EVALUATING LOCALLY-DEVELOPED LANGUAGE TESTING

A PREDICTIVE STUDY OF 'DIRECT ENTRY' LANGUAGE PROGRAMS AT AN AUSTRALIAN UNIVERSITY

Nicholas Cope *Centre for Macquarie English, Macquarie University*
nicholas.cope@mq.edu.au

The study reported here investigates the predictive validity of language assessments by 'Direct Entry' programs at an Australian University – programs developed on site for Non English Speaking Background international students, principally to provide (i) pre-entry academic and language preparation and (ii) language assessment for university admissions purposes. All 138 students in the sample had entered degree studies via one of the three programs that made up the locally-developed Direct Entry pathway. Inferential statistics (correlation and regression) showed the assessments awarded by two programs to satisfactorily predict academic outcomes, while predictive validity for one was not demonstrated. Descriptive statistics (mean pass rates and academic averages) then revealed a pattern of relatively poor academic performance in certain university disciplines to which particular Direct Entry programs were dedicated. Informed by principles of language program evaluation, the study's outcomes were seen as both summative and formative: remedial strategies are accordingly recommended. While the specific relevance of the study's findings is to the particular institutional context in which the study was conducted, the study instantiates a perspective on language assessment validation of broader relevance in an Australian context where locally-developed Direct Entry programs – about which the research literature is largely silent – are increasingly widespread.

KEY WORDS: English for academic purposes, language tests, test validity and reliability, academic achievement, higher education, Australia

1. BACKGROUND

Assessment results that play a role in decision-making processes for university admissions, and so in part serve a gate-keeping function, necessarily carry high stakes. University admissions decisions have an impact on a number of stakeholders in tertiary and higher education, most obviously prospective students, and also a variety of others both in the receiving institution and beyond. Given their high-stakes nature, the validity of the

ARAL 34:1 (2011), 40-59 DOI 10.1075/aral.34.1.03cop
ISSN 0155–0640 / E-ISSN 1833–7139

assessments that contribute to admissions decisions is frequently subject to the scrutiny of research. The study reported here is an example.

Language tests used by English medium universities in the admissions process for non English speaking background (NESB) students have frequently been evaluated – more so in recent years, concurrent with trends towards internationalisation in higher education – by predictive validity studies, which assess the extent to which test results predict subsequent academic outcomes. The USA-based Test of English as a Foreign Language (TOEFL) and the more recent UK-based International English Language Testing System (IELTS), both widely recognised tests of English language ability, in particular have attracted considerable research attention (see Hale, Stansfield & Duran, 1984 and Graham, 1987, for example, for illuminating reviews of early TOEFL-based predictive validity studies; the IELTS Research Reports series, Volume 1 [Wood, 1998], Volume 2 [Tulloh, 1999], Volume 3 [Tulloh, 2000] and Volume 7 [McGovern & Walsh, 2007] exemplify IELTS-based predictive validity research).

In the Australian context, predictive validity studies almost without exception have concerned the IELTS test, widely-recognised by universities in Australia. A frequent measure is to assess the relationship between, on the one hand, the IELTS grades achieved by NESB students seeking university admission, and, on the other hand, the end-of-semester Grade Point Averages (GPAs) they subsequently achieve on university award programs. ‘Grade correlation’ studies of this type for the purposes of investigating predictive validity have been done at, for example, universities in New South Wales (Woodrow, 2006), South Australia (Feast, 2002), Tasmania (Cotton & Conrow, 1995), Victoria (Hill, Storch & Lynch 1999; Kerstjens & Nery, 2000), Western Australia (Dooey & Oliver, 2002), and across Australia generally (Huong, 2001). Despite inconsistencies in their findings (see 1.3 below), such studies may be said to have collectively played a role in the validation of IELTS as means of language assessment for international students seeking admission to a range of Australian universities. No comparable study, however, has been done for the predominant means of language assessment for international students used at the site of the present study, the English Language Programs (ELP) unit within the National Centre for English Teaching and Research (NCELTR)[1], at Macquarie University (MU): NCELTR-developed academic preparation and language assessment programs, recognised by MU, generically known as ‘Direct Entry’ programs.[2]

The role of Direct Entry programs as a means of satisfying university entry language requirements is significant not only at NCELTR and MU: comparable ‘in-house’ academic preparation and language assessment programs have become widespread among Australian universities. A website search of universities in metropolitan Sydney and the Sydney region, for example, shows that comparable programs are offered by language centres at the universities of New South Wales; Sydney; Technology, Sydney (UTS); and Western Sydney (UWS); as well as Newcastle and Wollongong. In addition, as in the case of Macquarie University, admissions offices at some of these universities also recognise language assessment scores awarded by EAP programs offered by local independent language colleges, such as the Australian College of Languages (ACL). Disturbingly, however, no validation research has been published on the Direct Entry assessments that have become generalised in the Australian context.

1.1 THE 'DIRECT ENTRY' RESEARCH CONTEXT

Direct Entry programs provided by NCELTR at MU, like comparable programs at other Australian universities, are 'Direct Entry' in a descriptive sense: they offer an in-house university-specific qualifying pathway, as distinct from an external language test such as IELTS. NCELTR's Direct Entry programs may also be seen as 'Direct Entry' in a conceptual sense: founded on task-based and text-based principles, they offer course participants experience of some of the concepts, content and language conventions of their respective future fields of study. A descriptive label 'academic language preparation program' is more appropriate than 'English language entry test', given that the curriculum emphasis falls on the use of language in tasks that are closely related to future MU academic programs, and performance on these tasks rather than in a 'test' provides the basis for assessment.

In terms of authenticity in communicative language testing, the NCELTR Direct Entry approach may be located near the direct / authentic end of a hypothesised direct-indirect continuum (following Bachman, 1990). It was a guiding principle of assessment task design that assessment tasks be embedded in the curriculum with a view to eliciting instances of performance that closely approximate target language performance in destination university disciplines. The focus of assessment, as at MU, falls on instances of written and spoken performance; the means of assessing writing and speaking is criterion-referenced measurement. Gradings of student performances over the ten-week duration of Direct Entry programs, using the MU F/PC/P/Cr/D/HD grading protocol, are consolidated in week 10 into a single overall Direct Entry grade – which is then reported, as the final result, to the student concerned and to MU admissions offices.

The first such program offered by NCELTR was the eponymous Direct Entry English Program (DEEP), which between 1997 and 2000 catered for all NCELTR's international students on the Direct Entry as opposed to IELTS pathway into MU (subsequently, following the introduction of discipline-specific NCELTR Direct Entry programs for business-related MU programs, DEEP has principally catered for Direct Entry candidates with orientations in the humanities and sciences). The Business Preparation Program (BPP) was established in late 2000 to cater for the burgeoning numbers of overseas students with orientations in commerce. In turn, the Accounting Preparation Program (APP)[3] was established in 2001 in response to the need for a Direct Entry program linked to the content and the differing commencement dates of MU's academic programs in accounting.

Since the inception of NCELTR's Direct Entry pathway in late 1997, Direct Entry enrolments have experienced dramatic growth: there were 12 Direct Entry students at program completion in mid 1998, and in mid 2002, following the introduction of BPP and APP, the total number of Direct Entry students at program completion was 202. A corresponding growth did not occur for IELTS preparation programs: the almost universal first preference of NCELTR students intending to study at MU has been for the Direct Entry pathway. For the cohort of international students on an NCELTR Direct Entry or IELTS university-entry language preparation pathway in late 2002, which was the present study's target group, less than 20% were IELTS candidates and over 80% were Direct Entry candidates.

The principal aim of this study, in its original formulation, was to conduct at MU a predictive validity study, including within its scope both the IELTS and highly subscribed Direct Entry programs that together constituted NCELTR's university entry pathway, and so replicate at MU what has been done for IELTS at universities elsewhere in Australia. The IELTS aspect of the study was not to proceed beyond the data collection phase of the study, however, since effective rates of participation for IELTS students proved too low for IELTS data to be statistically meaningful; consequently, the Direct Entry aspect was to be the study's sole concern.

1.2 PATTERNS IN PREVIOUS RESEARCH

Internationally, predictive validity studies with a common concern with the specifying of optimal language levels for admissions purposes have generated a wide range of correlation coefficients signifying the relationship between language ability (the predictor variable) and academic performance (the criterion variable). A coefficient of $r = 0.30$, which signifies what is termed a 'weak' positive relationship between variables, and which on the basis of the corresponding coefficient of determination $r2 = 0.09$ signifies that the predictor variable explains only 9% of the variance in criterion variable, seems generally to be seen as a standard sufficient for validation purposes (Alderson, Clapham & Wall, 1995; Criper & Davies, 1988, as cited in Lynch, 1994; Davies, 1988, as cited in Hill, Storch & Lynch, 1999). Positive correlations have nonetheless been observed as high as $r = 0.52$ for IELTS (Bellingham, 1993)[4] and $r = 0.59$ for TOEFL (Gue & Holdaway, 1973, as cited in Graham, 1987); in contrast, other studies have found no statistically significant correlation between overall language test scores and academic outcome (Dooey & Oliver, 2002; Kerstjens & Nery, 2000), or a negative overall correlation (Cotton & Conrow, 1995). Predictive validity studies based on grade correlation methods have thus been marked for the inconsistency of their findings, an observation often repeated in the literature, and this inconsistency has hindered the making of generalisations on the relationship between language ability and academic performance. This is not altogether surprising given inconsistencies between studies in the variables tested for correlation: it is questionable whether correlations between one language test and GPA at one university and another language test and GPA as calculated by another university are unproblematically comparable. More fundamentally, in the recent view of Ingram and Bayliss (2007, pp. 141-142), a relationship between language ability and academic performance can be no more than "hypothetical" given that academic performance is subject to multiple other influences, such as intellect, motivation and acculturation.

A number of studies have considered how one or more moderator variables may affect the primary relationship between predictor and criterion variables, however, and from these studies patterns have emerged across studies which would seem to allow for the making of generalisations. In all cases, the moderator variables considered have a close relationship to the criterion variable. A 'level of study' moderator variable has been considered, for example, though more often perhaps by default than by design: studies conducted by Elder (1993), Ferguson and White (1998) and Lynch (1994, 2000) were all based exclusively on postgraduate samples. It is notable that all found positive relationships between language ability and academic performance with correlation coefficients approximating or exceeding $r = 0.30$; the findings of Huong (2001), which were based on a sample made up of both undergraduates and postgraduates, moreover generally reflect stronger positive relationships for the postgraduate group. A 'discipline of study' moderator variable has also been considered: studies by Elder (1993), Bellingham (1993) and Ferguson and White (1998), which were all based on homogenous samples in terms of disciplinary orientation (education, commerce and life sciences respectively), all produced positive correlations that exceeded $r = 0.30$. Huong's (2001)study, in which disciplines were grouped into 'Linguistically Demanding' and 'Linguistically Less Demanding' categories, produced generally similar findings, adding weight to an emerging sense that disciplinary orientation is a meaningful moderating influence. [5] In addition, what may be termed an 'elapsed time' moderator variable has frequently been an element of research designs. While the majority of studies use GPAs measured after the first semester of study, some have also included second (e.g. Huong, 2001) and subsequent (e.g. Feast, 2002) semesters, and still others have also included a one-year overall GPA (e.g. Cotton & Conrow, 1995; Dooey & Oliver, 2002). A pattern is evident in findings that the most meaningful measure is GPA after one semester of study, when in terms of elapsed time the predictor measure is closest to the criterion measure and the influence of the former is least dissipated by the passage of time (see, for example, Elder, 1993; Hill, Storch & Lynch, 1999; Huong, 2001).

While the research methods and findings of previous predictive validity studies closely informed the present study's principal aim (see 1.1 above), they were not the sole influence. Given that this principal research aim emerged in a most immediate sense from recommendations made in an earlier NCELTR DEEP evaluation study (see endnote 2), the intended outcome of the research question was originally framed in the theoretical context of a language program evaluation. To adopt terminology used in Lynch's review of rationales driving language program evaluation research (Lynch, 1996), research objectives were construed as both summative and formative – where summative objectives focus on the specification of the relationship between language assessments and subsequent academic performance, and formative objectives concern findings that may be fed back into program improvement. With the exception of studies conducted at Brock University, Ontario (Black, 1991) and the University of Edinburgh (Lynch, 1994, 2000), which similarly concern locally-developed language assessments, there are remarkably few precedents in the literature of predictive validity studies that engaged with dual summative / formative research objectives.

The present study was, nonetheless, also responsive to a tendency in the literature that suggests that predictive validity studies based on grade correlation methods may be of questionable real value. Indications have emerged from research – including studies involving international students in Australia (e.g. Cotton & Conrow, 1995; Dooey & Oliver 2002; Feast, 2002; Hill, Storch & Lynch, 1999) – that there might be other factors, apart from or in addition to English language ability (such as country of origin, motivation, previous study, and social adaptation to the host country) that may be significant predictors of subsequent academic performance. A secondary aim of this present study, therefore, was to explore the extent to which selected non-linguistic factors may predict the academic performance of overseas students at MU. Given constraints of space, however, findings in relation to this secondary research aim are not reported here but await separate treatment.

1.3 RESEARCH OBJECTIVES

The study's principal aim expressed in the form of research questions were as follows:

- In the MU setting, what relationship exists between the academic *language ability* of international students (as measured by NCELTR Direct Entry grades) and their *academic performance* (as measured by MU GPAs after one semester of study)?

- In the MU setting, what levels of *academic performance* (as measured by MU GPAs after one semester of study) are associated with international students who entered MU via NCELTR Direct Entry programs?

The central intention was to generate information on the predictive validity in the MU setting of language assessment scores gained on completion of one of the three NCELTR Direct Entry programs (DEEP, BPP or APP). Research objectives were seen as twofold: it was envisaged that the present study's 'grade correlation' findings, if positive, could generate summative evidence of some value in the validation of NCELTR's Direct Entry programs for international students; and, if negative, would serve formative ends through providing an empirical starting point for remedial action.

2. METHOD

2.1 PARTICIPANTS

All target participants had already satisfied the academic requirements for entry to their chosen MU programs, and had offers of enrolment from the university's admissions office for early 2003 conditional on their satisfying English language requirements. A total of 17 Direct Entry learner groups were involved, made up of 4 classes for DEEP, 9 for BPP and 4 for APP, and a total of 179 students. From this population, 154 participants were recruited to the study, constituting a participation rate of 86% at the recruitment stage.

This 86% figure nonetheless represented only a potential participation rate, since participation in the study also required enrolment on an MU program and completion of a semester of study so as to generate a first semester MU GPA. It transpired that 16 of the 154 potential participants either did not enrol at MU or did so only briefly before withdrawing, as shown in Table 1, leaving an adjusted participation rate of 77% (138 out of 179).

Table 1

Participation rates: DEEP, BPP and APP

DirectEntry Program	**Direct Entry program 'finishers', 2003[a]**	**Participants: potential**	**Non-enrolments/ withdrawals[b]**	**Participants: adjusted[c]**
DEEP	36	34	1	33
			(3%)	(92%)
BPP	96	81	13	68
			(16%)	(71%)
APP	47	39	2	37
			(5%)	(79%)
Total	179	154	16	138
	(81%)	(86%)	(10%)	(77%)

[a] Percentages in brackets are relative to all Direct Entry program 'finishers'.
[b] Percentages in brackets are relative to potential participants.
[c] Percentages in brackets are relative to program 'finishers' for respective programs.

The 138 participants that made up the sample comprised 33 from DEEP, 68 from BPP and 37 from APP. The participant group as a whole was predominantly of east Asian origin, with the majority destined for postgraduate study (83% overall; 94% DEEP, 70% BPP, 100% APP). A wide range of disciplines was represented, of which those relating to commerce and accounting (catered for by BPP and APP) were particularly well subscribed. The participant group was mainly in the 20 to 30 age group (84% overall; 76%DEEP, 95% BPP, 80% APP), with women slightly outnumbering men. Selected characteristics of the sample are given additional detail in Tables 2 and 3.

Table 2: Distribution of participants by country

	DEEP n=33		BPP n=68		APP n=37		All Direct Entry Programs	
Country	**Number**	**Percent**	**Number**	**Percent**	**Number**	**Percent**	**Number**	**Percent[a]**
China	13	39%	60	88%	34	92%	107	78%
Korea	9	27%	1	2%	2	5%	12	9%
Thailand	4	12%	2	3%	--	--	6	4%
Japan	3	9%	--	--	1	3%	4	3%
Turkey	2	6%	1	2%	--	--	3	2%
Indonesia	--	--	2	3%	--	--	2	1%
Taiwan	2	6%	--	--	--	--	2	1%
Germany	--	--	1	2%	--	--	1	1%
France	--	--	1	2%	--	--	1	1%
Total	33	99	68	103	37	100	138	100

[a] Percentages rounded to the nearest whole number in both Tables 1 and 2.

Table 3: Distribution of participants by discipline[a] of study

MU Division	DEEP n=33		BPP n=68		APP n=37		All Direct Entry Programs	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Economic and Financial Studies (EFS)	--	--	65	96%	37	100%	102	74%
Linguistics and Psychology (L&P)	10	30%	--	--	--	--	10	7%
Society, Culture, Media and Philosophy (SCMP)	7	21%	--	--	--	--	7	5%
Australian Centre for Educational Studies (ACES)	6	18%	--	--	--	--	6	4%
Humanities (Hum)	5	15%	--	--	--	--	5	4%
Graduate School of Management (GSM)	--	--	3	4%	--	--	3	2%
Environmental and Life Sciences (ELS)	3	9%	--	--	--	--	3	2%
Law (Law)	2	6%	--	--	--	--	2	1%
Total	33	99	68	100	37	100	138	99

[a] 'Discipline' here is broadly defined in terms of MU's academic structure, in which disciplines are grouped into Divisions.

2.2 PROCEDURE

For both research questions, the following data were used:

NCELTR Direct Entry grades (measuring the construct language ability, the independent or predictor variable).

MU GPAs after one semester of study (measuring the construct academic performance, the dependent or criterion variable).

The process of data collection and data analysis broadly occurred in three main stages from 2002 to 2004. In stage one, participants were recruited from the late 2002 deliveries of NCELTR Direct Entry (and IELTS) programs, and records for all participants of final grades achieved on their respective NCELTR programs were assembled. Stage two, implemented in the second half of 2003 after participants had completed their first semesters / trimesters of study at MU, involved the collection of MU academic records for all participants. In stage three, Pearson's product moment correlations (r) were computed for the two variables *language ability* (as measured by Direct Entry grades) and *academic performance* (as measured by MU GPAs) and – as a further, descriptive means of appraising the predictive validity of NCELTR's Direct Entry programs – mean MU GPAs and overall pass rates achieved by study participants were calculated.

Standard procedures were performed to assess the suitability of the data for the statistical test used. Scatterplots providing a visual impression of relationships between the variables *language ability* and *academic performance* (see Appendix A) were generated for each of the DEEP, BPP and APP data sets to assess for linearity; histograms were generated reflecting the distributions of Direct Entry grades and MU GPAs to assess distributions for normality; and data were examined for homogeneity of variance.

3. DATA ANALYSIS AND DISCUSSION

3.1 PREDICTIVE VALIDITY OF DIRECT ENTRY GRADES

It needs to be noted that the correlations between Direct Entry grades and MU GPAs set out in Table 4 below relate to data sets in which two potential moderator variables are, to a significant extent, controlled. First, since all MU GPAs were collected after one semester / trimester of study, a potential 'elapsed time' moderator variable is taken into account. Second, since DEEP, BPP and APP are to varying extents programs with discipline-specific orientations (DEEP for humanities and sciences; BPP for commerce; APP for accounting), a potential 'discipline of study' moderator variable is, to varying extents, taken into account.

Of the three programs, as already noted (see 1.2 above), DEEP is the least discipline-specific and APP the most.

Table 4

DEEP, BPP, APP grades and MU GPAs: Pearson product moment correlations (r)

	MU GPAs at one semester/trimester
DEEP Grades (N=33)	r = 0.361[a]
BPP grades (n=68)	r = 0.156
APP grades (n=37)	r = 0.418[a]

[a] correlation significant at the 0.05 level

Table 4 shows all correlations to be positive, reflecting for the samples general agreement between the two variables *language ability* (as measured by DEEP, BPP and APP grades) and *academic performance* (as measured by MU GPAs). In terms of strength of correlation, the coefficients for the DEEP and APP data sets are located near the borderline between value ranges held to represent ‘weak’ (r = 0.20 – 0.40) and ‘substantial’ (r = 0.40 – 0.70) relationships between variables.[6] Both are statistically significant, enabling extrapolation to wider DEEP and APP populations. The correlation for the BPP data sets, however, though positive, reflects a relationship so weak as to be negligible and did not achieve statistical significance.

The highest correlation is recorded for APP data (r = 0.418, significant at the 0.05 level), signifying that approximately 17.5% of variance in MU GPAs is accounted for by variance in APP grades. This, seen in the context of findings generally reported in the literature, represents an unusually strong relationship. The correlation recorded for the DEEP data (r = 0.361, significant at the 0.05 level) correspondingly signifies that 13% of variance in MU GPAs is accounted for by variance in DEEP grades. This finding, too, is relatively strong. [7] The BPP correlation finding, by contrast, was unusually weak.

So as to investigate whether BPP correlation findings may differ for particular subgroups within the BPP cohort, additional correlations were done with ‘level of study’ and ‘program of study’ moderator variables taken into account. Since APP enrols only students destined for a postgraduate level of study and essentially for only one program of study, the Master of Accounting degree (MU Division of Economic and Financial Studies), these moderator variables were already built in to the APP data sets. Correlations for ‘level of study’ were consequently done only for DEEP and BPP data sets.

Of the 33 DEEP participants, 31 were postgraduates, representing 94% of the whole DEEP sample. The DEEP correlation with the moderator variable ‘postgraduate level of study’ taken into account (r = 0.372) was marginally stronger relative to the correlation for the whole DEEP data set (r = 0.361), and DEEP grades were correspondingly improved predictors of MU GPAs (explaining approximately 14% of the variance). Of the 68 BPP participants, 47 were postgraduates, representing approximately 70% of the whole BPP sample.The BPP correlation with ‘postgraduate level of study’ taken into account represented a decrease relative to the finding for the whole BPP data set, and statistical significance was again not achieved.

Correlation analysis could not be done for a DEEP undergraduate subgroup since it included only 2 participants. The BPP correlation finding with ‘undergraduate level of study’ taken into account represented a marginal increase relative to the finding for the whole BPP data set; once again, however, statistical significance was not achieved.

Correlations with the ‘program of study’ moderator variable taken into account were limited to the BPP data set; the largest group of participants enrolled on a single program of study in the DEEP sample (n = 9) constituted a sample size too limited for statistical analysis. The largest group of participants enrolled on a single program of study in the BPP sample (n = 17) were enrolled on the Master of Commerce in Accounting and Finance program (MU Division of Economic and Financial Studies). The BPP correlation finding with ‘program of study’ taken into account represented a marked decrease relative to the finding for the whole BPP data set; this BPP correlation finding was moreover negative – signifying that, for this particular n = 17 sample, an increase in language ability as measured by BPP grades was associated with a decrease in academic performance as measured by MU GPAs. Statistical significance was again not achieved.

The correlation findings for NCELTR’s DEEP, BPP and APP Direct Entry grades and MU GPAs were therefore mixed. APP, in particular, and DEEP were shown to be programs whose assessments of academic *language ability* correlate well (in the context of findings reported in the literature) with MU assessments of *academic performance*. The APP findings seem also to add weight to the emerging view that grade correlations are stronger when the potential moderator variables ‘level of study’ and ‘discipline of study’, and even more particularly ‘program of study’, are taken into account; it should be noted, nonetheless, that the comparatively strong correlations found for the APP cohort (almost exclusively enrolled on the Master of Accounting program) might be partly explained by a likely greater reliability of the criterion variable (MU GPAs computed from grades awarded only within the MAcc program) rather than solely a relative increase in the predictive validity of the predictor variable (APP grades). BPP, in contrast, was shown to be a program whose assessments correlate poorly with MU assessments. Given the formative purposes that were included in the aims of this study, the BPP findings raised questions that the study sought to address – as the following discussion shows.

3.2 ABSOLUTE MEASURES OF PREDICTIVE VALIDITY

As a further, descriptive means of appraising the predictive validity of NCELTR Direct Entry programs that study participants had completed as a means of entering MU, absolute measures of *academic performance* – overall pass rates and mean MU GPAs achieved by study participants at the end of their first semester / trimester of study – were considered. Following the formula for the calculation of MU GPAs in the Handbook of Undergraduate Studies (Macquarie University, 2002), pass rates were established using a mean MU GPA of 2.000 as the pass / fail watershed (Table 5).

Table 5

DEEP, BPP and APP participants: pass rates at one semester / trimester

	Number	**Percent**
DEEP participants (n=33)	33	100%
BPP participants (n=68)	48	71%
APP participants (n=37)	27	73%

These findings show a discrepancy between overall pass rates at the first semester of study for DEEP participants in comparison to BPP and APP participants: whereas the pass rate was 100% for DEEP, pass rates for BPP and APP were little more than 70%.[8] Accordingly, the academic performance of BPP and APP participants may be seen as comparable, and the status of BPP and APP 'pass' grades collectively (made up of the Pass / Credit / Distinction / High Distinction grades in use at MU) may in turn be seen as comparable in terms of their ability to predict subsequent academic performance in MU programs; BPP in this light appears no poorer an academic preparation program than APP despite the stark difference between grade correlation findings for the two programs. This reasoning seems supported by further descriptive statistics in the form of mean MU GPAs achieved by DEEP, BPP and APP participants (Table 6).

Table 6

DEEP, BPP and APP participants: mean MU GPAs at one semester / trimester

	Mean MU GPA
DEEP participants (n=33)	3.042
BPP participants (n=68)	2.229
APP participants (n=37)	2.382

The figures shown in Table 6 similarly reflect a discrepancy between mean MU GPAs achieved at the first semester of study by DEEP participants in comparison to BPP and APP participants: the DEEP mean of 3.042 reflects a mean level of performance marginally exceeding the MU ‘Credit’ range (65—74%), while the BPP and APP means of 2.229 and 2.382 respectively reflect mean levels of performance exceeding, but proximate to, the MU ‘Pass’ range (50—64%).

The similarities between the academic performance of participants who entered MU via BPP and APP, both of which prepare students almost exclusively for degree programs offered by the same MU academic division, seemed to indicate that the poor BPP grade correlation findings were related to BPP assessment practices. Interviews were consequently conducted with BPP teaching staff, from which it became apparent that a systematic process of cross marking of student assignments so as to monitor inter-rater reliability, as existed in DEEP and APP, was not in place in BPP. Recommendations for amendments in this domain were accordingly made, as will be outlined in the conclusion that follows. Given the comparable performance of participants who entered MU via BPP and APP, however, and that these two Direct Entry programs almost exclusively service only the MU Division of Economic and Financial Studies, this study raises formative questions that go beyond program-internal recommendations, as will also be outlined below.

4. CONCLUSIONS AND RECOMMENDATIONS

A key motivation for the study was to replicate for Direct Entry programs provided by NCELTR on the MU campus the kind of predictive validity studies that have been conducted at universities elsewhere in Australia and internationally for more widely known standardised language assessments, such as TOEFL and IELTS. Given that locally-developed and unvalidated Direct Entry methods of language assessment predominate in the NCELTR / MU context, however, the study was construed as having both summative and formative purposes.

Concerning the study's summative outcomes, substantial evidence was generated for the validation of DEEP and APP, but less so for BPP. The study's correlation analyses show the predictive validity of DEEP and APP assessments to both meet and exceed (especially in the case of APP) expectations conventionally held for language test predictive validity studies; BPP assessments, in contrast, were shown to be erratically associated with subsequent assessments awarded by MU. The findings of the study's investigation of predictive validity through considering absolute measures of academic performance (pass rates and mean MU GPAs achieved by students who had completed one of three NCELTR Direct Entry programs), however, validate DEEP but raise questions about the academic performance of students who had completed either BPP and APP.

The study's formative outcomes, deriving in the first instance from problematical aspects of its summative findings recapped immediately above, led to recommendations for program improvement and further inquiry. A key recommendation emerging from the grade correlation phase of the study was to introduce into BPP a more rigorous process of cross marking between program teaching staff so as to increase the reliability of BPP assessments. This has been acted on in subsequent BPP deliveries, and allied to this the frequency of assessment tasks over the duration of BPP has been reduced so as to enable teaching staff more time for careful marking. The impact of these amendments on the predictive validity of BPP grades has yet to be investigated. A further recommendation of a longer-term formative nature is for investigation into the comparatively lower mean academic performance of BPP and APP participants relative to DEEP participants. Such an investigation might validly proceed from two distinguishing characteristics of BPP participants (n=68) and APP participants (n=37) seen as one sample group (n=105): as the sample description (Tables 2 and 3) reflects, the overwhelming majority were completing programs offered within the MU Division of Economic and Financial Studies (97%, as opposed to 0% for DEEP), and were of Chinese origin (90%, as opposed to 39% for DEEP).

While the specific relevance of the study's findings is to the particular MU context in which the study was conducted, the study has exemplified an approach to language assessment validation that has a more general significance. The study has sought to link formative outcomes to the customary summative outcomes of predictive validation studies of language assessment, and in so doing has further instantiated an approach to predictive validity research (following Black, 1991; Lynch, 1994, 2000 – see 1.3 above) that increases the capacity of predictive validity studies to implement change, if and where found to be necessary, in assessment practices. This potential in predictive validity research for research-driven program-renewal outcomes seems of immediate relevance in the current Australian context, in which locally-developed, and evidently unvalidated, 'Direct Entry' approaches to language assessment have become widespread.

ENDNOTES

1. Following organisational restructuring in 2008, ELP NCELTR is now named the Centre for Macquarie English (CME). I would like to acknowledge the financial support for this study provided by NCELTR research funds. In addition to the project leader (the author), project personnel included four other NCELTR language program convenors (Anna Poros, Cintia Agosti, Peter McCulloch, Rosemary Costley) who contributed in the recruitment of participants / analysis of data relating to the programs they convened.
2. A small-scale predictive study of NCELTR's longest-established Direct Entry program has been conducted as an element of a broader program evaluation (Cope, N. and Hennessy, M. [2002] Validation study and evaluation of the Macquarie University 'Direct Entry English Program'. [Unpublished report, ELP, NCELTR]), but the sample size (21) did not allow extrapolation beyond the sample group.
3. APP is delivered three times yearly to link with the trimester intakes for programs offered by the MU Department of Accounting and Finance, in particular the Master of Accounting; DEEP and BPP are delivered twice yearly, directly preceding commencement dates of MU's conventional first and second semesters.
4. A higher finding of $r = 0.540$ for IELTS (Hill, Storch & Lynch, 1999) has been reported but without confidence: as the authors observe (1999, p. 55), an examination of the study's data confirmed that "assumptions of the regression model had been violated".
5. There are many possible explanations for this effect; one may be the possibly increased reliability of the criterion measure, GPA, when generated by a single discipline or even department. As so often with language test predictive validity studies, however, there is also counter evidence: Dooey and Oliver (2002, p. 49) took 'discipline of study' into account but found "not a major difference … between IELTS scores and academic success across disciplines".
6. These coefficient ranges and interpretations follow Burns (1997, p. 198).

7. The correlation coefficient for the DEEP data set is based on a restricted range of the MU GPA scores in comparison to the BPP and APP data sets: the MU GPA range is 2.00 for DEEP, and 3.50 and 3.80 respectively for BPP and APP (see 3.2 above). This serves to restrict the potential of the DEEP grade / MU GPA correlation statistic.
8. Scatterplots of DEEP, BPP and APP grades in relation to MU GPAs (Appendix A) provide a visual impression of the pattern of pass rates for the respective programs.

REFERENCES

Alderson, J., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: CUP.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: OUP.

Bellingham, L. (1993). The relationship of language proficiency to academic success of international students. *New Zealand Journal of Educational Studies*, *30*(2), 229-232.

Black, J. (1991). Performance in English skills courses and overall academic achievement. *TESL Canada Journal*, *9*(1), 42-53.

Burns, R. (1997). *Introduction to research methods*. Melbourne: Longman.

Cotton, F. & Conrow, F. (1995). *An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania*. IELTS Research Reports (Vol. 1). Canberra: IELTS Australia.

Dooey, P. & Oliver, R. (2002). An investigation into the predictive validity of the IELTS test as an indicator of future academic success. *Prospect*, *17*(1), 36-54.

Elder, C. (1993). Language proficiency as a predictor of performance in teacher education. *Melbourne Papers in Applied Linguistics*, *2*(1), 68-85.

Feast, V. (2002). The impact of IELTS scores on performance at university. *International Education Journal*, *3*(4), 70-85.

Ferguson, G. & White, E. (1998). *A predictive study of IELTS*. Institute for Applied Language Studies, University of Edinburgh.

Graham, J. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly*, *21*(3), 505-521.

Hale, G., Stansfield, C. & Duran, R. (1984). *Summaries of studies involving the Test of English as a Foreign Language, 1963—1982* (Report no. RR-84-03, TOEFL-RR-16). Princeton, N. J.: Educational Testing Service.

Hill, K., Storch, N. & Lynch, B. (1999). *A comparison of IELTS and TOEFL as predictors of academic performance*. IELTS Research Reports (Vol. 2). Canberra: IELTS Australia.

Huong, T. (2001). The predictive validity of the International English Language Testing System (IELTS) test. *Post-Script*, *2*(1), 66-94.

Ingram, D. & Bayliss, A. (2007). *IELTS as a predictor of academic language performance, Part 1*. IELTS Research Reports (Vol. 7). Canberra: IELTS Australia.

Kerstjens, M. & Nery, C. (2000). *Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance*. IELTS Research Reports (Vol. 3). Canberra: IELTS Australia.

Lynch, B. (1996). *Language program evaluation: Theory and practice*. Cambridge: Cambridge University Press.

Lynch, T. (1994). The University of Edinburgh Test of English at Matriculation: Validation report. *Edinburgh Working Papers in Applied Linguistics*, No. 5. Scotland: Edinburgh University.

Lynch, T. (2000). An evaluation of the revised Test of English at Matriculation at the University of Edinburgh. *Edinburgh Working Papers in Applied Linguistics*, *10*, 61-71.

Macquarie University. (2002). *Macquarie University handbook of undergraduate studies*. Sydney: Macquarie University.

McGovern, P. & Walsh, S. (Eds.). (2007) *IELTS research reports* (Vol. 7). Canberra: IELTS Australia.

Tulloh, R. (Ed.). (1999). *IELTS research reports* (Vol. 2). Canberra: IELTS Australia.

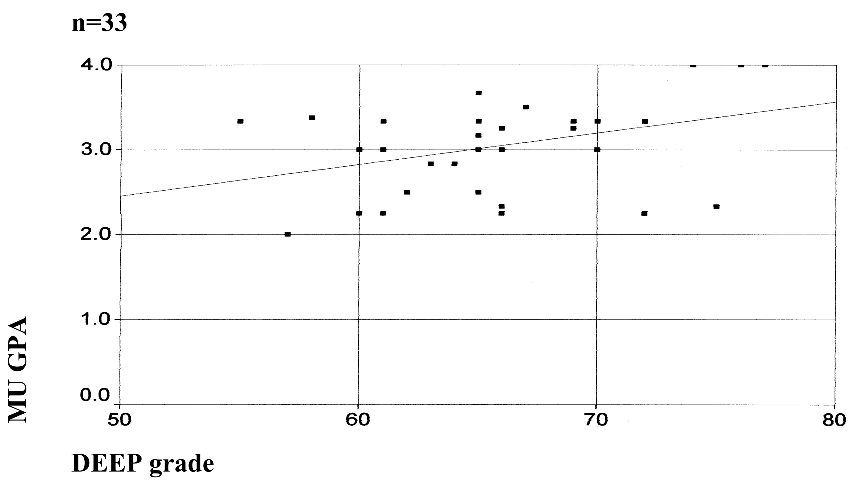
Tulloh, R. (Ed.). (2000). *IELTS research reports* (Vol. 3). Canberra: IELTS Australia.

Wood, S. (Ed.). (1998). *IELTS research reports* (Vol. 1). Sydney: ELICOS Association.

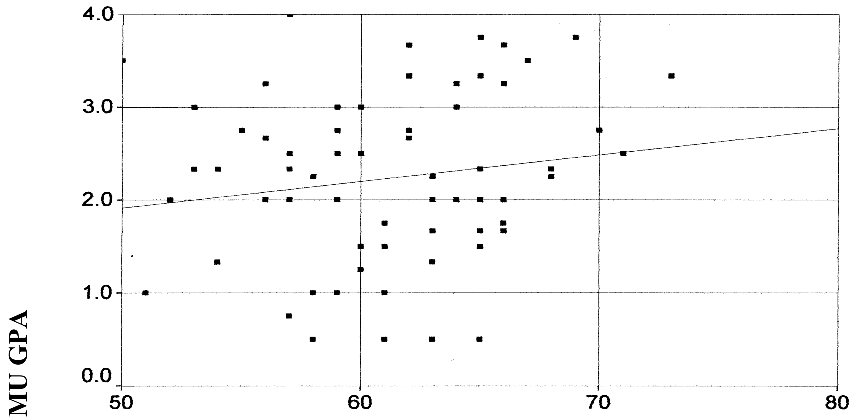
Woodrow, L. (2006). Academic success of international postgraduate education students and the role of English proficiency. *University of Sydney Papers in TESOL*, *1*, 51-70.

APPENDIX A

SCATTERPLOTS OF DE GRADES / MU GPAS

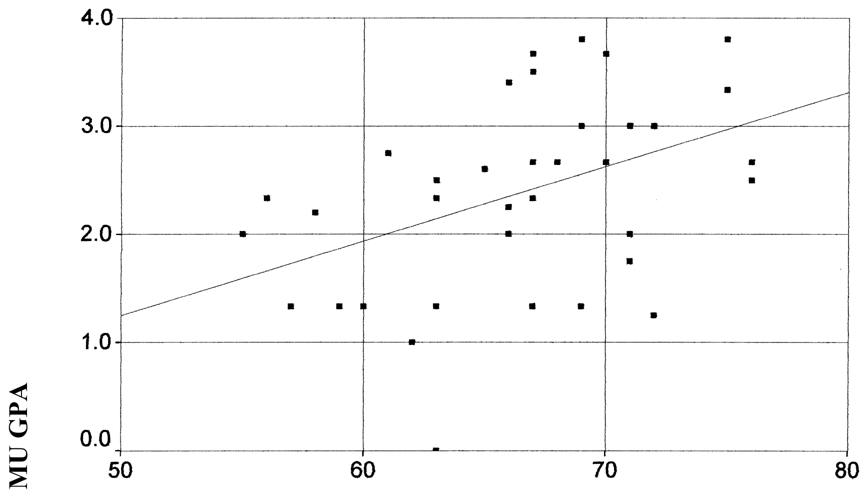


n=68



BPP grade

n=37



APP grade