NAPLaN test data, ESL Bandscales and the validity of EAL/D teacher judgement of student performance

SUE CREAGH
Institute for Social Science Research (ISSR), University of Queensland

Abstract: *Teachers are now experiencing the age of quantitative test-driven assessment, in which there is little weight accorded to teacher-based judgement about student progress. In the Australian context, the NAPLaN test has become a driving force in school and teacher accountability. The language of NAPLaN is one of bands and numerical scores and comparative performance, measuring schools against other 'similar' schools and against a national average. The consequences of this are troubling for all teachers. For EAL/D teachers whose specialised professional knowledge relates to building the academic English language of EAL/D learners, NAPLaN is highly problematic because it takes no account of second language factors which might impact on test performance. Yet, NAPLaN data do offer rich yet largely unexploited opportunity to highlight the validity of teacher judgement in the classroom. This paper will use ESL Bandscale data in an analysis of the NAPLaN performance of EAL/D students to show how teacher judgement (captured by the ESL Bandscales) is valid and aligned with NAPLaN performance. I will demystify the power and the fallibility of large-scale assessment like NAPLaN: to identify in which contexts large-scale data analysis is useful, and its limitations in micro-settings which include the classroom. The second goal of this paper is to stress the utility of teacher data, measured quantitatively, but based on qualitative observation, when it is grounded in teacher professional knowledge. Ultimately, such arguments serve to highlight the importance of remaining grounded in strong TESOL pedagogy, despite the intense pressure to follow mainstream English (as first language) literacy responses to NAPLaN testing.*

Keywords*: NAPLAN, Language Background Other Than English, LBOTE, English as a Second Language, ESL Bandscales*

ISSN 1030-8385

Introduction

In this paper I am presenting empirical evidence that there is an association between performance on standardised tests of literacy and numeracy and the English language level of English as a Second Language (ESL)[1] students. I am presenting this argument in response to the existing government processes of disaggregating national Australian test data in relation to language background and am arguing that a poor statistical category – Language Background Other Than English - does not provide sufficient information about the relationship that logically exists between English language level and performance on English-only tests of literacy and numeracy. In building my empirical analysis I am drawing on measures of English language proficiency which have been developed by ESL teachers in Australia, reflecting the specialist knowledge contained within the pedagogical field of ESL. The implementation of standardised testing is generally constructed by policy makers and governments as one in which testing is able to achieve an objective appraisal of student ability (Comber, 2012; Polesol, Rice & Dulfer, 2013); such a view, conversely (and erroneously), suggests that teacher judgement is potentially flawed, biased or unable to achieve the same level of objectivity. In education policy literature, this negative reformulation of teacher judgement is symptomatic of current global education reform movements in which "complex social processes" which occur in learning and teaching are seemingly able to be transformed into categories and numbers, which are then able to be measured (Ball, 2006, 144). The devaluing of teacher judgement in national assessment discourse will be of dire consequence to classroom pedagogy and will potentially reduce, rather than enhance, the capacity of schools and education systems to *argue for* and respond to the unique learning needs of the diverse groups of students who constitute schools and classrooms. In this paper I am specifically concerned with the implications of this situation for the continued policy support of ESL, both in terms of funding and pedagogy.

By way of background, Australia has held national tests in literacy and numeracy since 2008 for all children in schools years 3, 5, 7 and 9[2]. The National Assessment Program: Literacy and Numeracy (hereafter NAPLaN) test has preceded the implementation of an Australian national curriculum, (currently in progress). Instead of linking to a national curriculum, the NAPLaN test embodies nationally agreed

[1] ESL is being replaced by the term English as an Additional Language/Dialect (EAL/D) in Australian policy documentation. For its familiarity I will retain the term ESL in this paper.

[2] Exemptions are allowed for students for whom English is a second or additional language, in their first year of residency in Australia only.

Statements of Learning, upon which all State and Territory curricula are based (Australian Curriculum Assessment and Reporting Authority (ACARA) 2011b), representing common English and Mathematics knowledge, skills, understandings and capacities. The Statements of Learning contain no reference to students for whom English is a second or additional language and assume continuity of education in the Australian context: at year 9, a student being tested in NAPLaN is assumed to speak English as first language and to have continuous and unbroken education in Australia (Curriculum Corporation, 2005). NAPLaN tests constitute a suite of exam papers in reading, writing, grammar, spelling and numeracy. There is a time gap of some five months before schools and parents receive NAPLaN results, a factor which has been strongly criticised in recent senate reviews of the test as significantly impacting on the capacity of teachers and schools to usefully respond to student performance (The Senate Education and Employment References Committee, 2014, p.10).

NAPLaN test results are disaggregated across a number of statistical categories: gender, Indigenous status, geo-location, socio-economic status and language background. The latter, Language Background Other than English (hereafter LBOTE) is the only indicator of language. It has a broad definition: the child or their parents speak a language other than English at home (ACARA 2009); however LBOTE fails to provide any indicator of *language proficiency level.* The consequences are that the category is extremely broad, and captures students who range from highly proficiently bilingual, through to students in the very early stages of English language development. The statistical consequence of this breadth is that the category average or mean is influenced by its range and equates to the non-LBOTE average. In other words, there is virtually no difference, at a national level, between LBOTE and non-LBOTE average performance on NAPLaN tests although standard deviation is uniformly greater for LBOTE; this is the case for all year levels, for all years of the test. When the data are examined at state and territory levels there is greater variation across LBOTE and non-LBOTE. In terms of policy response to these data, the consequences are worrying for ESL, because they suggest that there is no relationship between language background and NAPLaN performance and potentially make opaque the need both to fund ESL programs and to support teacher professional development in ESL pedagogy. Indeed LBOTE NAPLaN performance runs counter to recent research with NSW teachers which identified a ‘pressing need’ for access to professional development in teaching ESL (Watkins, Lean, Noble & Dunn 2013, p 25). This paper will now report on research which interrogates the LBOTE data in order to determine whether there is, in fact, a relationship between English language level and NAPLaN test performance.

The paper will proceed in the following way. First, I will provide a brief overview of the research project from which this paper is drawn. I will then focus particularly on the analysis of the NAPLaN performance of a group of LBOTE students from Queensland schools, utilising data collected during 2010 and 2011. Importantly, I will demonstrate that assessment which relies on teacher judgement (application of ESL Bandscales and A to E grades) is clearly associated with NAPLaN performance. I will also critique the NAPLaN test in relation to its utility and relevance for students who are in the process of acquiring English. Finally, I will highlight the importance of maintaining ESL knowledge, and the use of Bandscale data, both for classroom pedagogy, and systemically, for statistical analysis which identifies and quantifies the ESL learner and is able to inform policy in relation to classroom practice, teacher professional development and funding allocation.

The Research Project

The research project I am reporting is drawn from my PhD, completed in 2013. In my research I was exploring the relationship between NAPLaN test performance and English language level in order to determine the capacity of the LBOTE category to adequately represent this relationship. This project brings new knowledge to the fields of applied linguistics and education in the Australian context. To date, there are no published studies which describe the association between English as a second language level, other language related variables and performance on the NAPLaN test.

The project is quantitative in its methodology, and involved the collection of considerable data about a large number of students and the analysis of these data in relation to NAPLaN performance. For the analysis I used multiple regression. Multiple regression is a statistical tool used to explore associations between an outcome or dependent variable (in this case, NAPLaN result) and a number of explanatory variables (language level, education background and socio-demographic factors). This method allows the exploration of the *combined* association of a group of explanatory variables with an outcome variable, whilst isolating the *unique* association of *each* of the explanatory variables in the model (de Vaus 2014). In my datasets, which constitute cross-sectional data drawn from enrolment and test data, the research outcomes cannot support causality (Yang, 2010).

Rather the purpose of the research is to determine statistically whether associations exist, which can then provide empirical support for the theories which have informed the model construction (Yang, 2010).

The kinds of data I have collected, the construction of the dataset and the analysis has been influenced specifically by second language acquisition research. The theoretical projects from this field suggest that the test performance of ESL students will be influenced by their level of proficiency in the test language (Cummins, 1981; Thomas & Collier, 1997). *Proficiency* is defined in relation to Cummins' (1981) conceptualisation of a binary model of language which differentiates basic, routine everyday language from the academic language demands of school. It is the latter, Cummins' Cognitive Academic Language Proficiency (CALP), which should be measured to determine school students' levels of language proficiency (Cummins 1981; Hakuta, Butler & Witt 2000; Thomas & Collier 1997). Research projects, which studied the performance of English language learners using standardised English tests, found that students took from 5 to 7 years to achieve test results equivalent to their English speaking peers (Cummins, 1981) but this time frame was only possible if the students had received foundational schooling up to year 6 in their first language (Thomas & Collier, 1997). Cummins identified an *interdependency* across first and second language such that development of literacy in first language facilitates development of literacy in second language (Cummins, 2000, p.173). However, students who arrived as adolescents had insufficient time to catch up to their English speaking peers during the remaining years of schooling, and Thomas and Collier (1997) identify the issue of increasing complexity of academic work across the school years as a factor which impacts all ESL learners once ESL support programs are completed. Further research by Hakuta et al. (2000) identified that the development of CALP is also associated with socio-economic status (SES), with students from low socio-economic status needing longer. Extrapolating from these studies suggests therefore that level of proficiency in the academic test language will be related to age on arrival and educational history, which will be associated with socio-economic status and years of education in first language and in total (Garcia 2000; Hakuta, Butler & Witt 2000; Thomas & Collier 1997). Further and more recent studies with refugee students in Australian schools indicates that this group of students, who may be characterised by limited schooling and minimal or no literacy in first language, require additional time to achieve academic proficiency (see Miller & Windle, 2010). Drawing on these foundational studies, and using multiple regression as my statistical tool, I am thus examining the unique impact of a selection of education, demographic and socio-economic factors on NAPLaN reading scores.

For the research I collected enrolment and assessment data from a number of schools in the Brisbane metropolitan region of Queensland. The data collection in 2010 and 2011 involved visiting 25 primary, secondary and P-12 state schools in the Brisbane metropolitan region. These schools constitute a representative sample for the purposes of the study because settlement of migrant and refugee communities largely occurs in Australia in urban locations like Brisbane and the students from these communities are more likely to attend government schools. However, there are some issues which pertain to the problematic definition of LBOTE which impacted on the capacity to create a sample representative of LBOTE, as defined for the NAPLaN test. Because of the breadth of its definition, that the child or their parent/s speaks a language other than English at home, LBOTE is not used for educative purposes in Queensland schools. Instead, those students who would satisfy the LBOTE definition are identified as ESL, defined by the Queensland education department as being in the process of acquiring English as a second or additional language and learning curriculum content through this language (Department of Education, Training and Employment, 2013). In other words, the only process for representing the LBOTE category was to include ESL students, even though the category definitions mean that ESL students are a subset of LBOTE. This is problematic because it suggests that the sample may not represent the full breadth of the LBOTE definition; however, this was an insoluble problem. The sample however, is justified on the grounds that this group should be represented by the broader LBOTE category, if the category is to be a fair representation of the impact of language on test performance. This dataset will be referenced as ESL Data. In this paper I will be reporting the results of the analysis of year 9 NAPLaN reading performance. It is the year 9 sample who will be engaging with the most complex academic language, and who have the least time to reach equivalence of educational outcome with their English speaking peers.

Whilst all students satisfied the definition of the LBOTE category, not all were *identified* in the test as being LBOTE. This is an interesting problem associated with the category, which in Queensland is only identified on the test paper, by the teacher administering the paper. For the year 9 group, 18% of these students were not identified on test papers as LBOTE and consequently, are not actually counted in the LBOTE category.

Of relevance to this paper, for each student I collected the following data:

- year level, gender, birthdate, country of birth, cultural background, language/s, visa subclass, entry status, date of arrival to Australia, or indication of birth in Australia or New Zealand
- parent 1 and 2 education levels (school and post school)
- education history, including age of commencing school, years of education before arrival to Australia, language of this education, countries where school was attended, access to English lessons
- Australian education history including years of education in Australia (both in Queensland and interstate) and access to ESL programs
- A to E school results in English and Maths for semester 1 of 2010 or 2011
- ESL Bandscales (proficiency levels in English) in listening, speaking, reading and writing: all historical records, including during semester of NAPLaN test
- NAPLaN scores and band levels for 2010 and 2011, and identification as LBOTE (or not)

Year 9 Descriptive statistics[3]

In this section I will describe the characteristics of the sample group in terms of their average NAPLaN reading performance, demographics, language, and education background. These factors constitute the variables used in the multiple regression models, and align with the theoretical basis for my models which suggests that NAPLaN performance will be associated with language level, and may also be influenced by demographic, education background and socio-economic factors.

Numerically, the NAPLaN test results are measured on a common scale with a mean score of 500. Numerical performance is also translated into 10 achievement bands, and the width of a band is approximately 50 numerical marks. The band levels are anchored somewhat by the designation of a national minimum standard for each year level. For example, in year 9, a student is performing at the national minimum standard in reading, if their numerical score places them in band 6 (approximately 470-520). ACARA (2011a) advises that students *below* the national minimum standard require considerable support to achieve success, and students *at* the national minimum standard may also require targeted interventions as well.

[3] A full list of descriptive statistics for the year 9 group is provided in appendix A.

Table 1 shows the average mean score, standard deviation and number of cases for the year 9 sample group. The reading mean places the group within the range of the national minimum standard for year 9.

Table 1. Year 9 sample reading NAPLaN results, 2010 and 2011.

	Mean	Sd	n
Reading	497.7	56.2	241

Source: ESL Data

The year 9 students were mostly recently arrived to Australia, and the majority (59%) had been here less than 3 years. For the year 9 group, 58% (143) were female. Parent education levels were collapsed into those who had completed year 12 and those who had less than year 12 level of schooling. 33% of parents had completed year 12, 18% had less than this, and 49% of parents had not provided information regarding their schooling. Most (80%) of the group spoke one language other than English; the remaining 20% spoke more than one other language. Most of the group were in or had recently exited an ESL support program and so had ESL Bandscale levels for the time of the NAPLaN test. The provision of ESL services in secondary school in the Queensland metropolitan region means that students move through intensive ESL programs until they enter mainstream classrooms, where they may continue to access ESL support. ESL teachers routinely track student progress using ESL Bandscales as a measure of language progress.

For this analysis, the key variables aimed at disaggregating the *influence of language* on NAPLaN performance include visa category, specific world region of birth, years of education, and language proficiency level at the time of the NAPLaN test. Both visa category and world region of birth are associated with socio-educational opportunities which relate to quality and continuity of schooling experience prior to arrival to Australia. In terms of visa categories, the year 9 group is over-representative of refugee category (43%), but included Australian and NZ residents (15%), business visa families (13%) and family visas (12%).

In addition to visa category, region of birth has been included in the analysis in order to count the influence of language background related to world region of birth. For example, students of Pacific Island background are often invisible in terms of language learning needs, because Pacific Island peoples who migrate to Australia via New Zealand are able to settle in Australia as New Zealand citizens without a specific visa (Amit, Borowski and DellaPergola, 2011;

Cuthill and Scull, 2011). Hence, the capacity to identify this group is problematic, and despite often having English as a second or additional language, this group has had limited access to ESL services (Cuthill and Scull, 2011). The Queensland Department of Education, Training and Employment now recognises students of Pacific Island background as one of the various groups of learners who may have ESL learning needs (DETE, 2013). The inclusion of world regions enables identification of the Pacific island cohort, and further, makes it possible to determine whether source country and all that the source country represents in terms of socio-education opportunities is associated with NAPLaN performance. Students came from all world regions, but predominantly from South East Asia (28%), Sub-Saharan Africa (22%), North Africa and the Middle East (12%) and North East Asia (12%). 8.5% of students came from New Zealand and the Pacific region.

Associated with refugee status, interrupted schooling impacts on the processes of cognitive development, which occur over uninterrupted years of schooling, and on the accumulation of specific learning, including subject specific vocabulary, register and genres (Brown, Miller and Mitchell, 2006). A dichotomous variable which measures students with age appropriate years of education, in comparison to those without, has been included in the year 9 analysis. Worryingly, 28% of these year 9 students had had insufficient schooling for their age.

Table 2 presents statistics in relation to age appropriate education levels. Students whose schooling experience has been incommensurate with their age are achieving average results below those who have experienced nine years of schooling in total. For reading, those with less schooling are on average achieving results below the national minimum standard for year 9.

Table 2. NAPLaN summary statistics by years of schooling, Year 9.

	Reading		
	mean	sd	n
< Age appropriate schooling	465.1	41.0	64
Age appropriate schooling	509.3	55.1	167
Unknown	512.2	79.0	10

Source: ESL Data

The key measure in the sample analyses is that of language proficiency level. This variable is measured using ESL Bandscales, a tool used by ESL teachers to map progress in English language development in the context of the classroom (DETE, 2013). The original National Languages and Literacy Institute of Australia (NLLIA) ESL Bandscales were developed in order to describe second language progress for the purposes of administration of and classroom pedagogy for ESL learners (McKay, Hudson & Sapuppo, 1994). The NLLIA Bandscales and more recent iterations of these (see DETE, 2013) continue to be used by ESL teachers for the purpose of monitoring ESL learner progress in Queensland. The Bandscales provide descriptions of the four macroskills of listening, speaking, reading and writing in primary and secondary aged groups and, within these age groups, enable the learner to be located along a spectrum of proficiency categories ranging from beginner through to advanced (McKay, 1996). Bandscales are usually allocated by an ESL teacher on the basis of a *range* of listening, reading, speaking and writing tasks (McKay et al., 1994; DETE 2013). In this project, the majority of teachers who worked with me in the data collection were highly experienced at allocating ESL Bandscales, suggesting that this data will provide a valid and reliable measure of student language levels.

For the purposes of the analysis I have isolated the reading Bandscale dated closest to the NAPLaN test, and this will be included as a group of dummy variables, representing the range of levels. Bandscale levels 2 and 3 represent students in the beginning stages of English language learning; Bandscale 4 indicates a developing level of English language capability; at level 5, students are consolidating language skills; and at level 6 and 7 they are becoming competent users of academic English.

Table 3 presents the mean scores, standard deviations and number of observations for each of the Bandscale levels included in the analysis. The lower Bandscale levels of 2, 3 and 4 are all achieving average grades which place them no higher than the national minimum standard. In fact, the majority of the group are placed on these Bandscale levels.

Table 3. NAPLaN summary statistics by Bandscale levels, Year 9.

	Reading		
	mean	sd	n
Bandscale 2&3	452.7	32.5	53
Bandscale 4	481.6	36.8	73
Bandscale 5	519.6	42.2	48
Bandscale 6&7	558.12	47.9	16
Unknown	527.7	68.4	51

Source: ESL Data

To conclude the section, I will again revisit the purpose of this dataset in the research project. If it is possible that LBOTE is not a good representation of the impact of language on test performance, is it possible to determine a relationship directly between second language level and NAPLaN result? Is language level associated with test performance?

Year 9 regressions

The regression models are now applied to this year 9 sample, characterised by significant numbers of students of refugee background, recently arrived, and a proportion with reduced years of schooling. The six models are built in such a way that they begin with control or background variables only, and then ESL Bandscales are introduced as the key measure of language proficiency. The remaining models see the inclusion of more indirect measures of language and education background.

The regression models are presented in table 4 followed by a discussion of the results. For those unfamiliar with reading and interpreting such statistical output, explanation of the following components will assist. First, adjusted R squared is a measure, given as a percentage, which shows the extent to which the explanatory variables are contributing to the variation in the NAPLaN reading scores. Secondly, the bulk of the variables in this analysis are categorical and are included in groups, as dummy variables. Reading and interpreting the correlation co-efficients (numerical output) should be done in relation to the reference variable in the group. For example, parent education (P1 Education) has parent with less than year 12 schooling as reference category. In Model 1, students whose parents had 12 years of schooling were, on average, achieving NAPLaN results 20.9 points above those whose parents had less than 12 years of schooling. Finally, statistical significance is indicated by the use of asterisks: a variable will be significantly associated with the NAPLaN results, if the data associated with that variable *contradict* the assumption that there is *no relationship* between that variable and the NAPLaN scores.

Table 4. Regression models 1 to 6 for reading, Year 9.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	485.1***	454.4***	470.7***	466.3***	476.5***	465.8***
Gender (ref: Male)						
Female	-5.4	-6.0	-5.7	-5.6	-4.9	-4.9
P1 Education (ref: < Yr 12)						
Yr 12	20.9*	9.1	5.3	5.6	1.0	-0.2
unknown	-6.8	-7.8	-8.9	-5.3	-6.9	-7.7
English Grades (ref: D/E)						
C	27.2*	18.8*	21.1*	22.0*	21.2*	20.2*
B	40.2***	28.0**	28.5**	29.9**	28.2**	28.7**
A	86.0***	62.1***	58.2***	64.1***	58.0***	59.4***
unknown	22.0	16.4	17.9	23.3*	23.1*	24.4*
NAPLaN LBOTE (ref: no)						
Yes	-21.5*	-11.7	-11.3	-7.0	-7.0	-5.6
Visa Group (ref: Aust/NZ)						
refugee				-24.3*	-46.7*	-46.9*
family				-17.2	-39.7*	-42.2*
business				-8.0	-26.2	-30.4
skilled				10.1	-10.5	-11.6
education				2.2	-15.8	-20.6
Birth Region (ref: Aust)						
Europe			7.3		19.4	23.8
Americas			-1.2		20.0	25.0
MidE/Nth Africa			-20.1		11.1	13.8
Sub-Sah Africa			-15.3		18.9	25.6
N E Asia			-2.5		14.2	16.9
Sthn & Cen Asia			-8.9		13.6	17.0
S E Asia			-20.6		7.2	11.8
NZ & Pacific			-32.1*		-28.1	-27.9
Years of Education (ref: < age appropriate)						
age appropriate						11.7
unknown						25.7
Reading Bandscales (ref: Bandscale 2&3)						
Bandscale 4		21.1*	21.0*	17.9*	22.0*	21.7*
Bandscale 5		52.4***	49.3***	45.8***	47.3***	45.0***
Bandscale 6&7		90.9***	85.8***	76.6***	73.0***	67.4***
unknown		62.2***	61.9***	48.1***	52.0***	50.9***
Adjusted R squared	0.17	0.38	0.39	0.39	0.40 (0.4014)	0.40 (0.4047)

Note: p* < .05; **p < .01; ***p < .001. Ref = reference category. n=240
Source: ESL Data

Model 1 explains 17% of the NAPLaN reading result, controlling for gender, parent education, English A to E grades, and LBOTE status. By model 6, the explanatory power of the model has more than doubled to 40%, as language and language proxy variables are built into the model. I will now explore this process in more detail.

Model 2 sees the introduction of the reading Bandscales and the effect of this is a reduction in effect sizes for A to E grades and parent education. To expand on this, students who are achieving A grades are advantaged by approximately 86 points over D and E grade students in Model 1, but this effect size is reduced in Model 2 to some 62 points. In Model 2, students who are at Bandscale levels 6 and 7 have an advantage of some 90 NAPLaN points above those who are at Bandscale levels 2 and 3. This initial reduction in effect of A to E grades is stabilised for remaining models, despite the introduction of the remaining language proxy variables.

In model 3 we see the introduction of world regions; in model 4, visa groups are included without world regions; and in model 5, both are included together. Finally in model 6, years of education are included in the regression. The introduction of these additional controls sees a slight reduction in the explanatory power of the Bandscales, though there is little variation between models 5 and 6.

For all models which analyse reading performance for this sample group, Bandscales remain significant ($p<0.001$). In comparison, none of the other language variables are statistically significant. For year 9 reading, the majority of the explanatory power therefore is coming from the inclusion of the reading Bandscales, and this is maintained, even when controlling for LBOTE status, visa, world region of birth, and years of education.

Generalised findings and discussion

To begin this section I want to revisit the key findings from second language acquisition research which informed this analysis. These findings identified: a relationship between level of education in first language and time required to achieve academic proficiency in second language; age of arrival as significant to capacity to 'catch up' to English speaking peers; low socio-economic status impacting negatively on rate of acquisition; and greater learning needs of students from refugee backgrounds with limited education. To address these findings I included in my analysis the following: years of schooling; source regions of the world and visa category (as socio-educational variables which help to capture the differing educational background experiences of the students) and English language proficiency level at the time of the test (unique to this project). I have presented the analysis for Year 9 students who, as young adolescents, have the least time but encounter increasingly complex academic language in their schooling. The findings of this analysis add important knowledge to the existing understandings about the performance of second language learners on standardised English tests in a number of ways.

World region of birth reflected variation in NAPLaN results, with world regions representing the global north (Europe/Americas, NE Asia and Australia) achieving above average results, and birth countries in the global South (Sub-Saharan Africa, SE Asia, North Africa and the Middle East) performing below the average. Of interest is the performance of students from the Pacific and New Zealand who are performing below the average. These students are not always identified as language learners by the Australian school system and do not always have access to ESL support, because New Zealand, as their primary source country, is considered an English speaking country. Their NAPLaN attainment, similar to students of refugee background and from countries with limited education services and therefore significantly disadvantaged, should be of concern to education systems.

Performance associated with visa category shows that students on skilled, business and education visas are generally performing above average and students on refugee and family visas are achieving below average results. This is not surprising given that the requirements related to language proficiency in English are quite different for entry to Australia as a skilled migrant in comparison to the refugee stream (Chiswick & Miller 2006). Further, the refugee population has been the subject of considerable research related to educational history, resettlement and attainment, and which I have reported elsewhere (Creagh 2013).

Descriptive statistics and the multiple regression models clearly indicate that NAPLaN attainment is associated with school achievement in A to E grades for English. In terms of Bandscale levels, language proficiency level is the most powerful predictor, along with A to E grade, of NAPLaN performance. The implications of this are important. Firstly, both A to E grades and Bandscale levels are allocated on the basis of teacher professional judgement. The alignment of these results with the NAPLaN test results suggests that teacher judgement is a sound and reliable indicator of learning outcomes. Further, test performance is clearly aligned with language *level*. For those students who are at Bandscale level 4 and below, the NAPLaN test is not a test of literacy, but a test of language and the results for these students are rendered invalid.

Given that I have established empirically that there is an association between language level and NAPLaN performance, there are important implications related to the reliability and validity of a test which assumes English as first language and up to nine years of uninterrupted schooling in Australia. There is considerable literature available which explores general issues of reliability and validity within large scale standardized testing (Koretz, 2008; Wiliam, 2001) and within the Australian context (see, for example, Wu, 2009, 2010). My focus here is particularly of relevance to ESL learners.

Reliability refers to consistency of measurement on the test, affected by variables related to the performance of the student, test conditions, marker consistency, length of test and, most significantly, by question choice (Koretz, 2008; Wiliam, 2001; Wu, 2009). The reading test has, for example, between 40 and 50 questions. From one test to the next, depending on the content covered in the limited range of test items, a student's score may fluctuate, depending on how well the test items align or do not align with the student's knowledge and understandings. There is an added complexity for language learners relevant to reliability. Whilst all students may or may not be familiar with the limited scope of test questions, this will largely be influenced by their literacy skills, if English is their first language. For language learners, there is an added dimension of difficulty directly related to their knowledge of the question field (or topic) and its associated vocabulary. The reliability of results for these students will be impacted by both their literacy *and* language skills.

Validity refers to an inference or conclusion which can be drawn from a test score, or from test data (Koretz, 2008, p.217). Conclusions that test data are valid mean that the test is measuring what it states it is measuring. Conversely, validity is undermined if a test fails to measure what it says it is measuring, or is measuring something else. For example, in the domain of reading, the limited number of test items (between 40 and 50) is insufficient to adequately measure the scope of the domain. Koretz (2008, p.220) refers to this as a problem of construct underrepresentation; insufficient questions mean that the domain of reading is under sampled, and some important knowledge is *not* included. ESL learners may encounter test items which are aligned with the English language knowledge they have encountered at school, or they may not, because of the enormous scope of adequately representing the domain of reading and the limitations placed on this by the test size. This is a 'hit and miss' factor which is particularly significant for second language learners. If the test items do not align with English knowledge, the test becomes one of language rather than literacy. The risk for ESL students who are in the process of learning English, and who are tested in English, is that their results will be confounded by their English language proficiency, rendering their test results invalid (Chalhoub-Deville & Deville, 2008; Genesee, Lindholm-Leary, Saunders & Christian, 2006; Lacelle-Peterson & Rivera, 1994;). Finally, I have already noted the problems associated with identification as LBOTE. The significant numbers of ESL students in my study who were *not* identified in the LBOTE category adds another dimension to the lack of validity and reliability of this data category.

My analysis shows clearly that the association between NAPLaN attainment and language background is not represented by the LBOTE category, and that the complexities of language background, educational experiences and socio-economic factors impact in integrated ways which are difficult to isolate and measure. Categories like visa and country of birth are insufficient to truly understand the performance outcomes of students, and this information is not always readily available to teachers, particularly for those students who are long term residents of Australia. NAPLaN performance is potentially indicative of a range of influential forces related to prior *quality* and *extent* of education and exposure to English. If this is so, these suggest deep causal mechanisms for disadvantaged students, which are not able to be *quickly* remedied by schools and teachers. Nor can schools and teachers be allocated responsibility for student NAPLaN output, which may well be the result of education experiences and education systems in other locations in the world. Rather, schools and teachers need greater knowledge and support in addressing these gaps in educational experiences.

Conclusion

This paper has reported the statistical analysis of the relationship between NAPLaN reading performance for an LBOTE sample of year 9 learners, controlling for a range of language, educational and socio-economic status variables. The analysis has been informed by research from the field of applied linguistics which suggests that second language learners take a number of years to reach parity with their English speaking peers and that this length of time is influenced by factors which relate to age, prior educational opportunities and socio-economic factors. These studies did not have the capacity to identify the language proficiency level of the student. In creating my model I have incorporated measures which account for years of schooling, socio-educational characteristics captured by world region background and visa category, *and* language proficiency level at the time of the test. The project gives a 'snapshot' of learner characteristics and the relationships between these and NAPLaN performance; it clearly identifies the significance of language proficiency level in accounting for NAPLaN performance, currently made invisible by the broad and simplistic LBOTE category.

The project is the first of its kind to make an association between a theoretically informed language proficiency scale (see Hudson, 2012) and NAPLaN performance. The scale has been in use in Queensland state schools for a number of years and is a source of knowledge concerning the progress of language development and related allocation of resources for ESL learners. It is of significance that the scale is educationally embedded in student learning, and is employed by teachers who have specialist knowledge about academic language development. In this sense, use of the Bandscales is akin to qualitative observation of, and documentation of, ESL learner development, and as such, represents valuable knowledge for research.

Importantly, the statistical methods I have applied in my research are less powerful, and potentially invalid if applied to very small numbers of cases. The power of statistical analysis rests in its application to a sizeable group of students, beyond the size of a single class or school. Whilst Bandscale data constitute 'shared understandings' between ESL teachers, they are less valued by education systems and have to date not been used for large scale analysis for educative purposes, although in some systems they are a guide to funding allocation. If I had not accessed Bandscale data from a number of schools and ESL teachers, it would have been impossible to complete this research project. If Bandscale data are not valued *systemically*, there is greater danger that the capacity to statistically argue that the language learner requires differentiated response and support will be lost. Further, it is important that these data are also valid and reliable. For this to be possible, education systems need to ensure that teachers (ESL and mainstream) are provided appropriate professional development in the use of ESL Bandscales, and that moderation across schools and groups of teachers is an ongoing process. The professional knowledge clearly already exists, as demonstrated by my analysis, but needs to be maintained by education departments.

When language as a factor impacting on NAPLaN is effectively silenced by a poor statistical category, and the students who sit NAPLaN are presumed homogenous in terms of English language capacity, how is NAPLaN underperformance by ESL learners to be remedied? Current response is that it can only be remedied by English (as first language) literacy. In contrast, Bandscale allocation is founded in theoretically informed understandings of the ways in which a second language develops in the academic setting. This is powerful knowledge which in turns informs the pedagogical choices made by the ESL teacher in the classroom. The two processes: assessment and pedagogy complement each other and in turn, support the ESL learner in reaching their full school potential.

Acknowledgements

I would like to thank the Queensland Department of Education teachers and principals for their support of this research in their schools.

References

Amit, K., Borowski, A., & DellaPergola, S. (2011). Demography - trends and composition. In A. Markus & M. Semyonov (Eds.), *Immigration and Nation Building: Australia and Israel Compared.* Cheltenham: Edward Elgar Publishing Limited.

Australian Curriculum Assessment and Reporting Authority. (2009). *National Assessment Program: Literacy and Numeracy Achievement in Reading, Writing, Language Conventions and Numeracy.* Retrieved 26 June 2014 from http://www.naplan.edu.au/verve/_resources/NAPLAN_2009_National_Report.pdf.

Australian Curriculum Assessment and Reporting Authority. (2011a). NAP National Assessment Program: Standards. Retrieved 26 June 2014 from http://www.nap.edu.au/Test_Results/How_to_interpret/Standards/index.html

Australian Curriculum Assessment and Reporting Authority. (2011b). National Assessment program: Statements of Learning. 2011, Retrieved 26 June from http://www.nap.edu.au/naplan/statements-of-learning.html

Ball, S. J. (2006). *Education Policy and Social Class.* London: Routledge.

Brown, J., Miller, J., & Mitchell, J. (2006). Interrupted schooling and the acquisition of literacy: Experiences of Sudanese refugees in Victorian secondary schools. *Australian Journal of Language and Literacy, 29*(2), 150-162.

Chalhoub-Deville, M., & Deville, C. (2008). National Mandated Testing for Accountability: English Language Learners in the US. In B. Spolsky & F. M. Hult (Eds.), *The Handbook of Educational Linguistics* (pp. 510-522). Malden, USA: Blackwell Publishing.

Chiswick, B. R., & Miller, P. W. (2006). Language Skills and Immigrant Adjustment: the Role of Immigration Policy. In D. A. Cobb-Clark & S.-E. Khoo (Eds.), *Public Policy and Immigrant Settlement.* Cheltenham, UK: Edward Elgar.

Comber, B. (2012). Mandated literacy assessment and the reorganisation of teachers' work: federal policy, local effects. *Critical Studies in*

Education, 53(2), 119-136. DOI: 10.1080/17508487.2012.672331

Creagh, S. (2013a). 'Language Background Other Than English': a problem NAPLaN test category for Australian students of refugee background. *Race Ethnicity and Education.* Published on line 16 Dec. DOI: 10.1080/13613324.2013.843521

Cummins, J. (1981). *Bilingualism and Minority Children.* Ontario: Ontario Institute for Studies in Education.

Cummins, J. (2000). *Language, Power and Pedagogy: Bilingual Children in the Crossfire.* Clevedon: Multilingual Matters Ltd.

Curriculum Corporation. (2005). *Statements of Learning for English.* Carlton South: Curriculum Corporation. Retrieved 06 November 2014 from http://www.curriculum.edu.au/verve/_resources/SOL_English_Copyright_update2008_file.pdf

Cuthill, M., & Scull, S. (2011). Going to university: Pacific Island migrant perspectives. *Australian Universities' Review, 53*(1), 5-13.

de Vaus, D. (2014). *Surveys in Social Research 6th Edition.* Crows Nest: Allen & Unwin.

Department of Education Training and Employment. (2013). *An introductory guide to the EQ Bandscales for English as an additional language or dialect (EAL/D) learners.* Queensland: State of Queensland (Department of Education Training and Employment) Retrieved 26 June 2014 from https://www.eqi.com.au/programs/bandscales.html

Garcia, G. N. (2000). *Lessons from Research: What is the Length of Time It Takes Limited English Proficient Students to Acquire English and Succeed in an All-English Classroom?* Washington: National Clearinghouse for Bilingual Education and minority Languages Affairs.

Genesee, F., Lindholm-Leary, K., Saunders, W. M., & Christian, D. (2006). *Educating English Language Learners: A Synthesis of Research Evidence.* Cambridge: Cambridge University Press.

Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How Long Does It Take English Learners to Attain Proficiency?* Santa Barbara: University of California Linguistic Minority Research Institute.

Hudson, C. E. (2012). *Teachers Write Theory: The Case of the NLLIA ESL Bandscales.* (Unpublished doctoral thesis), University of Queensland, St Lucia.

Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us.* Cambridge: Harvard University Press.

Lacelle-Peterson, M. W., & Rivera, C. (1994). Is It Real for All Kids? A Framework for Equitable Assessment Policies for English Language Learners. *Harvard Educational Review, 64*(1), 55-75.

McKay, P. (1996). ESL learners: How do we know them? *English in*

Australia, 115(March), 13-23.

McKay, P., Hudson, C., & Sapuppo, M. (1994). NLLIA ESL Bandscales. In P. McKay (Ed.), *ESL Development: Language and Literacy in Schools.* Canberra: National Language and Literacy Institute of Australia.

Miller, J., & Windle, J. (2010). Second Language Literacy: Putting High Needs ESL Learners in the Frame. *English in Australia, 45*(3), 31-40.

Polesel, J., Rice, S., & Dulfer, N. (2013). The impact of high-stakes testing on curriculum and pedagogy: a teacher perspective from Australia. *Journal of Education Policy.* Published online 16 Dec 2013. DOI: 10.1080/02680939.2013.865082

The Senate Education and Employment References Committee. (2014). *Effectiveness of the National Assessment program - Literacy and Numeracy Final Report.* Commonwealth of Australia Retrieved 26 June from http://www.aph.gov.au/~/media/Committees/Senate/committee/eet_ctte/naplan_2013/report/report.pdf.

Thomas, W. P., & Collier, V. (1997). *School Effectiveness for Language Minority Students.* Washington: National Clearinghouse for Bilingual Education.

Watkins, M., Lean, G., Noble, G., & Dunn, K. (2013). *Rethinking Multiculturalism, Reassessing Multicultural Education Project Report 1: Surveying NSW Public School Teachers.* Penrith South: University of Western Sydney.

Wiliam, D. (2001). Reliability, validity, and all that jazz. *Education 3-13: International Journal of Primary, Elementary and Early Years Education, 29*(3), 17-21.

Wu, M. (2009). Interpreting NAPLAN Results for the Layperson. Retrieved 26 June 2014 from http://www.edmeasurement.com.au/_publications/margaret/NAPLAN_for_lay_person.pdf

Wu, M. (2010). Measurement, Sampling, and Equating Errors in Large-Scale Assessments. *Educational Measurement: Issues and Practice, 29*(4), 15-27.

Yang, K. (2010). *Making Sense of Statistical Methods in Social Research.* London: Sage Publications Ltd.

Appendix A – NAPLAN reading, 2010 and 2011 (ESL Data)

	Reading		
	Mean	SD	N
Female	497.35	58.16	140
Male	498.13	53.64	101
Parent Education			
>12 years	489.00	49.12	43
Parent Education Year 12	515.25	57.60	81
Parent education unknown	488.70	55.21	117
English grades			
A	558.57	76.61	14
B	508.57	62.54	68
C	495.92	45.98	77
D/E	468.26	35.00	34
Unknown	488.12	51.47	48
LBOTE	491.97	53.82	195
Not identified as LBOTE	522.47	60.70	45
Visa status			
Refugee	472.59	42.34	100
Family	486.55	42.24	31
Business	508.92	58.53	37
Skilled	547.38	56.01	13
Education	525.87	56.60	23
Australia/NZ	528.57	61.98	37
Birth region			
Australia	534.87	59.27	16
Europe	552.17	53.83	6
Americas	525.62	46.88	8
Nth Africa & Middle East	485.86	40.65	29
Sub Saharan Africa	476.60	47.76	50
NE Asia	516.41	53.50	29
Sthn & Central Asia	531.80	93.16	15
SE Asia	487.72	49.47	67
NZ & Pacific	491.14	48.23	21
Years of education			
Age appropriate	509.3	55.1	167
< age appropriate	465.1	41.0	64
Bandscales			
Bandscale 2&3	452.7	32.5	53
Bandscale 4	481.6	36.8	73
Bandscale 5	519.6	42.2	48
Bandscale 6 & 7	558.12	47.9	16
Unknown	527.7	68.4	51

Sue Creagh is a post doctoral research fellow and lecturer in TESOL education at the University of Queensland in Brisbane, Australia. Sue completed her PhD at the University of Queensland in 2013. Her research interests include TESOL in the Australian school context, education policy, educational outcomes and disadvantage. screagh@uq.edu.au