaching for the different item types.

There are several other limitations, mostly due to sample characteristics and data availability. First, we were only able to correct for range restriction for PGT and UG student DET test-takers and not for analyses involving subject classification, nationality, or competitor ELP tests due to the unavailability of auxiliary data. Second, students who met language entry requirements through DET, IELTS, and TOEFL iBT represented independent samples in our study. Therefore, we could not statistically compare their predictive utility, nor could we explore whether lower academic success for DET test-takers was due to the nature of the test, score calibration, or external factors, such as test-takers' socioeconomic status or academic ability. Furthermore, the results were likely shaped by students with different background characteristics applying at different time points in the admissions cycle. Sample size differences across the three ELP tests are also notable, making it important to exercise caution when interpreting the test comparison results.

Unbalanced sample size was also a limiting factor. Due to the high proportion of DET test-takers from China and few students from European Union (EU) countries in the 2019–2020 academic year, potentially exacerbated by Brexit, we could only make nationality comparisons for Chinese students and students from all other nationalities. We had also wanted to examine programme-level data based on whether they required Advanced, Good, or Standard ELP thresholds for admissions as an indicator of linguistic demands. However, most programmes required a Good level, limiting our ability to use it as a grouping variable. Now that Duolingo routinely reports subscores, future research could examine subscore profiles. These data were only available for a small group of test-takers who had taken the test late in the application cycle in our sample.

Yet another limitation is using first-year course grades as the sole criterion measure. There is no placement test tradition at UK universities (e.g., Purpura et al., 2021),

precluding use of in-house ELP test scores as an additional criterion measure. Furthermore, we only tracked students' academic attainment over the course of their first year of study, with most PGT students in the data set undertaking 1-year master's programmes. In the case of multiyear degrees, it would be useful to investigate whether DET is more predictive of academic grades in subsequent years.

Despite these limitations, this study provides much-needed preliminary predictive validity evidence for DET in relation to academic attainment. However, criterion-related validity is not sufficient for construct validity (American Educational Research Association et al., 2014). Given its ongoing high-stakes use at many English-medium universities, different but complementary sources of evidence are necessary to build a robust evidence base to support the intended uses and interpretations of test scores (Chapelle, 2021). Duolingo has shown its commitment to this both through work conducted in house (e.g., Burstein et al., 2022), in some cases with newly hired personnel from the assessment and psychometric communities, and through funding validation projects such as this one.

Acknowledgements

We are grateful to Jill Burstein, Antony Kunnan, Geoff LaFlair, J. R. Lockwood, and Alina von Davier from the Duolingo Assessment Research team for their input on previous versions of this manuscript. We also thank Paula Winke and the *Language Testing* reviewers for their valuable comments on previous versions of our article.

Declaration of conflicting interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: During the period of undertaking and disseminating this Duolingo-funded project, the first author, Talia Isaacs, conducted assessment-related research, advisory, or consultancy work for the following organisations: British Council, Cambridge Assessment English, Educational Testing Service (ETS), National Centre for Excellence for Language Pedagogy (NCELP), and Organisation for Economic Co-operation and Development (OECD). Talia became the Co-Editor of Language Testing January 1, 2023. This co-authored paper was submitted and accepted before Talia assumed her editorial role for the journal. Co-Editor Paula Winke managed the peer review and editorial process for this paper.

Ethical Approval

The study received ethics approval from UCL IOE Research Ethics Committee, REC 1475. The project was jointly conceived by the first three authors. The fourth author provided access to admissions and academic attainment data, the second author conducted data analyses, and the first and second authors drafted the manuscript with reframing and editing from the third author. All authors approved this manuscript.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This article reports on a Duolingo-funded commissioned study.

ORCID iD


Talia Isaacs https://orcid.org/0000-0003-4302-3379

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Arcuino, C. L. T. (2013). *The relationship between the Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS) scores and academic success of international Master's students* [Unpublished doctoral dissertation]. Colorado State University. https://www.proquest.com/dissertations-theses/relationship-between-testenglish-as-foreign/docview/1413309058/se-2

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*, Article 903077. https://doi.org/10.3389/frai.2022.903077

Bolton, P., & Hubble, S. (2021). *Coronavirus: Financial impact on higher education*. House of Commons Library. https://commonslibrary.parliament.uk/

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

Burstein, J., LaFlair, G., Kunnan, A. J., & von Davier, A. (2022). *A theoretical assessment ecosystem for a digital-first assessment: The Duolingo English Test* (Duolingo Research Report DRR-21-04). Duolingo English Test. https://duolingo-papers.s3.amazonaws.com/other/det-assessment-ecosystem.pdf

Cardwell, R., LaFlair, G. T., Naismith, B., & Settles, B. (2022). *Duolingo English Test: Technical manual* [Duolingo research report, August 8, 2022]. Duolingo English Test. https://duolingo-testcenter.s3.amazonaws.com/media/resources/technical_manual.pdf

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. SAGE.

Cheng, L., Sun, Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching*, *48*(4), 436–470. https://doi.org/10.1017/S0261444815000233

Clark, T., Spiby, R., & Tasviri, R. (2021). Crisis, collaboration, recovery: IELTS and COVID-19. *Language Assessment Quarterly*, *18*(1), 17–25. https://doi.org/10.1080/15434303.2020.1866575

Clark, T., & Yu, G. (2021). Beyond the IELTS test: Chinese and Japanese postgraduate UK experiences. *International Journal of Bilingual Education and Bilingualism*, *24*(10), 1512–1530. https://doi.org/10.1080/13670050.2020.1829538

Cushing, S. T., & Ren, H. (2022). *Comparison of IELTS academic and Duolingo English Test* (IELTS Partnership Research Papers, 2021(1)). IELTS Partners: British Council, Cambridge Assessment English, & IDP: IELTS Australia. https://www.ielts.org/for-researchers/research-reports

Duolingo English Test. (n.d.). *Understanding your score report*. https://englishtest.duolingo.com/scores

Durrant, P. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics*, *38*(2), 165–193. https://doi.org/10.1093/applin/amv011

ETS. (2022). *Comparing TOEFL iBT scores*. https://www.ets.org/toefl/score-users/scores-admissions/compare

European Research Council. (2021). *Panel structure for ERC calls 2021 and 2022 (revised)*. https://erc.europa.eu/sites/default/files/document/file/ERC_Panel_structure_2021_2022.pdf

Franco, A., Malhotra, N., & Simonovits, G. (2014). Social science. Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening- speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing*, *29*(3), 345–369. https://doi.org/10.1177/0265532211424479

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

Gong, Y., Chow, I. H.-s., & Ahlstrom, D. (2011). Cultural diversity in China: Dialect, job embeddedness, and turnover. *Asia Pacific Journal of Management*, *28*(2), 221–238. https://doi.org/10.1007/s10490-010-9232-6

Harsch, C., Ushioda, E., & Ladroue, C. (2017). *Investigating the predictive validity of TOEFL iBT® scores and their use in informing policy in a UK university setting* (TOEFL iBT Research Report No. 30, ETS Research Report No. RR-17-41). ETS. https://files.eric.ed.gov/fulltext/EJ1168729.pdf

Hu, R., & Trenkic, D. (2021). The effects of coaching and repeated test-taking on Chinese candidates' IELTS scores, their English proficiency, and subsequent academic achievement. *International Journal of Bilingual Education and Bilingualism*, *24*(10), 1486–1501. https://doi.org/10.1080/13670050.2019.1691498

IELTS. (2022). *IELTS in CEFR scale*. https://www.ielts.org/about-ielts/ielts-in-cefr-scale

Ihlenfeldt, S. D., & Rios, J. A. (2022). A meta-analysis on the predictive validity of English language proficiency assessments for college admissions. *Language*, *40*(2), 276–299. https://doi.org/10.1177/02655322221112364

Isaacs, T. (2018). Fully automated speaking assessment: Changes to proficiency testing and the role of pronunciation. In O. Kang, R. I. Thomson, & J. Murphy (Eds.), *The Routledge handbook of English pronunciation* (pp. 570–584). Routledge. https://doi.org/10.4324/9781315145006-36

Isbell, D. R., & Kremmel, B. (2020). Test review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing*, *37*(4), 600–619. https://doi.org/10.1177/0265532220943483

Ishikawa, L., Hally, K., & Settles, B. (2016). *The Duolingo English Test and academic English* (Duolingo Research Report DRR-16-01). Duolingo English Test. https://s3.amazonaws.com/duolingo-papers/reports/DRR-16-01.pdf

Karadag, E. (2021). Effect of COVID-19 pandemic on grade inflation in higher education in Turkey. *PLOS ONE*, *16*(8), Article e0256688. https://doi.org/10.1371/journal.pone.0256688

Khabbazbashi, N., Xu, J., & Galaczi, E. D. (2021). Opening the Black box: Exploring automated speaking evaluation. In B. Lanteigne, C. Coombe, & J. D. Brown (Eds.), *Challenges in language testing around the world* (pp. 333–343). Springer. https://doi.org/10.1007/978-981-33-4232-3_25

Krausz, J., Schiff, A., Schiff, J., & Hise, J. V. (2005). The impact of TOEFL scores on placement and performance of international students in the initial graduate accounting class. *Accounting Education*, *14*(1), 103–111. https://doi.org/10.1080/0963928042000256671

LaFlair, G. T., Langenfeld, T., Baig, B., Horie, A. K., Attali, Y., & von Davier, A. A. (2022). Digital-first assessments: A security framework. *Journal of Computer Assisted Learning*, *38*(4), 1077–1086. https://doi.org/10.1111/jcal.12665

LaFlair, G. T., & Tousignant, J. (2020, June 8). Subscores: Improving how we report Duolingo English Test results. *Duolingo Blog*. https://blog.duolingo.com/subscores-improving-how-we-report-duolingo-english-test-results-2

Lim, G. S. (2018). Conceptualizing and operationalizing second language speaking assessment: Updating the construct for a new century. *Language Assessment Quarterly*, *15*(3), 215–218. https://doi.org/10.1080/15434303.2018.1482493

Muttaqin, S., Chuang, H.-H., Lin, C.-H., & Cheng, M.-M. (2022). When proficiency and education matter: The mediating role of English proficiency and moderating effect of parents' education in the SES–academic achievement relationship during EMI. *SAGE Open*, *12*(2), 21582440221103542. https://doi.org/10.1177/21582440221103542

Nakatsuhara, F., & Berry, V. (2021). Use of innovative technology in oral language assessment. *Assessment in Education: Principles, Policy & Practice*, *28*(4), 343–349. https://doi.org/10.1080/0969594X.2021.2004530

Ockey, G. J. (2021). An overview of COVID-19's impact on English language university admissions and placement tests. *Language Assessment Quarterly*, *18*(1), 1–5. https://doi.org/10.1080/15434303.2020.1866576

Oliver, R., Vanderford, S., & Grote, E. (2012). Evidence of English language proficiency and academic achievement of non-English-speaking background students. *Higher Education Research & Development*, *31*(4), 541–555. https://doi.org/10.1080/07294360.2011.653958

Papageorgiou, S., Davis, L., Norris, J. M., Garcia Gomez, P., Manna, V. F., & Monfils, L. (2021). *Design framework for the TOEFL® Essentials™ Test 2021* (Research Memorandum No. RM-21-03). ETS. https://www.ets.org/Media/Research/pdf/RM-21-03.pdf

Papageorgiou, S., & Manna, V. F. (2021). Maintaining access to a large-scale test of academic language proficiency during the pandemic: The launch of TOEFL iBT Home Edition. *Language Assessment Quarterly*, *18*(1), 36–41. https://doi.org/10.1080/15434303.2020.1864376

Pearson, W. S. (2020). The predictive validity of the Academic IELTS test: A methodological synthesis. *ITL—International Journal of Applied Linguistics*, *172*(1), 85–120. https://doi.org/10.1075/itl.19021.pea

Pearson PTE. (n.d.). *PTE Academic just got better!* https://www.pearsonpte.com/articles/pte-academic-just-got-better

Prevoo, M. J. L., Malda, M., Mesman, J., & van IJzendoorn, M. H. (2016). Within- and cross-language relations between oral language proficiency and school outcomes in bilingual children with an immigrant background: A meta-analytical study. *Review of Educational Research*, *86*(1), 237–276. https://doi.org/10.3102/0034654315584685

Purpura, J. E., Davoodifard, M., & Voss, E. (2021). Conversion to remote proctoring of the community English language program online placement exam at Teachers College, Columbia University. *Language Assessment Quarterly*, *18*(1), 42–50. https://doi.org/10.1080/15434303.2020.1867145

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*(1), 112–118. https://doi.org/10.1037/0021-9010.85.1.112

Schneider, J. W. (2014, June 9). *Using the multivariate truncated normal distribution*. https://assessingpsyche.wordpress.com/2014/06/09/using-the-multivariate-truncated-normal-distribution

Trenkic, D., & Hu, R. (2021). Teaching to the test: The effects of coaching on English-proficiency scores for university entry. *Journal of the European Second Language Association*, *5*(1), 1–15. http://doi.org/10.22599/jesla.74

Trenkic, D., & Warmington, M. (2019). Language and literacy skills of home and international university students: How different are they, and does it matter? *Bilingualism: Language and Cognition*, *22*(2), 349–365. https://doi.org/10.1017/S136672891700075X

Tsang, C. L., & Isaacs, T. (2022). Hong Kong secondary students' perspectives on selecting test difficulty level and learner washback: Effects of a graded approach to assessment. *Language Testing*, *39*(2), 212–238. https://doi.org/10.1177/02655322211050600

Wagner, E. (2020). Duolingo English Test, revised version July 2019. *Language Assessment Quarterly*, *17*(3), 300–315. https://doi.org/10.1080/15434303.2020.1771343

Wagner, E., & Kunnan, A. J. (2015). The Duolingo English Test. *Language Assessment Quarterly*, *12*(3), 320–331. http://doi.org/10.1080/15434303.2015.1061530

Whiteside, K. E., Gooch, D., & Norbury, C. F. (2017). English language proficiency and early school attainment among children learning English as an additional language. *Child Development*, *88*(3), 812–827. https://doi.org/10.1111/cdev.12615

Woodrow, L. (2006). Academic success of international postgraduate education students and the role of English proficiency. *University of Sydney Papers in TESOL*, *1*, 51–70. https://www.sydney.edu.au/content/dam/corporate/documents/faculty-of-arts-and-social-sciences/research/research-centres-institutes-groups/uos-papers-in-tesol/volume-1/article03.pdf