Received 7 January 2021
Revised 1 March 2021
Accepted 1 March 2021

Evaluating the efficacy of English proficiency on initial semestral results for higher education L1 English speakers in a multilingual context

Lyndon Lim and Wenjin Vikki Bo
Singapore University of Social Sciences, Singapore, Singapore

Abstract

Purpose – The purpose of this paper is to evaluate an English Proficiency (EP) programme and its efficacy with respect to students' academic performance in a university within a multi-lingual context, as the programme had been in effect for some years.

Design/methodology/approach – A quasi-experimental approach was used to study the efficacy of an EP programme in a university within a multilingual context. Data across two academic years were used, along with regression discontinuity design.

Findings – Results suggest that the EP programme had a significant and positive intervention effect on students' initial semester grade point average. The programme effect size was found to be medium to large.

Research limitations/implications – It might be useful to extend the study for one more year for more concrete conclusions. As the study was anchored upon the structure of the 2016 EP programme, any major curricular/structural change to the programme warrants another study.

Practical implications – This study demonstrated that the implementation of EP programmes in higher education institutions is essential not only for international students who are foreign language speakers of English but also for domestic students in English-speaking countries, especially for bi/multilingual speakers.

Originality/value – Previous studies related to the efficacy of EP within higher education have focused on international students who speak English as an additional/foreign language. Further, most studies have focussed on students' self-reported experiences and have yielded disparate findings. This study contributes to scholarship as it addresses the under-researched area related to domestic students who speak English as the first language in a bi/multi-lingual context.

Keywords Programme efficacy, English proficiency, L1 English speakers in multilingual contexts, Higher education

Paper type Research paper

Introduction

English proficiency (EP) programmes have long been established in English-medium universities, serving international students who have met the requirements of academic backgrounds, but not yet the language proficiency for admission (Crosthwaite, 2016; Keefe and

This research is supported by the Ministry of Education, Singapore, under its Tertiary Education Research Fund (MOE2019-TRF-020). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore.

Higher Education Evaluation and Development
Vol. 15 No. 2, 2021
pp. 82-97
Emerald Publishing Limited
2514-5789
DOI 10.1108/HEED-01-2021-0002

Shi, 2017; Storch and Tapper, 2009). EP programmes usually provide overseas students with a pathway into university studies and serve as proof of their English language proficiency.

In the context of Singapore, English has been the first language and sole medium of instruction in education at all levels, although the majority of Singaporeans are bilingual or multilingual speakers who could have a different mother tongue to use at home, such as Chinese Mandarin, Chinese dialects, Malay and Tamil. (Bolton *et al.*, 2017). Hence, for admission into undergraduate programmes in Singapore universities, all domestic students are also required to satisfy the EP requirement, such as attaining at least a Grade B4 in the General Certificate of Education Ordinary Level English or its equivalent (i.e. Grade 4 in English for International Baccalaureate (IB) diploma holders; International English Language Testing System (IELTS) academic score of 6.5; Test of English as a Foreign Language (TOEFL) score of 580 (paper-based) or 237 (computer-based) or 85 (Internet-based)).

Without meeting the EP requirement on admission, students would be provided with a pre-entry pathway programme in order to fulfil the language requirement. An example is the EP programme offered at the case university in Singapore, which is the focus of the present study. This EP programme includes a diagnostic course to indicate students' EP level on admission, followed by various enhancement courses to be completed if students fail the diagnostic course.

Review of the efficacy of EP programmes

To better understand the impact of EP programmes provided in higher education, previous studies have primarily focused on students' self-reported experiences and have yielded contradictory findings. On one hand, positive impacts were reported in a study under an Australian context (Baik and Greig, 2009) that students perceived usefulness in taking the EP programme, and those who attended the EP programme more frequently seemed to perform better in their subject course compared to those who rarely attended. Likewise, a subsequent study in Singapore (Xudong *et al.*, 2010) reported similar findings that students perceived the overall utility of the EP course positively as well as an increase of self-confidence after the course, despite the small to moderate progress in their writing quality when their writing products before and after the programme were compared. On the other hand, some recent studies have reported the perceived dissatisfaction with the effectiveness of the EP programme in Hong Kong universities (Bruce and Hamp-Lyons, 2015) and Iranian universities (Abdolerezapour and Tavakoli, 2013).

The differences in previous findings may be attributed to the measures meted out, e.g. questionnaires; these measures could be subjective in nature, and other confounding variables could also contribute to students' improvement in writing quality in addition to the intervention of an EP programme. Further, among the limited prior research attempting to evaluate the impact of EP programmes and courses, most studies focused on students' perceptions and writing products; little research has been done to explore the impact of EP courses on students' overall academic performance in the university such as their semester grade point average (GPA).

EP as a moderator of SGPA and why evaluate EP

English proficiency has been deemed as a critical condition for students' academic success in English-medium universities, so students who speak English as an additional language are usually required to fulfil certain EP conditions for admission into university studies, such as via IELTS or TOEFL. In addition, universities have also been accepting alternative forms of EP evidence such as locally developed language tests (Cho and Bridgeman, 2012; Oliver *et al.*, 2012), which has been described to be beneficial due to their alignment with the local curriculum (Coombe and O'Sullivan, 2012). It was argued that the results of locally developed tests could be more representative of students' true competence levels than those of internationally standardised EP test, considering that students might have intentionally prepared for the internationally standardised tests and hence, could perform above their actual proficiency levels (Ockey and Gokturk, 2019). Moreover, results of locally developed EP tests could serve as direct references about how to assign different students into relevant levels of EP courses in the university after being admitted.

To understand the role of students' EP levels in their academic performance in a university, various studies examining different EP tests have been conducted to understand their predictive validity in students' GPA during the first semester or year in the university. Predictive validity is described as "the correlation between test scores and later performance on something" (Carr, 2011). In terms of the EP tests, predictive validity would refer to the correlation between EP test scores and test takers' subsequent performance in the academic studies after admission into universities, such as GPA.

Findings for internationally standardised tests have been inconsistent. For example, Kerstjen and Nery (2000) discovered weak correlations between students' IELTS scores and GPA in an Australian university. Similarly, in the study of Cho and Bridgeman (2012), who examined the predictive validity of TOEFL in students' first-year GPA in American universities, results indicated that TOEFL scores only accounted for 3% of the variance in GPA among a sample of more than 2000 students. Those weak or even no correlations were also identified in other studies exploring the predictive validity of TOEFL/IELTS (Arrigoni and Clark, 2015; Bridgeman *et al.*, 2016; Dooey and Oliver, 2002). On the contrary, some studies identified a relatively stronger or more evident correlation. For instance, Woodrow (2006) found a moderate correlation between IELTS score and GPA in an Australia university; this was supported by Wait and Gressel (2009), who discovered that students' TOEFL score was moderately correlated with their GPA in an American university. This moderate correlation was even more evident when the first semester's GPA was used to indicate students' academic performance in the university (Yen and Kuzma, 2009).

Prior research exploring the relationship between the locally developed EP tests and students' university GPA also yielded inconclusive findings. Lee and Greene (2007) did not find noticeable correlations between the score of institutionally designed EP test and GPA, indicating that the internally designed EP test was a weak predictor of students' academic performance. However, a later study seemed to identify the significant and positive correlations between students' EP levels on entry into the university and their subsequent academic performance reflected by GPA (Ghenghesh, 2015).

Despite the disparity in research findings, what has been emphasised in the relevant literature is that English as a foreign language (EFL) status students were found to obtain lower GPAs in general than their non-EFL counterparts, so EP test scores have been deemed to have the capacity to ensure students' readiness to perform in their academic studies at English-medium universities (Eddey and Baumann, 2009; Neumann *et al.*, 2019).

This presents the importance of evaluating the EP programme and its efficacy with respect to students' academic performance in the case university, particularly when the EP programme has been in effect for some years. Lending support to this study is that most of the previous studies in this research area focused on international students who speak English as an additional language, and domestic students who speak English as the first language are under-researched, especially in a bilingual or multilingual context such as Singapore.

Methodology

A review of the relevant literature found that most studies related to evaluating EP have primarily focussed on student self-reports. Some of these may be faced with criticisms associated with biasness and these could partly be addressed by utilising complementary experimental or quasi-experimental approaches. While randomised controlled trials (RCTs) present the gold standard for measuring the efficacy of programmes, treatments or interventions (Hariton and Locascio, 2018), practical issues remain. In educational research, ethics and costs related issues dominate the discourse when selecting evaluation and research methods. For example, denying a group of students an intervention or remediation programme for the purposes of a research study might result in this group of students having a less desirable result; this calls into question the issue of ethics and fairness. Further, a retrospective study similar to this study would not be possible with RCTs.

In light of the issues faced by RCTs, regression discontinuity (RD) design, a quasi-experimental approach, offers a plausible if not excellent near-equivalent alternative given that it overcomes these issues including ethical concerns (Imbens, 2008; Shadish *et al.*, 2002; Smith, 2014). RD has gained widespread interest in recent years and is considered a rigorous approach to estimate programme impacts in situations where individuals are selected for treatment based on a numeric threshold (Jacob *et al.*, 2012). Nakamoto *et al.* (2017) have also suggested that RD affords stronger causal inferences than any other designs except for RCTs.

In RD, a cut-point or threshold, also commonly known as the forcing or assignment variable is set so that participants are split into two groups: one group will receive a treatment or intervention, while the other does not and serves as a control group. The situation where the entire group below the threshold receives treatment and the other group does not is known as a sharp RD design. A fuzzy RD design occurs when some participants from both groups undergo treatment or elect not to undergo treatment despite a cut-point. Particularly for sharp RD designs, participants at, just above and below the threshold, are considered systematically similar and hence, an unbiased estimate or effect of the intervention programme or treatment can be obtained from the difference in scores at the threshold on the premise that the functional form is specified appropriately. (Jacob *et al.*, 2012; Nakamoto *et al.*, 2017; Shadish *et al.*, 2002; Smith, 2014).

An alternative to RD is propensity score matching (PSM) that mimics some characteristics of RCTs (Austin, 2011). PSM functions on the basis that a propensity score is calculated for an individual based on the conditional probability of him/her seeking treatment or an intervention, and this propensity score is matched with another individual's with similar observed covariates to this individual before further tests, e.g. *t*-tests, are performed to evaluate the intervention (Adelson, 2013; Rosenbaum and Rubin, 1983). Nonetheless, PSM was less applicable to this study, as students below the cut-point or threshold were not given options other than offering the EP programme. Under this condition, RD design was chosen for this study.

In this study, the cut-point was pre-determined by policy as grade four on an ordinal scale. While Jacob *et al.* (2012) suggested that a cut-point would be from a continuous variable, Linden *et al.* (2006) and Nakamoto *et al.* (2017) suggested that a cut-point from either a continuous or ordinal variable would be tenable. Though grade four in this study should, by strict standards, be considered from an ordinal scale, these grade numbers have also been commonly used at the national level as an interval scale, i.e. adding the first language and top four or five subject grades provides an overall grade-score for placement purposes with a lower overall grade-score as the better. In this regard, the grade numbers could also be considered pseudo-continuous such that grade one is the best along the continuum from one to nine.

Participants and data

Accessible cross sectional data of students matriculated from 2017 to 2018 was used in this study. Non-accessible data included students who were exempted from the EP programme owing to the degree programme they read, e.g. Tamil language and literature, and students who met the EP requirement by means other than the General Certificate of Education Ordinary Level English, i.e. EL_GR, jointly awarded by the Singapore Ministry of Education and Cambridge Assessment. Across both years, 11 students who met the EP requirement and still offered the EP programme were not considered as they were suspended, terminated or withdrew from the case university by the end of the first semester of study; this enabled a sharp RD design for the study. Further, the cross sectional nature of the data afforded strong causal inferences for this study in that any differences in the outcome variable, i.e. SGPA, would be due to the intervention or random fluctuations, as the cut-point is known and applied; other factors could possibly cause effects other than the cut-point though this is unlikely (Luyten, 2006).

The various parallel tests in the EP programme have been judged as comparable by subject matter experts, and given the rigour of the programme and the highly limited evidence to suggest that students who matriculated in January would be advantaged or disadvantaged relative to students who matriculated in July, students matriculated within the same year were considered as a unit for analysis.

Analysis

The study was conducted based on recommendations by Jacob *et al.* (2012) and Smith (2014). SAS Version 9.4 was used for all analyses in addition to the sharp regression discontinuity design macro by Schoeneberger (2011).

Step 1 – Applicability of RD. The initial step for this research was to ascertain whether RD was an appropriate method to study the efficacy of the EP programme in the case university. Prior discussions about the participants indicated a sharp RD design in that those who did not meet the EP requirement, i.e. cut-point of EL_GR grade four, had to undergo the EP intervention while the others served as a control group. As the cut-point was set prior to the EP intervention and identifies those who need EP assistance, it is appropriate as an assignment variable. Further, there was no way the EP programme could have affected how students were assigned (Shadish *et al.*, 2002).

Suggestions of possible violations towards the applicability of RD by Bloom (2012), Jacob *et al.* (2012) and Shadish *et al.* (2002) due to crossover (always-takers or defiers) or absence (never-takers) including intervention drop out were either not observed or corrected by: (1) not considering the small number of students who were supposed to offer the EP programme but chose not to do it in the semester immediately after matriculation and (2) not considering the 11 crossover cases as the statuses of these students were either terminated, suspended or withdrawn by the end of the semester they offered EP (which was also their first semester of study).

Prior studies within the literature have demonstrated that English correlates with or has a predictive effect on GPA. This supports the appropriateness of applying of RD in this instance even though the outcome variable, i.e. SGPA, may not be required to be related to the assignment variable (Shadish *et al.*, 2002). The SGPA, being a continuous variable, further supports the use of RD in this study (Linden *et al.*, 2006).

Step 2 – Data suitability. Smith (2014) suggested two plots for this step to observe any potential threats to the validity of using RD with a set of data: (1) scatter plot of the raw data and (2) distribution of the assignment variable by frequency. Figures 1–6 present the scatter plot and distribution of the data of 2017 and 2018; the black solid line in Figures 3 and 4 represents the cut-point of EL_GR grade four.

Based on the figures, there was no evidence to suggest that there could be any way of manipulating the assignment variable, despite the cut-point being made known in the public domain (see Figures 1 and 2); Figures 1 and 2 do not indicate any stark contrast in frequencies

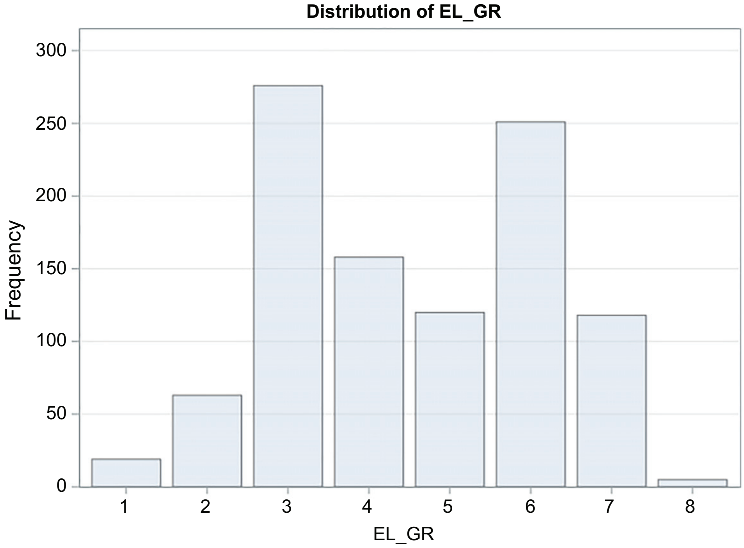

Figure 1.
Distribution of EL_GR (2017)

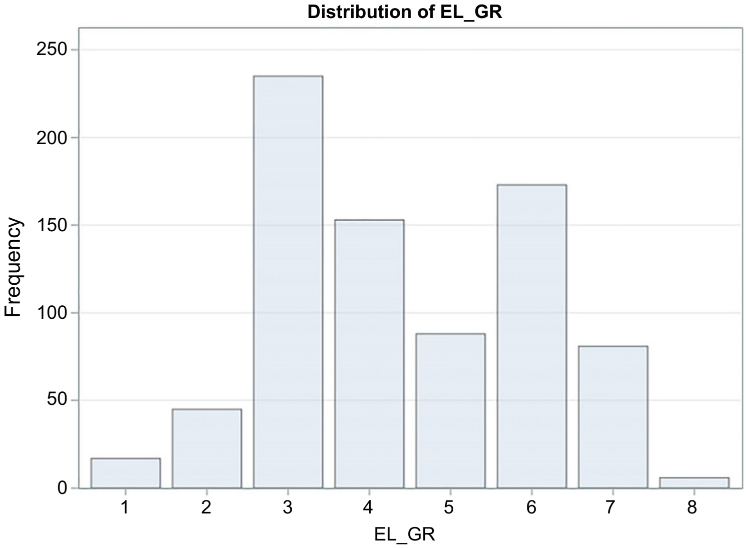

Figure 2.
Distribution of EL_GR (2018)

before or after EL_GR grade four and present both 2017 and 2018 cohorts as having similar trends. There was also no evidence to suggest that there were existing discontinuities that could be attributed to factors other than the cut-point for 2017 and 2018 (see Figures 3 and 4); it is visually clear for both years that a discontinuity exists at EL_GR grade four though the modest disjoints between grades one and two, and grades seven and eight were likely due to the substantially fewer observations.

The plot of mean and median SGPA against EL_GR for each year is presented in Figures 5 and 6, as any discontinuities occurring naturally is less obvious given the ordinal nature of

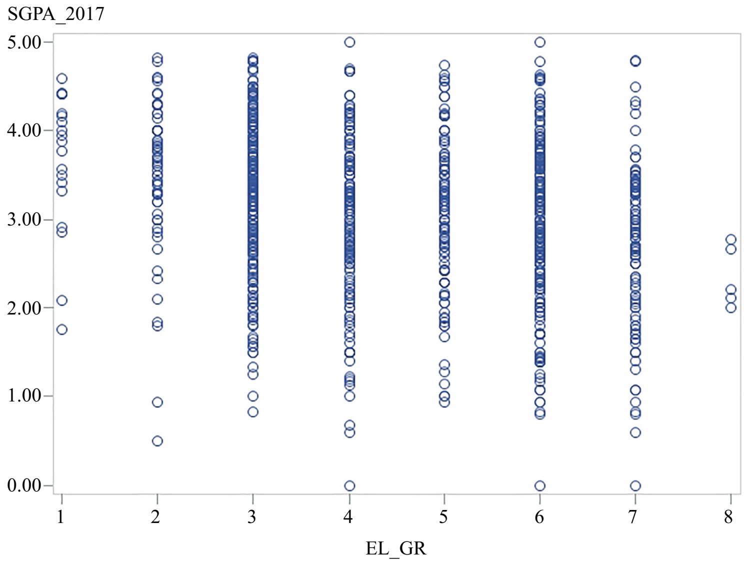

Figure 3.
Scatter plot of SGPA against EL_GR (2017)

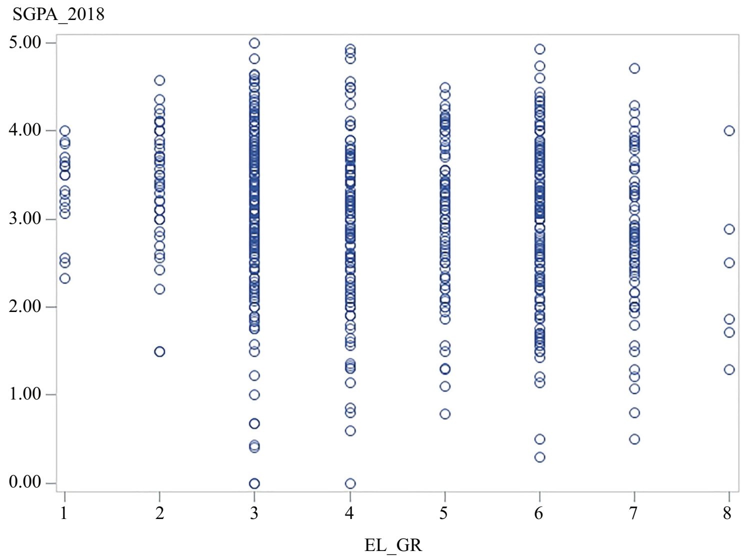

Figure 4.
Scatter plot of SGPA against EL_GR (2018)

the EL_GR that resulted in vertical stacks of data points; median plots were included to account for outliers. Figures 5 and 6 confirm that a clear discontinuity exists at EL_GR grade four for both 2017 and 2018.

As Figures 1–6 do not present evidence for any threat to validity, the data were deemed suitable for the application of RD.

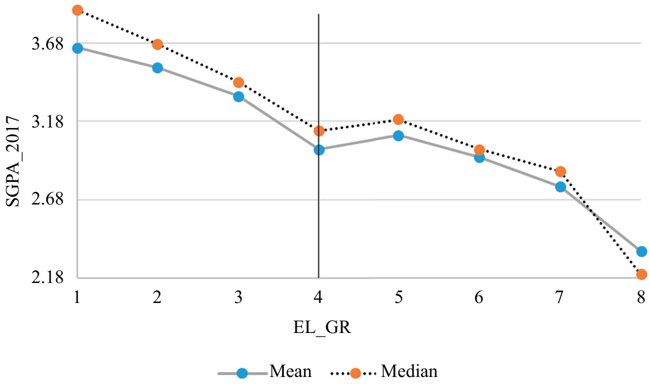

Figure 5.
Mean and median SGPA against EL_GR (2017)

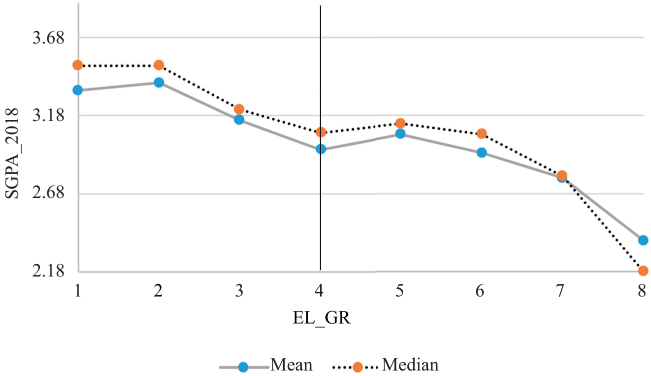

Figure 6.
Mean and median SGPA against EL_GR (2018)

Step 3 – Precision, functional form and bandwidth for RD. As with most recommendations, Shadish *et al.* (2002) suggested a larger sample to minimise discrepancies and errors and increase the chance of randomisation. Smith (2014) also discussed the literature related to having a larger sample size in the context of using RD as an alternative to RCTs. For this study, all accessible student data was used and hence, the sample was in effect the population. Further, the cut-point of grade four also resulted in 51.1 and 56.4% of students, forming the control group for 2017 and 2018, respectively. This enabled about half the number of students on each side of the cut-point to model the regression, sufficient for accuracy (Smith, 2014).

In determining the most suitable functional form for RD, the method of ordinary least squares (OLS) was used to fit the general linear, quadratic and cubic model, representing SGPA as the response variable and EL_GR as the explanatory variable. Based on the OLS method, quadratic and cubic terms of EL_GR were insignificant and could be removed from the model. This resulted in retaining only the linear term ($p < 0.0001$) (see Eqn (1)). Figures 7–10 show the insignificance of the quadratic and cubic terms, as the fit is almost linear despite overlaying the quadratic or cubic fit.

$$SGPA_i = \beta_0 + EP\beta_1 + \varepsilon_i \tag{1}$$

$SGPA_i$ represents the initial semester GPA of student *i*. The initial semester GPA results were used: (1) as an attempt to minimise confounds that increase with the number of semesters of

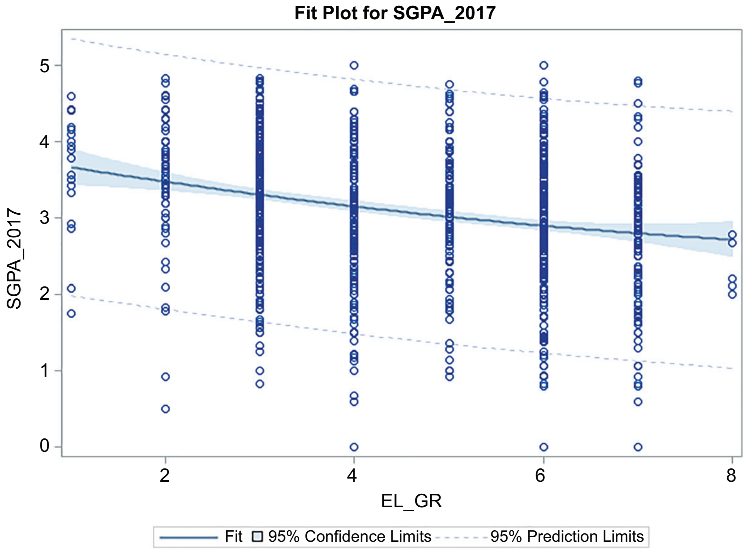

Figure 7.
Quadratic model for SGPA against EL_GR (2017)

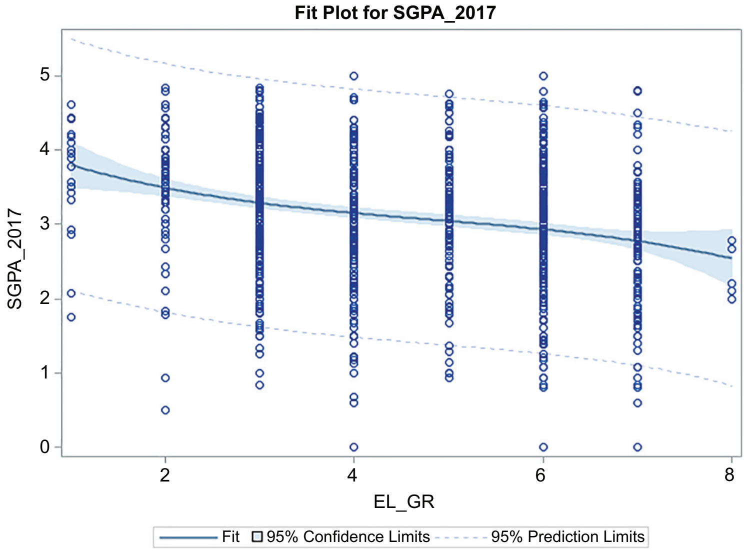

Figure 8.
Cubic model for SGPA against EL_GR (2017)

study and (2) as the case university strongly encourages students who just matriculated to offer the EP programme if they do not meet the threshold; most students complete the EP programme within the first semester of study. β_0 represents the intercept and EP_i is the intervention for student i. β_1 is the coefficient of the intervention and ε_i is the random error for student i.

Smith (2014) stated that while RD designs could be done with a full sample, selecting a sample around the cut-point i.e. using a smaller bandwidth might be less susceptible to effects

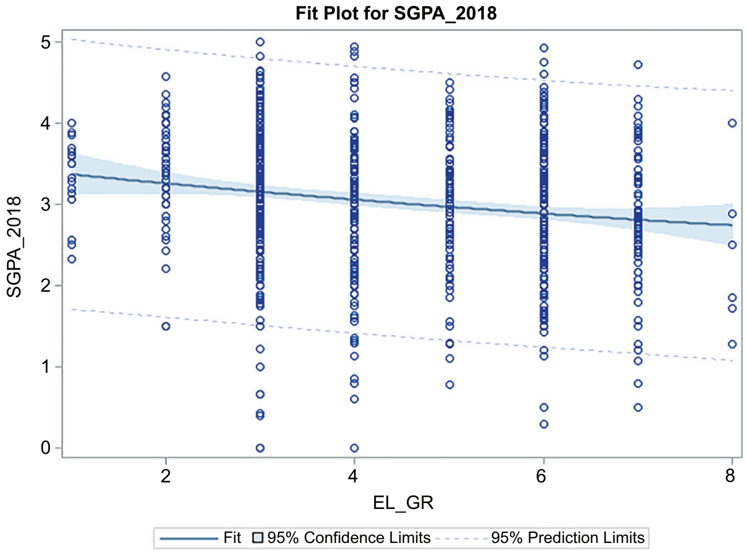

Figure 9. Quadratic model for SGPA against EL_GR (2018)

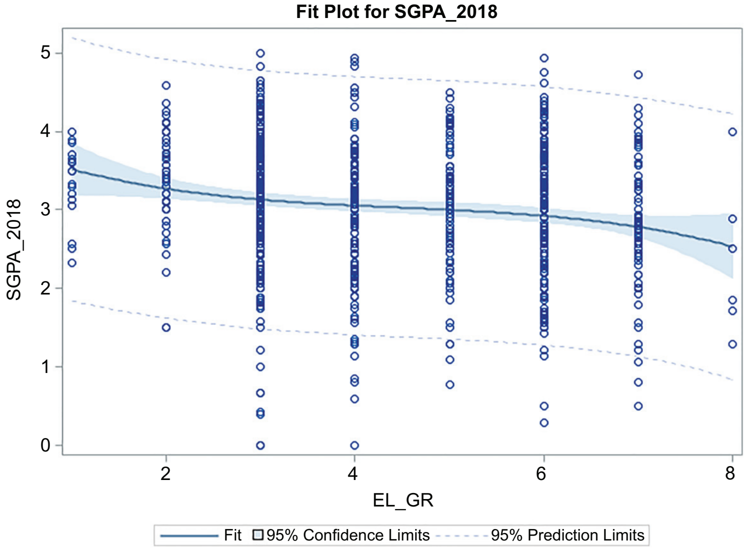

Figure 10. Cubic model for SGPA against EL_GR (2018)

brought upon by outliers. Nonetheless, given that EL_GR spans only eight grades, with four being the cut-point, it was decided that all data points would be considered for this study.

Step 4 – Applying RD. Given the suitability of data for RD and the linear functional form of the regression model, the sharp regression discontinuity design SAS macro by Schoeneberger (2011) could be used for this study. The RD analysis found that the intervention effect for both 2017 and 2018 was significant and positive (see Table 1, Figures

11 and 12). The intervention effect on the SGPA ranged between an average of 0.27–0.29 across both years.

Step 5 – Sensitivity tests. Sensitivity tests were conducted by shifting the cut-point to create pseudo discontinuities were conducted based on recommendations by Imbens and Lemieux (2008). Using EL_GR grade three and five as a cut-point and pseudo-assigning students to the EP programme resulted in non-significant intervention effects (see Table 2). These findings do not suggest any threats to the validity of the EP intervention effect and affirmed that the EP intervention effect was present only at EL_GR grade four, the publicly known cut-point.

Discussion and directions for future research

Results from the RD analyses suggest that the current EP programme in the case university within a multilingual context has a significant and positive intervention effect on the SGPA for students who did not meet the EL_GR grade four threshold. This supports the finding from reviewing the relevant literature that EP has the capacity to ensure students' readiness to perform in their academic studies.

Effect size estimates were calculated by comparing students who scored EL_GR three with those who scored EL_GR five, as these grades are adjacent to the cut-point for both years. For 2017, the effect size ($d = 0.25$) was considered small based on Cohen's (1988) widely used guidelines. This finding is similar to 2018 ($d = 0.11$). Nonetheless, empirical measures of effect size, e.g. Cohen's, may not apply owing to varying contexts (Bloom *et al.*, 2008; Sun *et al.*, 2010). In fact, in his seminal and contemporary work, Kraft (2020) posited that within educational research, effect sizes of 0.20 may be considered large and, to

Table 1.
Effect of EP intervention on SGPA

Year	Effect of EP intervention	Standard error	95% confidence interval	*p*-value
2017	0.29	0.12	0.06–0.52	0.01
2018	0.27	0.13	0.02–0.52	0.03

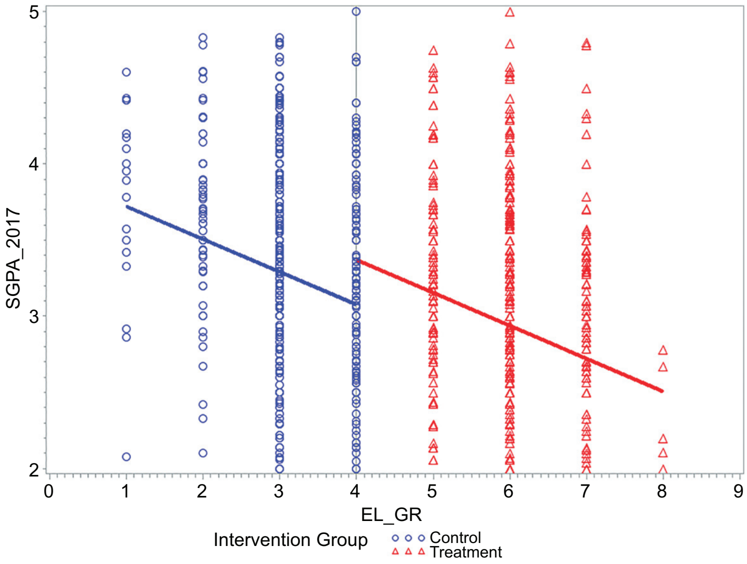

Figure 11.
RD analysis for 2017

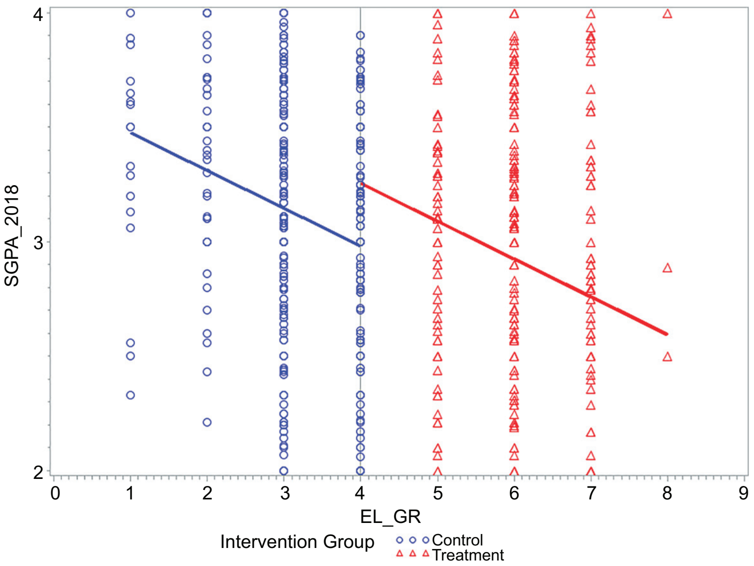

Figure 12.
RD analysis for 2018

Table 2.
RD analysis using pseudo cut-points

Year	EL_GR grade	Effect of EP intervention	Standard error	95% confidence interval	*p*-value
2017	3	−0.17	0.10	−0.37–0.02	0.07
2017	5	0.15	0.10	−0.06–0.35	0.15
2018	3	−0.08	0.10	−0.28–0.12	0.45
2018	5	0.08	0.11	−0.14–0.31	0.47

this end, the effect size across both years can be considered medium to large. Such effect sizes support the EP programme as an enabler and leveller such that students with lower levels of EP as reflected by the EL_GR grade would not be disadvantaged significantly on the premise that they offered the EP programme in the case university within the initial semester. Hence, while there may be significant and positive correlations between students' EP levels on entry into university and their subsequent academic performance, the difference in SGPA might be non-significant owing to improved levels of EP.

The analyses found that using EL_GR grade five as the cut-point presented a positive but non-significant effect of the EP programme, and using EL_GR grade three as the cut-point with the same dataset presented a regression of the SGPA results. These supported using EL_GR grade four as a threshold. Clearly, identifying students who need help in EP based on the EL_GR grade four threshold offers the greatest yield in that the RD analyses presented a positive and significant effect as compared with using EL_GR grade five or three as the threshold. Coincidentally, EL_GR grade four also identifies approximately 50% of the students from the dataset across both years for the EP programme intervention; this provided an acceptable number of observations for both the treatment and control group.

Similar to other RD analyses, the intervention effect should strictly be applicable to students centred around the threshold, EL_GR grade four in this case. Nonetheless, Jacob *et al.* (2012) discussed a more expansive view on the generalisability of RD findings. Owing to random error, the heterogeneous nature of the student population at the cut-point given that cross sectional data was used, and that RD design mimics RCTs, findings in this study could be extended to students beyond the threshold. Nonetheless, this generalisation should preferably be limited to the sample used in the analyses (Smith, 2014). This lends support to applying RD design consistently over the next few years to ascertain how the EP programme works across years.

The RD design applied in this study used a numeric ordinal cut-point owing to data accessibility issues. While an ordinal variable is accepted for RD designs, some including Jacob *et al.* (2012) have suggested that a continuous variable might be ideal. This is reflected in the vertical stacks of scatter plots for this study and plots of mean and median were included to ascertain any naturally occurring discontinuities. Subsequent studies could be expanded to include another continuous variable as the cut-point or other explanatory variables in the regression model, though this poses a challenge, as standardised national exam scores are neither made public nor readily available.

It should be noted that findings from this study apply for the current form of EP programme. Subsequent curricular or structural changes of the EP programme warrant further analyses, not necessarily RD, to determine its efficacy.

Conclusion and practical implications

This study served to provide information on the efficacy of EP on initial semestral results in a case university situated within a multilingual context (i.e. Singapore). The study presented results expected of the EP programme owing to the amount of invested resources. It also adds to the literature by showing, through a quasi-experimental approach (i.e. regression discontinuity design), an EP programme's capacity to positively contribute to the academic success of students in English-medium universities (Eddey and Baumann, 2009; Neumann *et al.*, 2019), particularly among the population of bilingual/multilingual students who speak English as their first language.

Results of this study demonstrate that the implementation of EP programmes in higher education institutions is essential not only for international students who are foreign language speakers of English but also for domestic students in English-speaking countries, especially for bi/multilingual speakers. This is crucial to the attention of policy makers and administrators at universities when they make informed decisions about the EP curriculum and corresponding assessment practices, as this would ensure that students attain an EP level sufficient to academically succeed in their subsequent studies.

Further, the significant efficacy of the EP programme in relation to students' SGPA in the current study echoed prior research that locally designed EP programmes and tests seemed to be advantageous due to their alignment with the institutional curriculum, and can be potentially more representative of students' true competence than internationally standardised EP tests that might be less representative due to coaching and test preparation effects (Coombe and O'Sullivan, 2012; Ockey and Gokturk, 2019). In the context of Singapore, in particular, it is of great importance for higher education institutions to develop and evaluate their EP programme for both admission and placement decisions, which would better cater to the bi/multilingual speakers speaking English as their first language.

References

Abdolerezapour, P. and Tavakoli, M. (2013), "University teachers and students perceptions of EAP methodologies and their effectiveness", *The Social Sciences*, Vol. 8 No. 1, pp. 49-54, doi: 10.36478/sscience.2013.49.54.

Adelson, J.L. (2013), “Educational research with real-world data: reducing selection bias with propensity scores”, *Practical Assessment, Research and Evaluation*, Vol. 18 No. 15, pp. 1-11, doi: 10.7275/4nr3-nk33.

Arrigoni, E. and 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*, Vol. 29, available at: https://www.ielts.org/-/media/research-reports/ielts_online_rr_2015-3.ashx.

Austin, P.C. (2011), “An introduction to propensity score methods for reducing the effects of confounding in observational studies”, *Multivariate Behavioral Research*, Vol. 46 No. 3, pp. 399-424, doi: 10.1080/00273171.2011.568786.

Baik, C. and Greig, J. (2009), “Improving the academic outcomes of undergraduate ESL students: the case for discipline-based academic skills programmes”, *Higher Education Research and Development*, Vol. 28 No. 4, pp. 401-416, doi: 10.1080/07294360903067005.

Bloom, H.S. (2012), “Modern regression discontinuity analysis”, *Methodological Studies*, Vol. 5 No. 1, pp. 43-82, available at: https://www.mdrc.org/sites/default/files/Regression_Discontinuity_embed.pdf.

Bloom, H.S., Hill, C.J., Black, A.R. and Lipsey, M.W. (2008), “Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions”, *Journal of Research on Educational Effectiveness*, Vol. 1 No. 4, pp. 289-328, doi: 10.1080/19345740802400072.

Bolton, K., Botha, W. and Bacon-Shone, J. (2017), “English-medium instruction in Singapore higher education: policy, realities and challenges”, *Journal of Multilingual and Multicultural Development*, Vol. 38 No. 10, pp. 913-930, doi: 10.1080/01434632.2017.1304396.

Bridgeman, B., Cho, Y. and DiPietro, S. (2016), “Predicting grades from an English language assessment: the importance of peeling the onion”, *Language Testing*, Vol. 33 No. 3, pp. 307-318, doi: 10.1177/0265532215583066.

Bruce, E. and Hamp-Lyons, L. (2015), “Opposing tensions of local and international standards for EAP writing programmes: who are we assessing for?”, *Journal of English for Academic Purposes*, Vol. 18, pp. 64-77, doi: 10.1016/j.jeap.2015.03.003.

Carr, N.T. (2011), *Designing and Analyzing Language Tests*, Oxford University Press, Oxford.

Cho, Y. and Bridgeman, B. (2012), “Relationship of TOEFL iBT® scores to academic performance: some evidence from American universities”, *Language Testing*, Vol. 29 No. 3, pp. 421-442, doi: 10.1177/0265532211430368.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Erlbaum, Hillsdale, NJ.

Coombe, C. and O’Sullivan, B. (2012), *The Cambridge Guide to Second Language Assessment*, Cambridge University Press, Cambridge.

Crosthwaite, P. (2016), “A longitudinal multidimensional analysis of EAP writing: determining EAP course effectiveness”, *Journal of English for Academic Purposes*, Vol. 22, pp. 166-178, doi: 10.1016/j.jeap.2016.04.005.

Dooey, P. and Oliver, R. (2002), “An investigation into the predictive validity of the IELTS Test as an indicator of future academic success”, *Prospects*, Vol. 17 No. 1, pp. 36-54, available at: http://www.ameprc.mq.edu.au/__data/assets/pdf_file/0017/230075/17_1_3_Dooey.pdf.

Eddey, P. and Baumann, C. (2009), “Graduate business education: profiling successful students and its relevance for marketing and recruitment policy”, *Journal of Education for Business*, Vol. 84 No. 3, pp. 160-168, doi: 10.3200/JOEB.84.3.160-168.

Ghenghesh, P. (2015), “The relationship between English language proficiency and academic performance of university students – should academic institutions really be concerned?”, *International Journal of Applied Linguistics and English Literature*, Vol. 4 No. 2, pp. 91-97, doi: 10.7575/aiac.ijalel.v.4n.2p.91.

Hariton, E. and Locascio, J.J. (2018), “Randomised controlled trials - the gold standard for effectiveness research: study design: randomised controlled trials”, *BJOG: An International*

Journal of Obstetrics and Gynaecology, Vol. 125 No. 13, p. 1716, doi: 10.1111/1471-0528.15199.

Imbens, G. (2008), "Special issue editors' introduction: the regression discontinuity design-theory and applications", *Journal of Economics*, Vol. 142 No. 2, pp. 611-614, doi: 10.1016/j.jeconom.2007.05.008.

Imbens, G.W. and Lemieux, T. (2008), "Regression discontinuity designs: a guide to practice", *Journal of Econometrics*, Vol. 142, pp. 615-635, doi: 10.1016/j.jeconom.2007.05.001.

Jacob, R., Zhu, P., Somers, M.A. and Bloom, H. (2012), *A Practical Guide to Regression Discontinuity*, MDRC, available at: http://www.mdrc.org/sites/default/files/regression_discontinuity_full.pdf.

Keefe, K. and Shi, L. (2017), "An EAP programme and students' success at a Canadian university", *TESL Canada Journal*, Vol. 34 No. 2, pp. 1-24, available at: http://files.eric.ed.gov/fulltext/EJ1170742.pdf.

Kerstjens, M. and Nery, C. (2000), "Predictive validity in the IELTS test: a study of the relationship between IELTS scores and students' subsequent academic performance", *International English Language Testing System (IELTS) Research Reports*, Vol. 3, pp. 85-108, 2000, doi: 10.1075/itl.19021.pea.

Kraft, M.A. (2020), "Interpreting effect sizes of education interventions", *Educational Researcher*, Vol. 49 No. 4, pp. 241-253, doi: 10.3102/0013189X20912798.

Lee, Y.-J. and Greene, J. (2007), "The predictive validity of an ESL placement test: a mixed methods approach", *Journal of Mixed Methods Research*, Vol. 1 No. 4, pp. 366-389, doi: 10.1177/1558689807306148.

Linden, A., Adams, J. and Roberts, N. (2006), "Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design", *Journal of Evaluation in Clinical Practice*, Vol. 12 No. 2, pp. 124-131, doi: 10.1111/j.1365-2753.2005.00573.x.

Luyten, H. (2006), "An empirical assessment of the absolute effect of schooling: regression discontinuity applied to TIMSS-95", *Oxford Review of Education*, Vol. 32 No. 3, pp. 397-429, available at: https://www.jstor.org/stable/4618668.

Nakamoto, J., Wendt, S., Rice, J., Bojorquez, J. and Petrosino, A. (2017), "An evaluation of school improvement grants using regression discontinuity and quasi-experimental designs", *SAGE Research Methods Cases Part 2*. doi: 10.4135/9781473953987.

Neumann, H., Padden, N. and McDonough, K. (2019), "Beyond English language proficiency scores: understanding the academic performance of international undergraduate students during the first year of study", *Higher Education Research and Development*, Vol. 38 No. 2, pp. 324-338, doi: 10.1080/07294360.2018.1522621.

Ockey, G.J. and Gokturk, N. (2019), "Standardized language proficiency tests in higher education", in Gao, X. (Ed.), *Second Handbook of English Language Teaching*, Springer, Cham, pp. 1-17.

Oliver, R., Vanderford, S. and Grote, E. (2012), "Evidence of English language proficiency and academic achievement of non-English-speaking background students", *Higher Education Research and Development*, Vol. 31 No. 4, pp. 541-555, doi: 10.1080/07294360.2011.653958.

Rosenbaum, P.R. and Rubin, D.B. (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika*, Vol. 70, pp. 41-55, doi: 10.1093/biomet/70.1.41.

Schoeneberger, J.A. (2011), "RDPLOT: a SAS® macro for generating regression-discontinuity plots", *Proceedings of the annual Southeastern SAS Users Group Conference*.

Shadish, W., Cook, T. and Campbell, D. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin Company, Boston, MA.

Smith, W. (2014), "Estimating unbiased treatment effects in education using a regression discontinuity design", *Practical Assessment, Research and Evaluation*, Vol. 19 No. 9, pp. 1-9, doi: 10.7275/7911-vd52.

Storch, N. and Tapper, J. (2009), "The impact of an EAP course on postgraduate writing", *Journal of English for Academic Purposes*, Vol. 8 No. 3, pp. 207-223, doi: 10.1016/j.jeap.2009.03.001.

Sun, S., Pan, W. and Wang, L.L. (2010), "A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology", *Journal of Educational Psychology*, Vol. 102 No. 4, pp. 989-1004, doi: 10.1177/0959354312436870.

Wait, I.W. and Gressel, J.W. (2009), "Relationship between TOEFL score and academic success for international engineering students", *Journal of Engineering Education*, Vol. 98 No. 4, pp. 389-398, doi: 10.1002/j.2168-9830.2009.tb01035.x.

Woodrow, L. (2006), "Academic success of international postgraduate education students and the role of English proficiency", *University of Sydney papers in TESOL*, Vol. 1 No. 1, pp. 51-70, available at: https://faculty.edfac.usyd.edu.au/projects/usp_in_tesol/pdf/volume01/article03.pdf.

Xudong, D., Cheng, L.K., Varaprasad, C. and Leng, L.M. (2010), "Academic writing development of ESL/EFL graduate students in NUS", *Reflections on English Language Teaching*, Vol. 9 No. 2, pp. 119-138, available at: http://www.nus.edu.sg/celc/research/books/relt/vol9/no2/119to138_deng.pdf.

Yen, D. and Kuzma, J. (2009), "Higher IELTS score, higher academic performance? The validity of IELTS in predicting the academic performance of Chinese students", *Worcester Journal of Learning and Teaching*, No. 3, available at: http://eprints.worc.ac.uk/811/1/YenKuzmaIELTScores.pdf.

About the authors

Lyndon Lim, EdD (ORCID 0000–0002–8199–5761), is Senior Lecturer with the Singapore University of Social Sciences. His research and publications focus on assessment and evaluation, and the social psychology of education.

Wenjin Vikki Bo, PhD, is a Senior Lecturer with the Singapore University of Social Sciences. Her teaching and research interests are in Applied Linguistics, Educational Psychology and Technology-enhanced Learning. In particular, she has a research interest to explore the predictors of students' academic performance in higher education. Wenjin Vikki Bo is the corresponding author and can be contacted at: wenjinbo@connect.hku.hk