Meta-analysis

LANGUAGE TESTING

Language Testing
1–24

DOI: 10.1177/02655322221112364
journals.sagepub.com/home/ltj
SAGE

A meta-analysis on the predictive validity of English language proficiency assessments for college admissions

Samuel Dale Ihlenfeldt
The University of Minnesota, USA

Joseph A. Rios
The University of Minnesota, USA

Abstract

For institutions where English is the primary language of instruction, English assessments for admissions such as the Test of English as a Foreign Language (TOEFL) and International English Language Testing System (IELTS) give admissions decision-makers a sense of a student's skills in academic English. Despite this explicit purpose, these exams have also been used for the practice of predicting academic success. In this study, we meta-analytically synthesized 132 effect sizes from 32 studies containing validity evidence of academic English assessments to determine whether different assessments (a) predicted academic success (as measured by grade point average [GPA]) and (b) did so comparably. Overall, assessments had a weak positive correlation with academic achievement ($r = .231$, $p < .001$). Additionally, no significant differences were found in the predictive power of the IELTS and TOEFL exams. No moderators were significant, indicating that these findings held true across school type, school level, and publication type. Although significant, the overall correlation was low; thus, practitioners are cautioned from using standardized English-language proficiency test scores in isolation in lieu of a holistic application review during the admissions process.

Keywords

Academic success, IELTS, meta-analysis, predictive validity, TOEFL iBT

Each year, a growing number of students choose to study internationally around the world at higher education institutions where English is the primary language of instruction (Baer

Corresponding author:
Samuel Dale Ihlenfeldt, Department of Educational Psychology, University of Minnesota, 250 Education Sciences Bldg, 56 E River Rd, Minneapolis, MN 55455, USA.
Email: ihlen010@umn.edu

& Martel, 2020; Institute of International Education [IIE], 2018).[1] From the perspective of international students, studying abroad can offer a better education, exposure to another culture, and greater employment opportunities (Bodycott, 2009; Mazzarol & Soutar, 2002). From the perspective of universities, international students bring culture, innovation, and increasing revenue to the university each year (Hegarty, 2014; IIE, 2018). In addition, international students have a profound economic impact for the countries where they are studying. For instance, in the United States alone, international students brought in $39 billion in foreign dollars and were responsible for the creation of hundreds of thousands of jobs (Israel & Batalova, 2021). Due to these advantages, many schools are increasing efforts to attract international students (Sá & Sabzalieva, 2018), and a number of countries have set long-term goals to increase the number of international students attending higher education in their country by as much as 100% (IIE, 2018).

In the 2019 academic year, there were more than one million international students in the United States (5.5% of enrollments; Israel & Batalova, 2021). Since 2010, there has been a nearly 60% increase in international student enrollment in the United States, the majority from China and India (IIE, 2018). These trends were comparable in other English-speaking nations: Canada (642,480 international students), the United Kingdom (485,645), and Australia (370,269; IIE, 2018). Even smaller destinations such as Ireland and New Zealand have seen increases in recent years (IIE, 2018). With the high demand to study internationally, coupled with the value of international students to universities, admissions decision-makers need to use all available evidence to select students who will succeed academically and remain enrolled.

International students wishing to attend English-instruction schools must prove their ability to communicate successfully in the English language at an advanced level in both written and oral/aural context. Research has shown that students who are below a certain threshold in these skills will not succeed in an English learning environment (Graham, 1987). Standardized English language proficiency (ELP) assessments serve as a way for admission decision-makers to determine at what level applicants can read, write, listen, and speak in English, skills that are critical for communication at a collegiate level (Graham, 1987). Academic institutions establish a minimum score to be considered for acceptance for each assessment based on the language demands of that institution.

The two most universal ELP admissions assessments are the Test of English as a Foreign Language (TOEFL, Educational Testing Service [ETS]) and the International English Language Testing System (IELTS, Cambridge English). Worldwide, the TOEFL has been administered to more than 35 million people since its inception and is accepted by more than 11,000 universities (https://www.ets.org/toefl/test-takers/ibt/about), and the IELTS serves more than 3.5 million people a year and is accepted by more than 10,000 institutions (British Council, 2017). Although the TOEFL is the most widely accepted worldwide, with the IELTS in close second, many universities accept other ELP assessments, such as Cambridge Assessment English (CAE) and the Pearson Test of Academic English (PTE). Each assessment quantifies English skill, but there are still differences in the assessments; according to a review by Wood (2022), the TOEFL focuses on academic English, whereas the IELTS assesses both academic and everyday English; another view is that the tests differ on task types (see Bright, 2020). Although every English assessment for admissions is produced by a different company, each

company has provided evidence that scores from their assessments are indicative of underlying academic English ability. This collection of evidence allows admissions decision-makers the flexibility to accept multiple exams for the same purpose, that is, understanding how well applicants can read, write, listen, and speak in academic English.

Despite the clearly stated purpose of these assessments, they may be interpreted as predictors of academic success because they exist in the same realm as other more general admissions tests such as the Scholastic Aptitude Test (SAT) or the Graduate Records Examination (GRE). For example, Arcuino (2013) researched admissions practices for three universities in the United States and found that students slightly below the cut-off in terms of undergraduate grade point average (GPA) and GRE scores were admitted with a strong TOEFL or IELTS score. Predicting criteria other than academic English with scores from these assessments is neither explicitly supported by ETS (the creators of the TOEFL; Educational Testing Service [ETS], 2018), nor by the creators of the IELTS (n.d.). Consequently, there is a need to collect some form of validity evidence to support this interpretation.

Predictive evidence is a form of criterion-related validity evidence and has been traditionally defined as the degree to which some measures (in this case, an assessment) predict a criterion such as academic success (Cronbach & Meehl, 1955). According to the Standards for Educational and Psychological Testing (American Educational Research Association [AERA] et al., 2014), "if validity for some common or likely interpretation for a given use has not been evaluated . . . potential users should be strongly cautioned" (p. 23). Furthermore, "if a test score is interpreted for a given use in a way that has not been validated, it is incumbent on the user to justify the new interpretation for that use, . . . collecting new evidence, if necessary" (p. 24). The claim that a score on an academic English assessment predicts academic success has not been sufficiently or broadly validated. Thus, it is imperative that evidence is evaluated to determine if this interpretation is sufficiently supported by evidence.

The link between academic language and academic success is not direct, and one can imagine any number of scenarios where students who excel in academic English do worse in school due to poor quantitative skills, for instance. As noted in Cho and Bridgeman (2012), if the correspondence was one-to-one, native English speakers would all be successful in higher education, which is obviously not the case. Nonetheless, these assessments are being used for admissions decisions; so in the present meta-analysis, we investigated studies that correlated scores on admissions language assessments to measures of academic success (e.g., GPA). Meta-analytic techniques aggregate the findings of primary research and provide necessary validity evidence to support the salient interpretation that English entrance exams are predictors of academic success. Thus, in the present meta-analysis, we synthesized the evidence to support or oppose the claim that different admissions English assessments (a) predict success in higher education and (b) do so comparably.

Literature review

Primary research has demonstrated markedly mixed results when predicting academic success from English admissions exams. Although many researchers have had success finding a strong correlation with various metrics of academic achievement (Daller &

Phelan, 2013; Koys, 2010), others had little or none (Dooey & Oliver, 2002; Person, 2002). In some cases, although rare, researchers have even found negative correlations between these assessments and academic success (Arcuino, 2013; Cotton & Conrow, 1998). These studies were built upon each other, citing the discrepancies as justification for more primary research with the body of research surrounding the predictive evidence of these assessments ever-increasing.

Given the expansive selection of primary research on this topic, there is no surprise that several qualitative and meta-analytic syntheses have been conducted related to these assessments. Two prior meta-analyses have explored the capability of the TOEFL to predict academic success in higher education (Abunawas, 2014; Wongtrirat, 2010). The first, by Wongtrirat (2010), contained 22 effect sizes from 22 studies and focused on undergraduate and graduate international students studying in the United States. Abunawas (2014) expanded this analysis and found 47 TOEFL effect sizes across 41 studies, with school level, TOEFL version, and school location as moderators. In both studies, the authors found TOEFL to have a positive but small correlation (Wongtrirat, $r = .18$; Abunawas, $r = .21$) with GPA. More recently, Gagen (2019) linked the IELTS to GPA in a meta-analysis with 29 effect sizes across 18 studies, finding a similar positive small correlation ($r = .23$). Lastly, Pearson (2021) conducted a review of the methods of predictive IELTS studies, finding that, methodologically, there was a large degree of heterogeneity between studies. However, this was not true for every variable: 90% of primary IELTS studies had GPA as the outcome variable, and 84% of studies employed correlational methods.

Only a few primary studies have compared the predictive power of different assessments. Of the few that have, some did not find any difference between TOEFL and IELTS in predicting final cumulative GPA (Arcuino, 2013; Johnson & Tweedie, 2017, 2021; Lahib, 2016). On the contrary, Hill et al. (1999) found IELTS to have a significantly stronger correlation with first year GPA than TOEFL. In that vein, independently, the results of Abunawas (2014), Wongtrirat (2010), and Gagen (2019) indicated that IELTS and TOEFL similarly predict academic success, but no prior meta-analyses have combined different assessments and treated assessment type as a moderator.

In the context of meta-analytic research on other admissions assessments, the correlations of .18–.23 (Abunawas, 2014; Gagen, 2019; Wongtrirat, 2010) are comparable with some, although far weaker than others. One of the highest correlates of undergraduate GPA, the American College Testing (ACT), has a strong correlation by Cohen's (1988) standards (operational validity $\hat{\rho} = .51$; Westrick et al., 2015). In contrast, estimators of graduate success are slightly weaker, but still stronger than admissions English assessments: the GRE ($\hat{\rho} = .27$–$.38$; Kuncel et al., 2010) and the Graduate Management Admission Test (GMAT) ($\hat{\rho} = .35$–$.47$; Kuncel et al., 2007). High-school and undergraduate GPA ($\hat{\rho} = .58$ and .31) have also been shown to be stronger predictors of undergraduate and graduate success, respectively. Thus, although English assessments for admissions may predict success in higher education, other better suited assessment scores should be used also as part of a holistic admissions review process.

Study objectives and rationale

Prior meta-analytic research on the predictive evidence of admissions English assessments has been fairly narrow and only focused on either the TOEFL (Abunawas, 2014; Wongtrirat,

2010) or the IELTS (Gagen, 2019) without considering the large variety of other commercially available assessments. Because there is little research comparing the predictive evidence of different assessments, when used for this purpose, admissions decision-makers must make the unsubstantiated assumption that each assessment is equivalent in this regard. This is of particular interest to the admissions decision-makers at post-secondary institutions, who may apply the results of this study to improve their selection process. Additionally, these prior studies also did not consider the age of the research. The current Internet-based iterations of the TOEFL and IELTS were released in 2005, so there may be issues in comparing more current research to that released before 2005.

Considering the meta-regression, prior meta-analyses were all limited that their moderators were not modeled concurrently; thus, the results do not reflect the covariance between moderators when estimating their significance. Furthermore, only Abunawas (2014) included publication bias as a moderator. Finally, although previous researchers used random-effects approaches to weigh the impact of each study on the average effect size (Abunawas, 2014; Gagen, 2019), they did not employ procedures to account for the dependencies in standard errors within each study, leading to a potentially inflated type I error rate (Becker, 2000).

This study was built upon and expanded on previous meta-analytic research on the predictive evidence of admissions English assessments on academic success. By combining research from multiple assessments, this allowed for (a) a general statement on the association between admissions English assessments and academic success, and (b) a comparison between different assessment types after accounting for moderators such as school type (public vs. private), school level (undergraduate vs. graduate), and publication bias. In fact, this was the first meta-analysis to consider whether schools are public or private, even though significant differences have been found in the academic success of students at different institution types (Scott et al., 2006). Because the current iterations of the most popular assessments (TOEFL and IELTS) began offering their Internet-based assessments in 2005, we only considered research in which samples were collected after that year. This has the potential to make the results of this study more meaningful to admissions decision-makers than prior research, given current ELP assessment characteristics. Finally, we accounted for effect size dependencies within studies when calculating standard errors to minimize type I errors.

We utilized rigorous meta-analytic methods as outlined in the PRISMA framework (Moher et al., 2009) to answer the following research questions:

Research Question 1: To what degree does performance on a standardized admissions ELP assessment predict GPA in higher education?

Research Question 2: To what degree is this prediction moderated by school level (graduate vs. undergraduate), publication type, and school type (public vs. private university)?

Research Question 3: After controlling for significant moderators, are there differences in predictive evidence between ELP admissions exams?

To support admissions decision-makers, we provided (a) the average correlation between ELP exam scores and GPA, (b) the magnitudes of potential moderators of that

correlation, and (c) a comparison of the predictive evidence of different assessments. Colleges and universities around the world may find the results of this study valuable when deciding whether the proposed use of predicting academic success from these assessments is supported, and if so, whether this prediction should be considered more strongly for certain assessments. This study was designed so that the results could be generalized to students at English-speaking public and private higher education institutions around the globe (dependent on the availability of primary research and the representativeness of the sample).

Methods

An abridged methods section is presented here. A comprehensive methods section in line with rigorous PRISMA guidelines (Moher et al., 2009) is included in Appendix A of the supplementary materials.[2]

Search strategy

Two databases in the field of education and educational psychology were searched for relevant studies: Education Resources Information Center (ERIC) (through EBSCO) and Education Source (through EBSCO) using controlled searches. A reproduction of the search used in ERIC is produced in Appendix C of the supplementary file, along with the number of primary studies it produced. Search terms used in this search included "predict*" (using an asterisk in a database search expands the search by including all forms of the word predict, such as predictive and predicting), "validity," "undergraduate," "graduate," and the names of all the exams that met the inclusion criteria. Several other search strategies were employed. First, the reference lists for the qualitative and meta-analytic syntheses published by Wongtrirat (2010), Abunawas (2014), Gagen (2019), and Pearson (2021) served as the initial literature sources. Next, Google Scholar was searched with the same terms employed in the database search. As publication bias poses a risk to the quality of meta-analyses (Borenstein et al., 2009), unpublished literature were searched for in the ProQuest Digital Thesis database and on OSF preprints using the same search terms. Finally, the ETS and the IELTS research repositories were searched. Subsequent forward and backward citation searches were conducted on Google Scholar to identify other relevant sources.

Inclusion and exclusion criteria

To be included, the primary study had to correlate an international or English learner's score on one of several (a) large-scale, (b) standardized, (c) commercially available, (d) securely, and (e) currently administered English language proficiency assessments (a full list is provided in the section below) with their academic achievement in the school they were admitted to as measured by some form of GPA (also described below).

Assessments. To be included, the assessment in the study had to be used for admissions decisions at a college (2 year, 4 year, or technical) or university where English is the

primary written and spoken language of instruction. Studies including the following commercially available assessments were included:

- The Test of English as a Foreign Language (TOEFL iBT or TOEFL PBT),
- Any version of the International English Language Testing System (IELTS),
- The Pearson Test of Academic English (PTE),
- Cambridge English: Advanced (CAE) or C1,
- Cambridge English: Proficiency (CPE) or C2,
- Canadian Academic English Language Assessment (CAEL),
- The Examination for the Certificate of Proficiency in English (ECPE).

Studies that utilized internally developed language assessments administered by colleges and universities (e.g., Lee & Greene, 2007) and studies investigating English language exams that were not developed for admissions purposes (e.g., Daller & Phelan, 2013) were excluded. Other assessments not listed here were also excluded.

Given that the IELTS and TOEFL are the most widely available and accepted exams, one extra constraint was placed on study inclusion with these exams in mind. In 2005, ETS discontinued the TOEFL CBT (computer-based test) and introduced the TOEFL iBT (Internet-based test) alongside the original paper-based test (PBT) (ETS, 2020). In the same year, the developers of the IELTS introduced a computerized version of their assessment for the first time (Green, 2007). Today, both exams are offered either digitally or on paper. To strengthen the findings of this synthesis, only samples of students who took an ELP entrance exam after 2005 were included. Studies that were conducted after this date but analyzed historic data sets were excluded (e.g., Itaya et al., 2008).

Participants. The target population included students who were required to take one of the listed assessments to gain admission to a college or university and attended that college or university for at least one term. Mainly, these were students from a primarily non-English speaking country applying to a school within their country where English is the primary language of instruction (e.g., O'Dwyer et al., 2018).

Academic achievement. There are many measures of student success, such as satisfaction, career success, persistence and achievement of learning outcomes, but the most common metric in educational research is GPA (York et al., 2015). This is not without controversy: there have been researchers both supporting (e.g., Gershenfeld et al., 2016) and opposing (e.g., Young, 1990) the use of GPA as a metric of success. Despite the somewhat controversial nature of the variable, it is prevalent in relevant predictive validity studies and prior meta-analytic research (Abunawas, 2014; Gagen, 2019; Wongtrirat, 2010) likely because it is the most accessible academic outcome metric (Cho & Bridgeman, 2012). Throughout this study, the following terms are used interchangeably: student success, student achievement, and GPA.

Researchers of included studies measured academic achievement using GPA, or another numeric outcome that could be converted to GPA, as the outcome variable. For example, Gochev (2013) correlated IELTS scores with a non-traditional GPA measured on a 100-point scale. The time frame could be any time in which students were attending

the college or university for which they applied with the corresponding assessment. Included studies could report (a) GPA for one or two semesters (e.g., Arrigoni & Clark, 2015), (b) cumulatively over any amount of time (e.g., Johnson & Tweedie, 2017), and (c) within one or multiple majors or course loads (e.g., Müller & Daller, 2019). Studies with qualitative metrics of success, such as a survey or a teacher evaluation, were also excluded.

Study quality and design. To be included, researchers conducting primary studies were required to rely on observational (i.e., existing student data were analyzed) research designs. In addition, two pieces of information were required: (a) the correlation between test score and GPA or the information needed to compute the correlation, and (b) the sample size, in order to calculate the standard error of the correlation.

Screening and coding procedure

The initial search was completed by the primary author and the subsequent screenings were conducted with the aid of a PhD student who had taken a graduate-level course on meta-analysis. A free systematic review website, Rayyan (Ouzzani et al., 2016), was utilized by both raters to identify studies for inclusion based on title/abstract review. After the primary search in 2018, all articles were reviewed by both raters. The full text of each study identified in the primary screening process was retrieved and reviewed in its entirety to ensure it met the inclusion criteria. The percentage of agreement for article inclusion was 100%, indicating that both reviewers agreed entirely on the selection of the final 32 included articles.

Coded variables

The following categories of variables were coded: (a) study, (b) assessment, (c) sample, (d) school, and (e) effect size variables. A description of each of these categories is provided in the following paragraphs.

Study variables. Several identifying variables related to the study itself were recorded, including first author and year of publication. Whether the article was published in a peer-reviewed journal was also coded dichotomously to allow for publication bias analyses.

Assessment variables. The central moderator variable for this study was the assessment being evaluated. Because many of these tests have multiple versions (e.g., TOEFL PBT vs. TOEFL iBT), this information was also recorded. For the IELTS, there was no distinction between paper-based and computer-based assessments, so this was coded solely as IELTS throughout.[3]

Sample variables. Prior researchers noted the difference in rigor between undergraduate and graduate-level programs (Fu, 2012; Woodrow, 2006), so a categorical moderator recorded whether each sample contained primarily participants studying at the

undergraduate level, graduate level, or a combination of both. Additionally, prior researchers noted that the majority of international students come from China and India (IIE, 2018), so the primary country of origin for the sample was coded. Because of the diversity covered by primary research, this variable was coded only for descriptive purposes only.

School information. The majority of predictive validity research considered student success at a single institution, which was noted for descriptive purposes, as was the country in which the school was located. Although Abunawas (2014) found that institutional setting (United States vs. non-United States) was marginally significant, it was not included as a moderator due to the fact that it was highly conflated with assessment type. That is, most research on the IELTS took place in a non-United States context, whereas most TOEFL research was conducted on samples of students in the United States. Another categorical variable described if the institution was public (i.e., government funded and publicly controlled) or private as there are many differences between public and private colleges, such as cost to attend, size of the institution, and other factors that are associated with student success in college (Scott et al., 2006). Woodrow (2006) observed that the majority of studies on predictive validity focused solely on GPA in the first semester of college. As noted by Gagen (2019), GPA must be considered throughout the collegiate experience to determine the true predictive validity of assessments. Thus, a dichotomous variable was included to describe the type of GPA (one/two terms vs. cumulative).

Effect size variables. A number of variables related to effect sizes were included for either descriptive or meta-analytic purposes. The desired effect size was the correlation, r, between English entrance assessment score and college or graduate level GPA.[4] If there were multiple effect sizes (e.g., an effect size for students from different countries, or in different programs), each was coded separately. From any given study, the largest number of non-overlapping disaggregated samples were recorded. For instance, Shbeeb (2019) reported effect sizes at the aggregate level and disaggregated by program. For this study, the correlation for each program was collected. The number of participants was also noted in order to calculate effect size sampling error variances.

Interrater reliability

Coding for effect sizes and all moderators was completed by the first author. A doctoral student coded an approximate 20% overlap ($k=8$). Any inconsistencies were resolved through discussion, with the first author ultimately determining the resolution. The average percentage agreement between the two raters across all coded variables was 96%. Of the variables included in the moderator analysis, none had an interrater reliability below 87.5%. The only variables to suffer in terms of interrater reliability were country of origin and GPA type, which were included only descriptively (62.5% and 75% agreement, respectively). The interrater reliabilities for each coded variable can be found in Appendix E of the supplementary materials.

Statistical analyses

Calculating individual effect sizes. Although correlations can serve as an effect size, each outcome was converted to Fisher's *z* to normalize the sampling distribution (Fisher, 1915). The transformation from *r* to Fisher's *z* and the associated standard error are depicted here:

$$z = 0.5 \times ln\left(\frac{1+r}{1-r}\right) \tag{1}$$

$$v_z = \frac{1}{n-3} \tag{2}$$

where *n* is the size of the sample, *r* is the correlational effect size, and v_z is the sampling error variance of the estimated Fisher's *z*. Effect size comparisons and potential moderator analyses were completed using Fisher's *z*, although for presentation they were converted back to *r* to ease interpretation.

There were no range restriction corrections made to these data even though it is very likely that this occurred (i.e., students with high TOEFL scores were selected partially based on these high TOEFL scores and high school/undergraduate GPA, which correlated highly with TOEFL scores; Woodrow, 2006). Due to this, the overall calculated average effect size was likely an underestimation of the true mean effect size.

Missing data analysis. The *Amelia* package in R (Honaker et al., 2011) identified moderator variables that were missing from many studies, as well as the effect sizes that were missing many moderator variables. Missing moderators were obtained from other available data whenever possible (e.g., missing institution variables were gathered via institution websites) but other more formalized imputation methods were not implemented. When moderators could not be inferred, those effect sizes were removed from the moderator analysis. Missing data were a concern during the coding procedure due to the fact that the moderator model deletes any cases with missing values, a shortcoming that is elaborated on in the study limitations. Thus, moderator variables that were missing information for more than 10% of effect sizes were not included in the moderator analysis so as to retain as much information as possible. This percentage was chosen a priori based on the expertise of the authors.

Examining outliers and identifying publication bias. Figure 1 depicts a funnel plot which displays the effect size plotted against the inverse of the standard error. Studies with high standard error and effect sizes close to zero are less likely to be published, so an absence of data in this portion of the graph may indicate publication bias. Egger's test, which is a test of symmetry of the funnel plot (Egger et al., 1997), was also employed. A forest plot (Figure 2) was also produced in order to visualize effect sizes by study/sample. Publication status was also considered as a moderator in the meta-regression model described below; a significant effect of publication status would indicate notable publication bias.

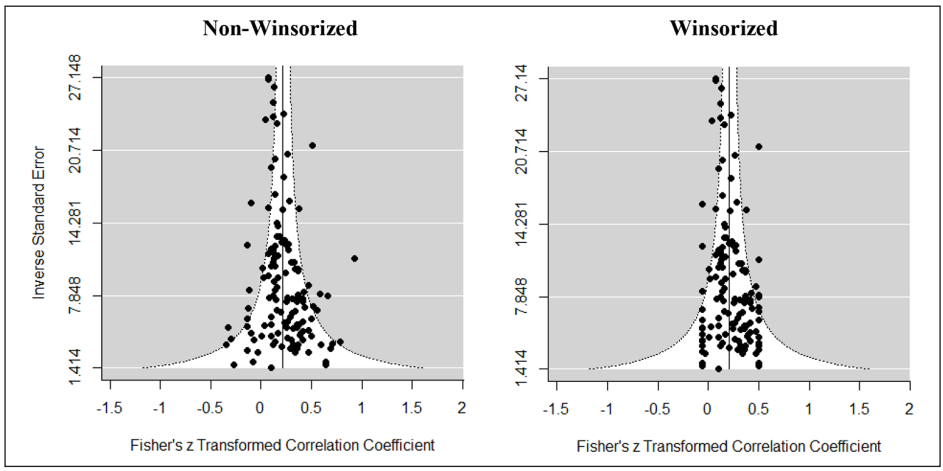

Figure 1. Funnel plots for the non-winsorized and winsorized effect sizes, plotted against inverse standard error.
Note. The dotted line represents the weighted mean effect size assuming effect sizes are all independent.

A sensitivity analysis was also conducted by comparing mean effect size with and without outliers.

Meta-regression. The mean effect size across all effect sizes was calculated with an intercept-only meta-regression using the *robumeta* R package (Fisher & Tipton, 2015). This approach uses the inverse variance of primary studies to weigh each study. Additionally, this model allowed for the evaluation of the degree of heterogeneity between the effect sizes. To determine the degree of between-study heterogeneity, a modified Cochran's Q, the I^2 statistic, was calculated using the following formula (Higgins & Thompson, 2002):

$$I^2 = \frac{Q-(k-1)}{Q} \times 100\% \tag{3}$$

where Q is Cochran's Q and k is the number of studies. The I^2 statistic represents the percentage of total variability in a set of effect sizes due to between-study variability. Another advantage of the I^2 statistic is the ease with which it can be interpreted. An $I^2 < 50\%$ indicates low heterogeneity, $50\% \leqslant I^2 < 75\%$ medium heterogeneity, and $I^2 \geqslant 75\%$ large heterogeneity (Higgins & Thompson, 2002). Thus, if the null model has medium or high heterogeneity, a moderator analysis is warranted.

Effect sizes in this analysis from within the same study may be correlated with one another, as they typically sample different subpopulations of students (e.g., students from different countries of origin) at the same institution. Effect size dependency is a threat to the validity of interpretations of a meta-analysis, as it can increase type I error by deflating standard error estimates (Becker, 2000). To account for this, the *robumeta* package (Fisher & Tipton, 2015) employs robust variance estimation (RVE) procedures (Hedges

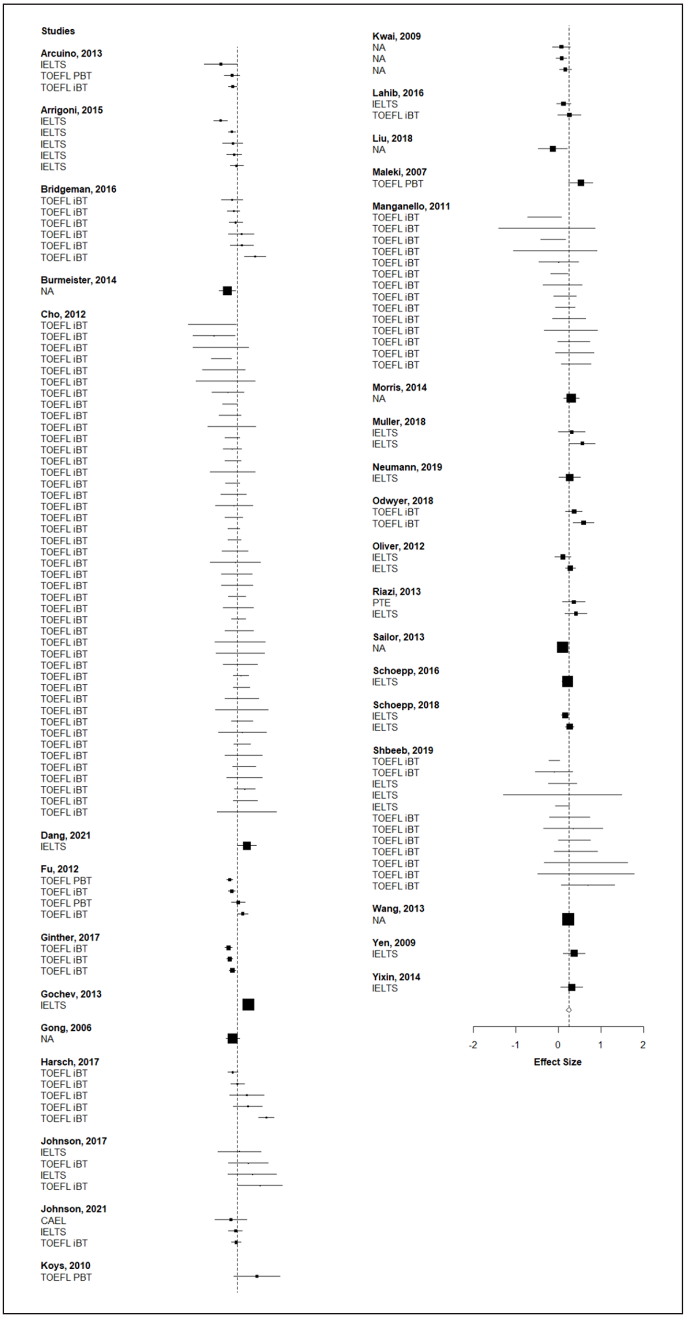

Figure 2. A forest plot displaying all 132 effect sizes from 32 studies.

Note. The dotted line represents the weighted mean effect size determined from the null model. Point size is determined by sample size of the effect.

et al., 2010) to estimate the variance of effect sizes. This approach approximates the dependence structure of effect sizes from the same study by estimating the within study covariances using the cross products of the residuals from within a given study.

Moderator analysis. A meta-regression model was fit with the following moderators:

$$\hat{z} = \beta_0 + \beta_1 (Private) + \beta_2 (Graduate) + \beta_3 (Published) + \beta_4 (TOEFL\,iBT) + \beta_5 (TOEFL\,PBT) + \epsilon \quad (4)$$

where β_0 is the intercept, β_1 is the partial slope associated with school status (reference is public), β_2 is the partial slope associated with school level (reference is undergraduate), and β_3 is associated with publication status (reference is unpublished). Coefficients β_4 and β_5 are associated with different assessments (reference is IELTS).

For both the null and moderator analyses, we have opted to employ a fixed-effects meta-regression model. Although this approach has been criticized for being less generalizable than a random-effects model (e.g., Gagen, 2019), there is still reason to believe that it is appropriate given the aims of this study (Rice et al., 2018). Rice et al. (2018) make the argument that, under certain minimal assumptions, fixed-effects analyses are equally useful in their ability to estimate mean effects. The *robumeta* package further mitigates any assumption violations by correcting for small sample sizes in their estimates (Fisher & Tipton, 2015).

Results

A summary of the literature search and screening process can be found in Figure 3 (Moher et al., 2009, The PRISMA Group). The search of ERIC and Education Source yielded 237 studies. Other sources, namely ProQuest, publisher websites for the TOEFL and IELTS, Google Scholar, and the citations from Wongtrirat (2010), Abunawas (2014), and Gagen (2019) led to 269 citations after duplicates were removed. Backward and forward citation searches yielded 16 additional studies that met the inclusion criteria, bringing the total number up to 51. These studies were subsequently narrowed down to 62 studies that met the inclusion criteria based on title and abstract screening. Final full-text screening resulted in a sample of 32 studies, including published journal articles, research reports, and unpublished dissertations.

The largest reason to exclude studies was age, with 38 otherwise relevant studies excluded during the initial screening because they were published prior to 2006. After the initial test screening, four more studies were excluded because, even though they were published beyond 2006, they referenced older samples. The other rejected studies consisted of research reviews, reviews of excluded assessments, non-correlational studies, and studies with prior high school or college GPA as an independent variable rather than a subsequent higher education GPA as a dependent variable. One source (Avdi, 2011) that could have been included based on title and abstract was excluded due to lack of access to the full paper.

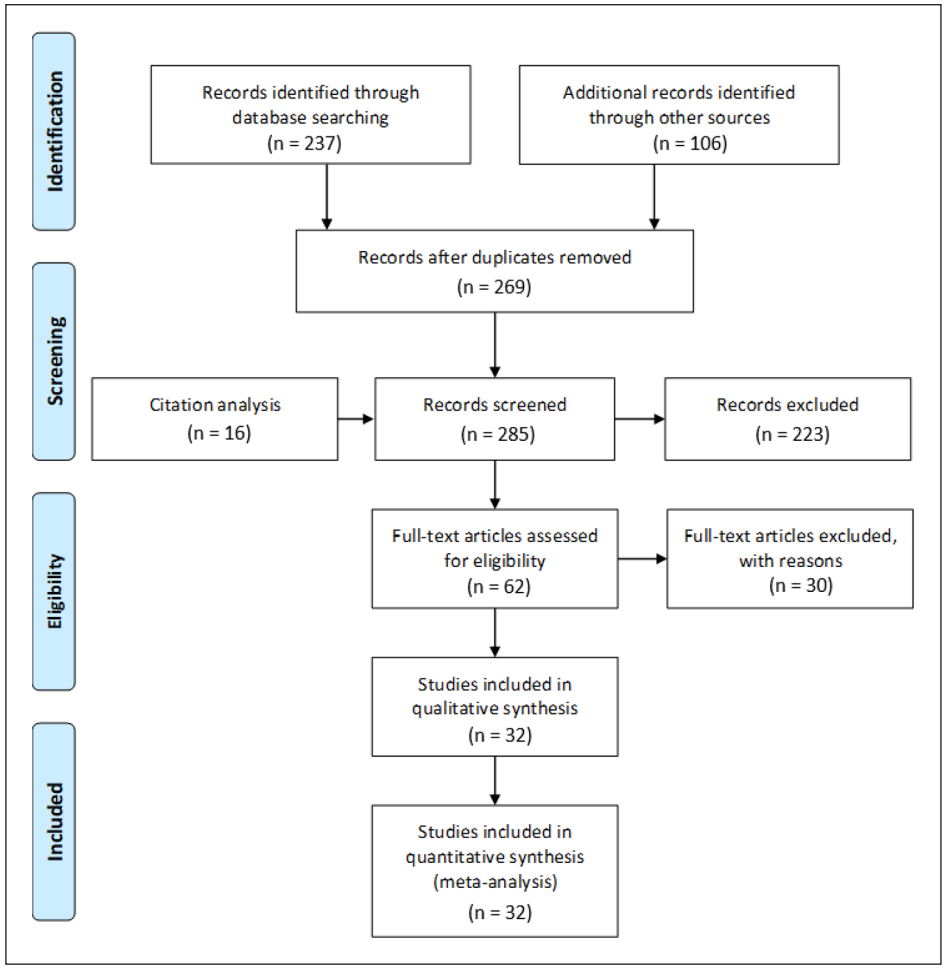

Figure 3. A PRISMA flow diagram detailing the search and screening process of the meta-analysis.

Coding results

Description of sample. All studies included in the meta-analysis are cited in Appendix B of the supplementary file and coded information from each is included in Appendix D of the supplementary file. The final sample of 32 studies yielded 132 effect sizes and ranged from the year 2006 to the year 2021, with the median study year being 2014. Studies contributed anywhere from 1 to 44 effect sizes, with a median of two effect sizes. The total number of participants across all studies was 15,691. Of the 132 effect sizes, 75 came from specific programs or schools, business being the most common field ($n = 17$), followed by some form of social science ($n = 13$). Nine countries produced primary research, the most common being the United States ($n = 92$ effect sizes) with the United

Kingdom in a distant second ($n = 9$). Most studies specified that their sample comprised students from all over the world ($n = 55$), although a small number of samples were primarily made up of Chinese students ($n = 12$). The TOEFL iBT and the IELTS were the main language assessments represented in the sample. The TOEFL iBT contributed 90 effect sizes (68.2%), whereas the IELTS contributed 26 effect sizes (19.7%). The TOEFL PBT contributed five effect sizes (3.8%). Two other exams, Canadian Academic English Language (CAEL) and PTE, were both represented with one effect size. The distribution of effect sizes can be seen in Figure 2. Other moderator variables included in the quantitative analysis, such as school level, are described below with the moderator analysis.

Most coded variables were available in every study. Every study identified the necessary effect size information, and the majority had information about the desired moderators. A total of nine studies did not specify which version of assessments they were using or converted scores from different assessments to the same scale. In terms of other moderators, 11 studies did not divulge the name of the school in question and were missing potentially identifying information on whether the schools were public or private. One moderator that was ultimately excluded was GPA type. Although it had been considered in prior meta-analyses (Gagen, 2019), it was missing from sufficient studies to warrant its exclusion from the meta-regression model. Of the studies that did include GPA type, 58% of effect sizes were based on one/two term GPA, whereas the rest were collected cumulatively at the end of the students' program.

Outlier analysis. Six effect sizes were identified as outliers, four of those indicating a large negative correlation between assessment score and academic success. A sensitivity analysis yielded no difference between Winsorized and non-Winsorized data sets (the difference in mean effect sizes was 0.005). Consequently, effect sizes were not Winsorized as it did not dramatically impact the presented outcomes.

Meta-regression

English university admission assessments (the TOEFL and IELTS, for the most part) had an average correlation of .23 with GPA (95% CI [.18, .28]). Overall, this mean correlation was found to be significantly different than zero ($p < .001$). This result was based on all 132 effect sizes from 32 studies. The results of this analysis can be found in Table 1. The meta-regression results indicated a moderate degree of heterogeneity present within the observed effect sizes ($I^2 = 68.53\%$), suggesting the need for moderator analyses.

Moderator analysis. The results of the moderator analysis are also depicted in Table 1. Because only one study analyzed the PTE and CAEL, they were excluded from the moderator analysis. Thus, the moderator analysis was conducted with the 111 TOEFL and IELTS effect sizes for 29 distinct samples. Overall, the majority of effect sizes came from public institution data sets ($n = 100$, 82%).[5] For both the IELTS and TOEFL assessments, the majority of research took place in public institutions, but IELTS research was more likely to occur in private settings (29% of IELTS research, 14% of TOEFL research). Whether a school was public or private did not significantly explain any of the heterogeneity in the effect size ($p > .05$). Similarly, graduate vs. undergraduate also failed to be a

Table 1. Results of null model and moderator analysis for meta-regression of GPA onto college English entrance exam assessment score.

Moderator	Unconditional model coefficient and *SE*	Moderator model coefficient and *SE*
Institution:		
Private[a]		.076 (.086)
Program:		
Graduate[b]		.037 (.079)
Test:		
TOEFL iBT[c]		–.048 (.070)
TOEFL PBT[c]		.035 (.151)
Publication Status:		
Published[d]		–.018 (.066)
Constant	.232* (.025)	.100* (.024)
Effect sizes (*n*)	132	111
Number of studies (*k*)	32	29
I^2	68.53	62.63

iBT: Internet-based test; PBT: paper-based test; SE: standard error.
Note. *$p < .05$.
[a]Reference is public, [b]reference is undergraduate, [c]reference is IELTS, [d]reference is unpublished.

significant moderator ($p > .05$). The collected effect sizes fairly evenly represented undergraduate and graduate students (53% graduate students), although this difference is more pronounced when considering assessment brand. For IELTS effect sizes, 63% of research came from undergraduate institutions schools; in contrast, 60% of TOEFL effect sizes were found in graduate programs. The majority of research on admissions English assessments took place at public graduate schools (50.4% of effect sizes). After accounting for moderators, admissions English assessments were positively correlated with academic success ($p = .001$) although no significant differences were found between each of the tests ($p > .05$); thus, the unconditional model average correlation of .23 is consistent across instruments. Based on the I^2 statistic, the model containing all moderators explained 8.6% more variation than the null model ($I^2 = 62.63$).

Publication bias. Publication bias poses a serious risk to the validity of meta-analyses, and as such, great care was taken to evaluate its presence. Of the 132 included effect sizes, 58 came from 14 unpublished works (44%). More specifically, eight studies ($n = 40$ effect sizes) were dissertations or theses, three were reports ($n = 12$), two were chapters ($n = 5$), and one was an unpublished manuscript ($n = 1$). The presence of publication bias was evaluated with funnel plots (Figure 2) and by Egger's test. The funnel plot appears to be symmetric across the median, and this is confirmed by Egger's test ($z = 1.63$, $p = .1$) indicating that publication bias was likely not present in the analysis. Additionally, publication status was considered as a moderator but was not found to be statistically significant after accounting for school level, type, and assessment type ($p > .05$).

Discussion

Overall, the results of this meta-analysis support the conclusion that admission English assessments do predict academic success in undergraduate and graduate school. This result is in line with previous meta-analyses analyzing just the TOEFL (Abunawas, 2014; Wongtrirat, 2010) or IELTS (Gagen, 2019). However, this conclusion should not be considered in a vacuum; although the correlation was positive and significant ($r = .23$, $p < .001$), it was lower than correlations from other more general admissions assessments, such as the ACT (Westrick et al., 2015) and the GRE (Kuncel et al., 2010). One must also consider these results in the context of other studies of language testing. Plonsky and Oswald (2014) synthesized the results of 175 correlational studies in the field of language testing, concluding that correlations of .25 are small, .40 medium, and any above .6 are large. By these standards, English assessments for admissions predict academic success to only a small degree.

Although it was found to be significant in Gagen (2019), similar to Abunawas (2014), level of study (graduate vs. undergraduate) was found to be non-significant. The contradictory nature of these findings could indicate that (a) after accounting for the other moderators in this study, school level was no longer significant or (b) there is a difference between the TOEFL and IELTS' ability to predict success at the graduate and undergraduate level that was not captured in the current study. This study was the only meta-analysis thus far to consider whether schools were public or private. Somewhat surprisingly, given the differences in the students who attend these institutions (Scott et al., 2006), after accounting for other moderators, school type was not found to be significant. This may be due to any number of other factors, such as grade inflation or differences in the types of majors typical to the student body. Finally, none of the methods for identifying publication bias (funnel plot, Egger's test, moderator analysis) seemed to do so, strengthening the generalizability of this study.

After controlling for the aforementioned moderators, the TOEFL (both versions) and IELTS were not found to differentially predict success in undergraduate and graduate school. These results are unsurprising given the similarity of the findings by Abunawas (2014), Gagen (2019), and Wongtrirat (2010). Still, one may wonder why this is the case as these tests have been found to have as many differences as similarities (Li, 2018; Wood, 2022; see also Bright, 2020). In fact, Li (2018) noted that the TOEFL iBT and IELTS differ substantially in terms of content and tasks. Overall, some estimate that the TOEFL iBT focuses more on academic English, whereas the IELTS looks for general and academic language (see Wood, 2022). Nonetheless, score comparison guides for these assessments have been developed by ETS (https://www.ets.org/toefl/score-users/scores-admissions/compare). Due to this, these results may offer some reassurance to colleges that have been accepting these assessments interchangeably if they are being used to predict who will succeed at an institution.

Limitations

Due to the many complex components of conducting a meta-analysis, several aspects of this study could be improved in future iterations of this and similar studies. For instance,

during the data collection process some number of studies were excluded because they were not accessible through any of the databases. Assuming these studies were inaccessible not due to some feature of the study, it is reasonable to expect that they would not have greatly influenced the results. Despite this, more exhaustive attempts could have been made to contact study authors.

During the coding process, many more moderators could have been explored and coded, from participant demographics to outcome variations (i.e., moderating for GPA vs. other outcome measures, or controlling for types of GPA). Other predictors, such as GRE/SAT score or High School/Undergraduate GPA, could also enrich the analysis from the perspective of an admission decision-maker. Without the introduction of imputation techniques, one issue related to coding more moderators is that *robumeta* handles missing data by listwise deleting each study missing any moderator variable. This introduces bias unless the data are missing completely at random; this assumption is untestable and typically untenable. Additionally, without imputation, each study that was missing a single moderator value was excluded from the moderator analysis, decreasing the total sample size by 21 correlations from 3 studies. This also led to the exclusion of certain moderators (e.g., GPA type; Gagen, 2019) that were included in prior meta-analyses. In general, these model assumptions need to be examined more thoroughly in future analyses.

A number of other analytical procedures were not included in this study. Correcting for range restriction would yield more accurate correlations. A number of factors may contribute to range restriction issues: (a) schools select students with higher English assessment scores; (b) collegiate grades tend to be inflated; and (c) negative skew may be observed in GPA in primary studies. Nonetheless, as noted previously, range restriction corrections may inadvertently add bias based on the authors assumptions about the full non-range restricted population. Some methods have been proposed for correcting correlations for range restriction when populations parameters are unknown (Cohen, 1959), but these corrections are not widely employed in practice leading to questions of their accuracy. Similarly, predictive power has shown to weaken over time, so correlations with later outcomes may have been smaller. Another limitation is that there was no power analysis conducted in this study. Although post hoc power analyses are typically not recommended, one could have shed valuable light on the generalizability of these findings. A meta-analysis with low power may not offer any valuable insight into practical implications. More effect sizes would need to be included in order to increase the power of the analysis, particularly if more moderators are added to the regression models.

Future research

Much of the primary research in this field is lacking in important information that could be relevant to future meta-analytic research. One assumption of this study was that authors correctly identified which type of exam score they were reporting. This assumption may not be tenable as in a number of studies, IELTS and TOEFL scores were equated based on guidelines published by IELTS (e.g., Kwai, 2009). For those

studies where it was not clear, effect sizes were not included in the moderator analysis because they were missing exam type. Still, researchers should make an attempt to specify this as much as possible to improve the findings of future meta-analyses. Authors of future primary analyses should make an effort to better describe the sample they are analyzing whenever possible. Subgroup analyses by nationality (similar to those in Bridgeman et al., 2016), for instance, may lead to better conclusions about the predictive evidence of these assessments in more specific contexts. This increased specificity could be seen as a helpful tool for schools when weighing large numbers of applications for a limited number of positions. Additionally, the body of primary predictive evidence research must grow considerably before comparisons can be made to assessments other than the TOEFL and IELTS. Only one study identified in this search incorporated research related to assessments other than these, and although that research may exist in some form, perhaps with another outcome variable, it was not captured in the comprehensive literature search.

Implications

It is clear from the results of this meta-regression results that these exams positively correlate with GPA. Still, it is a small correlation, so admission decision-makers may be wise to continue to use it solely as a measure of English skills rather than academic success, or as one piece of non-predictive evidence in a holistic review process. This conclusion is especially true when considering other predictors that are more correlated with success, such as SAT, ACT, or GRE scores, or prior academic achievement.

The hypothesis analyzed through this study was that not all admission English assessments are created equal, which, based on the results of this meta-analysis of three instruments, appears to be untrue. Although this meta-analysis had shortcomings, its results are still generalizable and still carry some weight. The sample of studies represents all available studies from a systematic search of research analyzing the validity evidence of the TOEFL and IELTS in predicting some form of GPA, and thus generalizes to those assessments. The results of this analysis indicate that admissions decision-makers do not need to differentially weigh admissions English assessments. Although the TOEFL is the most popular of its ilk, the IELTS is gaining popularity in the United States and functions equivalently for predicting success in higher education. Due to the lack of other assessments represented in the sample, it is unclear if these results extend to exams such as the CAE and PTE, among others.

Author contributions

The first author conceived of the presented idea, conducted all parts of the research, and wrote the manuscript. The second author provided critical feedback and guidance at each stage of the writing process. All authors approved of the final version to be published.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Samuel Dale Ihlenfeldt https://orcid.org/0000-0003-3642-8407

Supplemental material

Supplemental material for this article is available online.

Notes

1. The exception to this trend is 2020. As an example, there was a 16% decrease in international student enrollment in the United States due to the persisting COVID-19 travel restrictions (Baer & Martel, 2020).
2. The R code and associated data can be found on OSF: https://osf.io/xw2c5/?view_only = f2b0aad9355948a7b42ee77525d1f9d7.
3. At the request of a reviewer, we also recorded whether studies reported an internal consistency for the sample, a practice that is standard in language testing research. Only one study (Müller & Daller, 2019) did so, reporting an alpha of .83. It is likely that this information was not reported because researchers did not have access to the individual examinee responses required to calculate internal consistency. Still, we recommend that future researchers seek out and present this information when possible.
4. The decision was made to include studies that reported either Pearson or Spearman correlations. Although these two correlations are not interchangeable, this is an approach that has been taken by prior meta-analyses. In the final sample, only three studies provided Spearman correlations.
5. Note that some cases were missing this information, so totals may not add up to the full 132 effect sizes.

References

Abunawas, M. (2014). *A meta-analytic investigation of the predictive validity of the Test of English as a Foreign Language (TOEFL) scores on GPA* [Unpublished doctoral dissertation]. Texas A&M. https://oaktrust.library.tamu.edu/bitstream/handle/1969.1/154156/ABUNAWAS-DISSERTATION-2014.pdf?sequence=1

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-test-english-as-foreign/docview/1413309058/se-2

Arrigoni, E., & Clark, V. (2015). *Investigating the appropriateness of IELTS cut-off scores for admissions and placement decisions at an English-medium university in Egypt*. (IELTS Research Reports Online Series No. 3). IELTS. https://www.ielts.org/en-us/for-researchers/research-reports/online-series-2015-3

Avdi, E. (2011). IELTS as a predictor of academic achievement in a Master's Program. *English Australia Journal*, *26*(2), 42–49.

Baer, J., & Martel, M. (2020). *Fall 2020 international student enrollment snapshot*. Institute of International Education. https://www.iie.org/Research-and-Insights/Open-Doors/Fall-International-Enrollments-Snapshot-Reports

Becker, B. J. (2000). Multivariate meta-analysis. In S. D. Brown & H. E. A. Tinsley (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499–525). Academic Press. https://doi.org/10.1016/B978-0-12-691360-6.X5000-9

Bodycott, P. (2009). Choosing a higher education study abroad destination: What mainland Chinese parents and students rate as important. *Journal of Research in International Education*, *8*(3), 349–373. https://doi.org/10.1177/1475240909345818

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. https://doi.org/10.1002/9780470743386

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

Bright, L. (2020, January 15). IELTS vs. TOEFL: What's the difference? *StudyinCanada.com* https://www.studyincanada.com/Discover/Article/1/5116/IELTS-vs-TOEFL:-What's-the-Difference?

British Council. (2017, September 8). *IELTS numbers rise to three million a year*. https://www.britishcouncil.org/organisation/press/ielts-numbers-rise-three-million-year

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. *Language Testing*, *29*(3), 421–442. https://doi.org/10.1177/0265532211430368

Cohen, A. C. (1959). Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*, *1*(3), 217–237. https://doi.org/10.1080/00401706.1959.10489859

Cohen, J. (1988). Set correlation and contingency tables. *Applied Psychological Measurement, 12*(4), 425–434.

Cotton, F., & Conrow, F. (1998). An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania. *International English Language Testing System (IELTS) Research Reports*, *1*, 72–115. https://www.ielts.org/-/media/research-reports/ielts_rr_volume01_report4.ashx

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Daller, M. H., & Phelan, D. (2013). Predicting international student study success. *Applied Linguistics Review*, *4*(1), 173–193. https://doi.org/10.1515/applirev-2013-0008

Dooey, P., & Oliver, R. (2002). An investigation into the predictive validity of the IELTS test as an indicator of future academic success. *Prospect*, *17*, 36–54.

Educational Testing Service. (2018). *How TOEFL® test scores are used*. https://www.ets.org/toefl/institutions/scores/use/

Educational Testing Service. (2020). *TOEFL® program history* [TOEFL® Research Insight Series, 6]. https://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v6.pdf

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *The BMJ, 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507–521.

Fisher, Z., & Tipton, E. (2015). *Robumeta: An R-package for robust variance estimation in meta-analysis* (Version 2.0) [Computer software].

Fu, Y. (2012). *The effectiveness of traditional admissions criteria in predicting college and graduate success for American and international students* [Unpublished doctoral dissertation]. The University of Arizona. https://www.proquest.com/docview/917742486

Gagen, T. (2019). *The predictive validity of IELTS scores: A meta-analysis* [Unpublished doctoral dissertation]. The University of Western Ontario. https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=8762&context=etd

Gershenfeld, S., Ward Hood, D., & Zhan, M. (2016). The role of first-semester GPA in predicting graduation rates of underrepresented students. *Journal of College Student Retention: Research, Theory & Practice*, *17*(4), 469–488. https://doi.org/10.1177/1521025115579251

Gochev, N. (2013). *Criterion-related validity of strong and weak second language performance assessments in a university pathway programme* [Unpublished doctoral dissertation]. Lancaster University. https://www.baleap.org/wp-content/uploads/2016/03/Dissertation_Nikolay_Gochev.pdf

Graham, J. G. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly*, *21*(3), 505–521. https://doi.org/10.2307/3586500

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education* (*Vol. 25*). Cambridge University Press.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Hegarty, N. (2014). Where we are now—The presence and importance of international students to universities in the United States. *Journal of International Students*, *4*(3), 223–235. https://files.eric.ed.gov/fulltext/EJ1054975.pdf

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558.

Hill, K., Storch, N., & Lynch, B. (1999). A comparison of IELTS and TOEFL as predictors of academic success. *International English Language Testing System (IELTS) Research Reports*, *2*, 62–73. https://www.ielts.org/-/media/research-reports/ielts_rr_volume02_report3.ashx

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(1), 1–47. https://doi.org/10.18637/jss.v045.i07

IELTS (n.d.). *IELTS test score guidance*. https://www.ielts.org/-/media/pdfs/ielts-test-score-guidance.ashx

Institute of International Education. (2018). *A world on the move: Trends in global student mobility*. https://p.widencdn.net/w9bjls/A-World-On-The-Move

Israel, E., & Batalova, J. (2021, January 14). *International students in the United States*. Migration Policy Institute. https://www.migrationpolicy.org/article/international-students-united-states-2020

Itaya, L. E., Chambers, D. W., & King, P. A. (2008). Analyzing the influence of admissions criteria and cultural norms on success in an international dental studies program. *Journal of Dental Education*, *72*(3), 317–328. https://doi.org/10.1002/j.0022-0337.2008.72.3.tb04498.x

Johnson, R. C., & Tweedie, M. G. (2017). A comparison of IELTS, TOEFL, and EAP course results as predictors of English language learning success in an undergraduate nursing program. In C. Coombe, P. Davidson, A. Gebril, D. Boraie & S. Hidri (Eds.), *Language assessment in the Middle East and North Africa: Theory, practice and future trends* (pp. 36–53). TESOL Arabia.

Johnson, R. C., & Tweedie, M. G. (2021). "IELTS-out/TOEFL-out": Is the end of general English for academic purposes near? Tertiary student achievement across standardized tests and general EAP. *Interchange*, *52*(1), 101–113. https://doi.org/10.1007/s10780-021-09416-6

Koys, D. (2010). GMAT versus alternatives: Predictive validity evidence from Central Europe and the Middle East. *Journal of Education for Business*, *85*, 180–185. https://doi.org/10.1080/08832320903258618

Kuncel, N. R., Credé, M., & Thomas, L. L. (2007). A meta-analysis of the predictive validity of the Graduate Management Admission Test (GMAT) and undergraduate grade point average (UGPA) for graduate student academic performance. *Academy of Management Learning & Education*, *6*(1), 51–68. https://doi.org/10.5465/amle.2007.24401702

Kuncel, N. R., Wee, S., Serafin, L., & Hezlett, S. A. (2010). The validity of the Graduate Record Examination for master's and doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement*, *70*(2), 340–352. https://doi.org/10.1177/0013164409344508

Kwai, C. K. (2009). *Model of international student persistence: Factors influencing retention of international undergraduate students at two public statewide four-year university systems* [Unpublished doctoral dissertation]. University of Minnesota, Minneapolis. https://search-proquest-com.ezp2.lib.umn.edu/dissertations/docview/305213878/fulltextPDF/66262DE91244947PQ/1?accountid=14586

Lahib, S. (2016). *Testing tensions: The use of English Language Proficiency tests for the admission of Ontario high school applicants to one Ontario university* [Unpublished doctoral dissertation]. The University of Western Ontario. https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=5646&context=etd

Lee, Y. J., & Greene, J. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research*, *1*(4), 366–389. https://doi.org/10.1177/1558689807306148

Li, Y. (2018). A comparison of TOEFL iBT and IELTS reading tests. *Open Journal of Social Sciences*, *6*(8), 283–309. https://doi.org/10.4236/jss.2018.68023

Mazzarol, T., & Soutar, G. (2002). "Push-Pull" factors influencing international student destination choice. *International Journal of Educational Management*, *16*(2), 82–90. https://doi.org/10.1108/09513540210418403

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

Müller, A., & Daller, M. (2019). Predicting international students' clinical and academic grades using two language tests (IELTS and C-test): A correlational research study. *Nurse Education Today*, *72*, 6–11. https://doi.org/10.1016/j.nedt.2018.10.007

O'Dwyer, J., Kantarcıoğlu, E., & Thomas, C. (2018). An Investigation of the Predictive Validity of the TOEFL iBT® Test at an English-medium university in Turkey. *ETS Research Report Series*, *2018*(1), 1–13. https://doi.org/10.1002/ets2.12230

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*, *5*, Article 210. https://doi.org/10.1186/s13643-016-0384-4

Pearson, W. S. (2021). The predictive validity of the Academic IELTS test: A methodological synthesis. *International Journal of Applied Linguistics*, *172*(1), 85–120. https://doi.org/10.1186/s13643-016-0384-4

Person, N. E. (2002). *Assessment of TOEFL scores and ESL classes as criteria for admission to career & technical education and other selected Marshall University graduate programs* [Unpublished master's thesis]. Marshall University. https://files.eric.ed.gov/fulltext/ED473756.pdf

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Rice, K., Higgins, J. P., & Lumley, T. (2018). A re-evaluation of fixed effect (s) meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(1), 205–227. https://doi.org/10.1111/j.1467-985X.2008.00552.x

Sá, C. M., & Sabzalieva, E. (2018). The politics of the great brain race: Public policy and international student recruitment in Australia, Canada, England and the USA. *Higher Education*, *75*(2), 231–253. https://doi.org/10.1007/s10734-017-0133-1

Scott, M., Bailey, T., & Kienzl, G. (2006). Relative success? Determinants of college graduation rates in public and private colleges in the US. *Research in Higher Education*, *47*(3), 249–279. https://doi.org/10.1007/s11162-005-9388-y

Shbeeb, R. (2019). *The relationship of English language scores on international students' academic success* [Unpublished doctoral dissertation]. University of Central Florida. https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=7577&context=etd

Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M., & Schmidt, F. L. (2015). College performance and retention: A meta-analysis of the predictive validities of ACT® scores, high school grades, and SES. *Educational Assessment*, *20*(1), 23–45. https://doi.org/10.1080/10627197.2015.997614

Wongtrirat, R. (2010). *English language proficiency and academic achievement of international students: A meta-analysis* [Unpublished doctoral dissertation]. Old Dominion University. https://digitalcommons.odu.edu/cgi/viewcontent.cgi?article=1183&context=efl_etds

Wood, S. (2022, July 8). The complete guide to the TOEFL test. *U.S. News*. https://www.usnews.com/education/best-colleges/articles/the-complete-guide-to-the-toefl-test

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

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical Assessment, Research & Evaluation*, *20*, Article 5. https://doi.org/10.7275/hz5x-tx03

Young, J. W. (1990). Adjusting the cumulative GPA using item response theory. *Journal of Educational Measurement*, *27*(2), 175–186. https://doi.org/10.1111/j.1745-3984.1990.tb00741.x