NSW GOVERNMENT

NSW trial of the reliability and validity of the EAL/D (English as an Additional Language/Dialect) Learning Progression

Statistics Unit, Centre for Education Statistics and Evaluation

and Multicultural Programs Unit

NSW Department of Education and Communities

January 2013

Publication and contact details

Centre for Education Statistics and Evaluation 2015, *NSW trial of the reliability and validity of the EAL/D (English as an Additional Language/Dialect) Learning Progression*, Report prepared by L Lu & K Rickard. This report is available at: http://www.cese.nsw.gov.au/publications/reports-and-presentations

Statistics Unit, Centre for Educational Statistics and Evaluation and the Multicultural Programs Unit gratefully acknowledge the work and input of schools and school staff who participated in this trial project. Teachers undertook the significant extra work and coordination required to participate in the trial with enthusiasm, professionalism and generosity and provided valuable and insightful feedback during workshops and in the survey following the trial.

Acknowledgements

Statistics Unit, Centre for Educational Statistics and Evaluation and the Multicultural Programs Unit gratefully acknowledge the work and input of schools and school staff who participated in this trial project. Teachers undertook the significant extra work and coordination required to participate in the trial with enthusiasm, professionalism and generosity and provided valuable and insightful feedback during workshops and in the survey following the trial.

Executive Summary

Aims and research questions

In 2012, in NSW government schools, approximately 230,000 students were from language backgrounds other than English (LBOTE). These students made up around 30% of total enrolments. Over 136,000 students (18%) were learning English as an additional language or dialect (EAL/D).

The relationship between language background and educational disadvantage has been analysed for more than 30 years. Given limited resources, policy makers seek to target resources effectively to overcome specific aspects of disadvantage, such as language proficiency, to improve both the quality and equality of education outcomes.

Since 30% of students in NSW government schools are LBOTE but only 18% require additional English language support, it is evident that LBOTE does not by itself indicate educational disadvantage or support needs. Various measures have been developed over the last 10 years for diagnostic purposes to identify suitable students for specialist programs. Other measures focus on resource allocation and contribute to school funding formulas that take account of the varying needs of schools.

In 2011 ACARA developed the *English as an Additional Language or Dialect (EAL/D) Learning Progression* to support the implementation of the Australian Curriculum. NSW Department of Education and Communities (DEC) carried out a trial of the *EAL/D Learning Progression* in government schools in May and June 2012. The aim was to investigate whether the *EAL/D Learning Progression* (the instrument) was sufficiently valid and reliable for teachers to use to assess English language proficiency of EAL/D students, primarily as a broad resource allocation mechanism.

The trial considered three research questions:

1. *Can teachers with a diversity of experiences and expertise in ESL education, assess each of the four language modes consistently using the EAL/D Learning Progression, across a broad range of EAL/D students?*
2. *Is there sufficient evidence to support the intended interpretations and uses of teachers' EAL/D Learning Progression phase assessments?*
3. *What are the successful elements and useful resources identified by teachers from the trial process?*

Methodology

The trial included 97 teachers, both specialist ESL teachers and classroom teachers with diverse ESL (English as a Second Language) teaching and assessment backgrounds. The teachers came from 56 schools, including primary, secondary, and central schools as well as

Intensive English centres. Most of these schools were from metropolitan regions, with a few from provincial areas.

A stratified sample of 944 students, across key target grades (Kindergarten, Years 3, 5, 7 and 9), gender groups, sub-demographic groups (i.e., Aboriginal, international student, refugee), and representing the range of English proficiency levels based on the current NSW ESL phase assessment tool, was selected by participating teachers for inclusion in the trial.

A detailed program of professional learning prepared teachers to take part in the trial. Workshops and supporting documentation showed teachers how to identify suitable students, and collect work samples, observations and assessments for the trial.

Each student was assessed on four language modes – listening, speaking, reading and writing. Each language mode had four phases of proficiency – beginning, emerging, developing or consolidating. Teachers submitted assessment results of the four language modes for every student to a purpose-built website.

The trial used a double-marking process to investigate the reliability of assessments using the *EAL/D Learning Progression*: 639 of the sample of 944 students were assessed by two teachers on every language mode (listening, speaking, reading and writing) using the EAL/D instrument. Teachers' assessments for each student were then compared. Various types of inter-rater statistics, such as exact and adjacent agreement rates, correlations, Kappa rates and Dependability Index (score reliability coefficient) were examined.

The validity of the *EAL/D Learning Progression* for the purpose of determining resource allocation in NSW government schools was analysed in relation to four aspects of score validity – concurrent, discriminant, structural and measurement.

Following the trial, teachers were surveyed to assess the usefulness of the instrument and the quality of support provided throughout the trial.

Results

Results are reported against each of the three research questions.

1. *Can teachers assess each of the four language modes consistently using the EAL/D Learning Progression?*

The trial results showed that, on the whole, the *EAL/D Learning Progression* enabled teachers to make consistent judgements of English language proficiency across the four language modes (listening, speaking, reading and writing). There was also strong evidence that teachers were able to use the *EAL/D Learning Progression* to consistently discriminate between the four phases within each mode and between each of the modes.

However, some variations in the consistency of teacher judgements were also observed. For example, teacher judgements were less consistent with some student cohorts (e.g., boys

and students of relatively higher English language proficiency) and in speaking and listening modes. Teacher judgements were more consistent for other student groups, such as those at Intensive English centres. Reasons for variations could be attributable to factors including but not limited to: differences in teachers' prior experience in similar assessment tasks, in particular in assessing informal student interactions which dominate the speaking and listening mode indicators of the EAL/D instrument; teacher knowledge of students; professional learning and training received; and time available to collect and assess work samples.

2. *Is there sufficient evidence to support the intended interpretations and uses of teachers' EAL/D Learning Progression phase assessments?*

The trial has collected sufficient evidence to support the claim that the *EAL/D Learning Progression* provides a balanced and accurate reflection of English language development. The instrument can be used to derive a single measure of English language proficiency for each EAL/D student, for the purpose of allocating ESL funding to schools.

The trial found that there was a reasonable level of congruence between teachers' assessments using the EAL/D instrument and those based on a similar construct – the current NSW ESL Phase tool. The EAL/D instrument was identified as being a more refined and discriminating tool than the current ESL Phase tool and allowed better discrimination between students at the low end of the English proficiency continuum.

The trial analysed the relationship between EAL/D assessments and NAPLAN results for each matched student. As expected, higher NAPLAN results on reading or writing were generally associated with higher EAL/D reading or writing phases, for the same students assessed.

The clear evidence of the **concordance** between the EAL/D assessments and those from other similar or related constructs strengthens the argument that the EAL/D instrument is functioning as intended.

There was also strong **measurement** evidence from the trial that the four modes were measuring a single underlying proficiency and that the four EAL/D phases (beginning, emerging, developing and consolidating) were being used meaningfully and consistently by the teachers. This means that EAL/D assessments over all four modes can be summarised to derive a single measure of English language proficiency for each student.

The analysis of the **structural** patterns of teacher judgements showed that some EAL/D students were at different levels for academic aspects of language (reading and writing) than conversational aspects of the language (listening and speaking). These findings corroborate earlier studies on the differential pattern of development across different modes for some EAL/D students. This further supports the validity of the assessments using the EAL/D instrument.

3. *What are the successful elements and useful resources identified by teachers from the trial process?*

Teachers generally supported the use of the *EAL/D Learning Progression* instrument for resource allocation in place of the current ESL Phase tool. Teachers also indicated considerable interest in using the instrument for informing mainstream program planning.

Survey results indicated that the professional learning provided was adequate to support confident and consistent judgements using the *EAL/D Learning Progression*. However, teachers required substantially more time than expected to plan and complete student assessments to make informed judgements on the appropriate phase. The teachers also indicated that they would like more sample assessment tasks and annotated work samples to support effective *EAL/D Learning Progression* implementation.

Teachers provided detailed feedback regarding the layout and wording used in the *EAL/D Learning Progression.* This feedback may be useful for ACARA and the ACARA English Language Proficiency Working Group for future enhancements to the *EAL/D Learning Progression* and its supporting packages.

Conclusions and recommendations

In conclusion, the NSW trial of the *EAL/D Learning Progression* has provided sufficient reliability and validity evidence for the instrument to be used in NSW government schools as a broad measure of English language proficiency for resource allocation. In addition, the NSW trial has national implications in terms of informing the design and the associated cost-benefit analysis of a prospective national trial. Such a national trial would provide recommendations about the potential use of the progression to report English language proficiency across jurisdictions.

The following recommendations were made based on evidence, feedback and insight from the NSW trial.

Recommendations for NSW government schools

1 Implementation for resource allocation

1.1 Implementation of the *EAL/D Learning Progression* is recommended as a replacement for the ESL Phase tool currently used in NSW government schools. If adopted, the *EAL/D Learning Progression* will become the broad measure of English language proficiency used for EAL/D students, and will become the basis of the allocation formula used for the ESL funding component of the new Resource Allocation Model in NSW government schools.

1.2 Full implementation of the *EAL/D Learning Progression* is recommended for 2014 so that data can be collected and used in the resource allocation process for 2015.

1.3 An implementation plan is recommended that builds on the learning from the trial, and includes:

a) developing and conducting a program of professional learning in government schools during 2013 to prepare teachers to use the *EAL/D Learning Progression* phase assessments for all EAL/D students from the beginning of 2014

b) ensuring professional learning programs emphasise the importance of judging students' phases based on ongoing student assessment

c) developing resources to support teacher assessment using the *EAL/D Learning Progression* instrument, including samples of assessment tasks and annotated work samples demonstrating the evidence required to judge students against the *EAL/D Learning Progression* indicators

d) providing participants in professional learning programs with clear guidelines as to the purpose and potential uses of the *EAL/D Learning Progression*

e) providing participants in professional learning programs with clear guidelines as to the process of deriving an overall phase judgement from the four language mode phase judgements for each EAL/D student

f) building capacity in the new student administration and learning management system (SALM) to allow teachers to enter EAL/D mode and phase data. Teachers should be able to update at any time the EAL/D phase judgements for individual students in SALM by 2014

g) revising school data collection tools to collect *EAL/D Learning Progression* phases in each mode for every EAL/D student. This would replace the current ESL phase data collection tool.

2 Classroom support

The *EAL/D Learning Progression* should be mapped against the new NSW syllabuses and the literacy continuum. On the basis of this mapping, teaching resources should be developed where appropriate to support classroom teachers to program and plan for EAL/D learners.

National recommendations

3 National trial of the *EAL/D Learning Progression*

A national trial is recommended to test the generalisability of the evidence from the NSW trial. This national trial should include:

- a broader range of student and teacher demographic groups, including larger cohorts of specific student groups, in particular Aboriginal students
- examination of differential reliability in teachers' judgements across different types of schools, students and language modes to inform future programs of teacher training and professional development
- recommendations about the potential use of the *EAL/D Learning Progression* to report English language proficiency across jurisdictions.

4 Detailed information to ACARA

It is recommended that ACARA is provided with this report of the NSW trial and with access to the detailed feedback provided by teachers related to the instrument itself. This may be useful to ACARA as authors of current and possible future versions of the *EAL/D Learning Progression*.

Table of Contents

Executive Summary 2

Table of Contents 8

List of Tables 10

List of Figures 10

List of Appendices 11

1. Introduction and research questions 12

1.1 Background and context of the trial 12

1.2 Research questions 14

1.3 Trial limitations 15

2. Trial design and data collections 16

Stage 1: Establishing sample cohort parameters 16

Stage 2: Training and support for participating teachers 18

Stage 3: Student selection and ESL phase data entry 18

Stage 4: Data collection 18

Stage 5: Teacher survey 19

Stage 6: Data analysis 19

3. Descriptions of trial participants 21

3.1 Characteristics of participating teachers 21

3.2 Types of schools included in the trial 21

3.3 Types of students included in the trial sample 21

4. Data analysis and results 23

4.1 Reliability of teacher judgements using the *EAL/D Learning Progression* 23

4.2 Construct validity 30

4.3 Teacher feedback survey 52

5. Further discussions of the trial results 58

5.1 Reliability of teachers' judgements using *EAL/D Learning Progression* 58

5.2 Construct validity of the *EAL/D Learning Progression* 59

5.3 Implications of the trial results on the EAL/D implementation in NSW 61

6. Conclusions 62

6.1 Can teachers assess each of the four language modes consistently using the *EAL/D Learning Progression*? Does this consistency vary across different types of students? 62

6.2 Is there evidence of construct validity to support the intended interpretation of the assessed levels of the EAL/D students? 63

6.3 Considering the importance of training on teachers' approach to using the *EAL/D Learning Progression*, what types of training, how much training, and what additional support should be provided to teachers prior to the use of the instrument? 64

7. Summary of Recommendations 62

Final note 65

References 68

Appendices 71

List of Tables

Table 1 Background of the teachers 21
Table 2 Demographic statistics of the students included in the sample 22
Table 3 Exact agreement rates across modes and groups of students 24
Table 4 Adjacent agreement rates across modes and student groups 26
Table 5 Kappa rates across modes for selected student groups 27
Table 6 Dependability Index for teachers' judgements using the *EAL/D Learning Progression* 29
Table 7 Inter-correlations between two teachers' judgements on the same or different modes 30
Table 8 Distribution of standardised residuals and occurrences of unexpected responses 37
Table 9 Fit statistics and estimated difficulty of each mode, for each scholastic year cohort 39
Table 10 Average ability measures and Rasch-Andrich thresholds 43

List of Figures

Figure 1 A schematic representation of the trial stages 20
Figure 2 Relationship between NAPLAN writing results and EAL/D writing phases 32
Figure 3 Relationship between NAPLAN reading results and EAL/D reading phases 32
Figure 4 Proportions of students at each EAL/D writing phase who are either below, at or above the National Minimum Standards 33
Figure 5 Proportions of students at each EAL/D reading phase who are either below, at or above the National Minimum Standards 34
Figure 6 Proportions of students at each ESL phase who were either below, at or above the National Minimum Standards - Writing 35
Figure 7 Proportions of students at each ESL phase who were either below, at or above the National Minimum Standards - Reading 35
Figure 8 A functional EAL/D rating scale with ordered categories 37
Figure 9 ICC for reading from the Kindergarten dataset 40
Figure 10 ICC for Speaking from the Year 7 dataset 41
Figure 11 Category Probability Curve for the listening mode, from the Year 5 dataset 44
Figure 12 Factor loading of the second factor in EAL/D assessment data – primary students 46
Figure 13 Factor loading of the second factor in EAL/D assessment data – secondary students 47
Figure 14 Locations of students on the EAL/D Proression continuum with their ESL phases 49
Figure 15 Average locations of the Phase 1, 2 and 3 students – primary student data 50
Figure 16 Average locations of the Phase 1, 2 and 3 students – secondary student data 50
Figure 17 Locations of ability estimates of students at, below or above NMS - primary 51
Figure 18 Locations of ability estimates of students at, below or above NMS - secondary 52
Figure 19 Confidence in using the EAL/D Learning Progression in relation to confidence using the ESL Scales 53
Figure 20 Teacher feedback on the *EAL/D Learning Progression* trial process 55
Figure 21 Popularity of different types of support material for future use of the *EAL/D Learning Progression* 56

List of Appendices

Appendix A ESL LEARNER PHASES 71
Appendix B Samples of training and support material 73
Appendix C Student selection screen 76
Appendix D *EAL/D Learning Progression* Assessment Record 77
Appendix E *EAL/D Learning Progression* assessment data collection web screen 78
Appendix F Annotated work sample used for double marking 79
Appendix G Online teacher survey 81
Appendix H Agreement rates (based on five EAL/D levels) for reading and writing 85
Appendix I Histograms of the differences in the matched pairs of owner and non-owner teacher judgements 86
Appendix J Demographic statistics of the students included in the double-marking 87
Appendix K Counts of students in each EAL/D phase for each grade cohort 88

1. Introduction and research questions

1.1 Background and context of the trial

In 2012 in NSW government schools, approximately 230,000 students (30% of total enrolments) are from language backgrounds other than English (LBOTE). The LBOTE cohort is a heterogeneous group made up of students from a broad range of social and linguistic backgrounds. Over 136,000 (18%) of these students are learning English as an additional language or dialect (EAL/D) and require additional support to assist them to develop proficiency in English (hereunder referred to as EAL/D students).[1]

Analysis of NAPLAN data in NSW shows the overall LBOTE cohort as consistently achieving on par with, if not better than, the non LBOTE cohort[2]. However, within the LBOTE cohort, English language proficiency has been identified as an important source of educational disadvantage experienced by some EAL/D students. Analysis conducted by the NSW Department of Education and Communities (DEC) reveals that, after controlling for the effects of students' parental background and school-level effects, one of the most disadvantaged groups of students by far are newly arrived, refugee students with limited English proficiency[3].

Based on the evidence available at the state level, the new resource allocation model for NSW government schools, which is being developed as part of the educational reform*, Local Schools Local Decisions,* has identified a separate loading for EAL/D students requiring English language support.

Currently the NSW DEC allocates ESL (English as a Second Language) resources to schools based on each school's reported level of ESL need, assessed each year using the NSW ESL Phase tool. The ESL Phase tool provides a brief description (Appendix A) of three broad phases of language progression applying to learners across all stages of schooling. Phase 1 students are beginners in English and require intensive level of specialist ESL support. Phase 2 students have developing skills in English and require ongoing explicit English language instruction across the curriculum. Phase 3 students are consolidating their English language skills and require subject specific contingent language scaffolding. Lack of discrete descriptors for different aspects or modes of proficiency (such as listening, speaking, reading and writing) restricts its capacity to accurately reflect the diversity of EAL/D learners who may have variable performance across modes. In addition, the current phase descriptors do

[1] In NSW, the previous term used to describe this group of students is ESL (English as a Second Language) students. Throughout this report, the term ESL is retained whenever it is a part of the name of an existing assessment instrument, document or process.

[2] For example, refer to the *NAPLAN National Report for 2011* published at the following site for comparisons of results between LBOTE and non LBOTE students: http://www.nap.edu.au/verve/_resources/NAPLAN_2011_National_Report.pdf

[3] Refer to the Final *Report for the Review of the Funding for Schooling* (p. 117 – 119) published at the following site for more detailed information about the NSW analysis of factors influencing on the educational achievements of sub-populations of LBOTE students: http://foi.deewr.gov.au/node/30439/

not differentiate students with limited literacy proficiency in their first language. Known issues in the reliability and accuracy of teachers' judgements using the phase tool resulted in the need to incorporate a time factor into each phase description in order to moderate teacher judgement of English language progression. This meant that teachers' judgements using this tool did not always appropriately reflect the level of the students' English language proficiency.

Nationally, the *Final Report for the Review of the Funding for Schooling* released by the Australian Government in December 2011 also recommended basing funding for EAL/D students on their assessed levels of English language proficiency. However, a survey conducted in 2011 of government school systems for all states and territories has shown that no consistent measure is currently available for identifying or reporting English language proficiency of EAL/D students across jurisdictions.[4]

An interim measure developed by Australian Curriculum, Assessment and Reporting Authority (ACARA) to capture the needs of the EAL/D students is the disadvantaged LBOTE variable, which is incorporated in the calculation of the Index of Community Socio-Educational Advantage (ICSEA) (ACARA 2011). This variable is defined as the percentage of the parents in the school community who are both LBOTE and completed a school education of Year 9 equivalent or below. Although intended as a measure to capture the additional educational disadvantage experienced by EAL/D students, analysis conducted by NSW DEC in 2011 demonstrated that this variable was not a reliable proxy for students' ESL needs. DEC modelling revealed that, if used as a funding mechanism in NSW government schools, this measure would not only underestimate, by approximately 50%, the size of the target group of students needing ESL support, but it had a high probability of misdirecting funds to student populations and schools that did not require ESL support.[5]

Motivated by the desire to better identify and discriminate the ESL need at the school level to ensure equitable resource distributions both within the NSW government sector and nationally, DEC conducted a trial of the *EAL/D Learning Progression* (EAL/D) in NSW government schools between May and June 2012. The goal of this trial was to investigate whether the *EAL/D Learning Progression* was sufficiently valid and reliable for use by teachers to assess EAL/D students' English language proficiency, primarily for the purposes of resource allocation.

The *EAL/D Learning Progression* was developed by ACARA in 2011, with input from content experts across jurisdictions and academia. It describes the development of English language typical of students learning English as an additional language or dialect (EAL/D). It includes

[4] At the request of the Australian Education, Early Childhood Development and Youth Affairs Senior Officials Committee (AEEYSOC) and the Ministerial Council for Immigration and Multicultural Affairs (MCIMA), the Schools Data Sub Group conducted the survey during July and August 2011 on funding for ESL/EAL/D student support across States and Territories.
[5] Detailed results demonstrating the extent of the misalignment between EAL/D students' needs and the interim measure of Disadvantaged LBOTE will be contained in a NSW DEC report to the Strategic Policy Working Group in early 2013.

broad descriptions of the characteristics of learner groups at each of four phases of English language learning (Beginning, Emerging, Developing and Consolidating) differentiated by four language modes (listening, speaking, reading and writing) and by three stages of schooling (Kindergarten-Year 2, Years 3-6, Years 7-10). The Beginning phase includes a unique set of descriptors for students with limited literacy proficiency in their first language[6]. As the tool was developed to support teachers as they develop teaching and learning programs for the Australian Curriculum, it has the potential to be implemented nationally.

1.2 Research questions

Three broad research questions were formulated to direct study design, data collection and analysis. Since the trial was to investigate the potential for use of the *EAL/D Learning Progression* primarily as a tool for resource allocation, the first research question was related to reliability – *Can teachers assess each of the four language modes consistently using the EAL/D Learning Progression?* It was important to demonstrate that the tool could be used consistently and reliably by teachers with a diversity of experiences and expertise in ESL education across a broad range of EAL/D students. This question was analysed through inter-rater reliability statistics, based on teacher judgements collected from a double-marking process.

Secondly, as the *EAL/D Learning Progression* (hereunder also referred to as the EAL/D instrument) was trialled as a direct measure of English language proficiency of EAL/D students for the purpose of resource allocation, it was necessary to examine the question of construct validity – *Is there sufficient evidence to support the intended interpretations and uses of teachers' EAL/D Learning Progression phase assessments?* For example, do lower EAL/D phases correspond to lower English language proficiency, thus indicating more support or funding for the student assessed; and vice versa?

As validity is a multi-dimensional concept, this question was addressed through evidence pertinent to the following four aspects of score validity:

- Concurrent validity –
    - Is the relationship between assessments for the same students using the EAL/D instrument and those from other external constructs (e.g., NAPLAN) on a similar trait of language ability as expected?
    - Is the relationship between assessments for the same students using the EAL/D instrument and those made using other theoretically similar constructs (e.g., the NSW ESL Phase tool) as expected?[7]
- Discriminant validity –

[6] Refer to the following website for the general descriptions of the EAL/D *Learning Progression* http://www.acara.edu.au/verve/_resources/English_as_an_Additional_Language_or_Dialect_Teacher_Resource_05_06_12.pdf

[7] This type of evidence is also referred to as the external aspect of construct validity by Messick (1996).

- Can the four language modes of the EAL/D instrument be empirically discriminated by teachers? For example, is there any evidence of teachers' judgements on one mode being confounded by students' assessed performances on other modes?

- Structural aspect of validity –
  - Is the internal structure of the assessments produced by teachers using the EAL/D instrument consistent with the expected interrelations among the different modes of language, as derived from the relevant second language acquisition theory or from prior empirical studies?
- Measurement aspect of validity –
  - Is there evidence demonstrating the measurement quality of the assessments? For example, is there evidence of the four modes measuring a single underlying ability (i.e. the English language proficiency of EAL/D students)? Are the four rating scales used to assess the four modes functioning as intended?

Evidence of the measurement aspect of validity has additional importance to this project. For the purpose of resource allocation a single measure of English language proficiency for each student is desirable. However, before four separate mode assessments can be summarised into a single student assessment score, there must be supporting evidence for the abovementioned measurement questions (Andrich, 1988).

Finally, to support a prospective national implementation of the *EAL/D Learning Progression*, the trial sought to identify successful elements and useful resources from the trial process. An online evaluation survey was used to collect and analyse teacher responses to the instrument. The corresponding research question was – *What is teachers' feedback on the trial process, the tool and its usability*?

In summary there were three research questions for the trial:

1. *Can teachers assess each of the four language modes consistently using the EAL/D Learning Progression?*

2. *Is there sufficient evidence to support the intended interpretations and uses of teachers' EAL/D Learning Progression phase assessments?*

3. *What are the successful elements and useful resources identified by teachers from the trial process?*

1.3 Trial limitations

While the trial included the question of construct validity, detailed examination of the content validity —the appropriateness and relevance of the phase statements and descriptors used in the EAL/D instrument to the target ability that is intended to be measured— was considered out of scope from the beginning. However, the trial did seek

teacher comments on the content and the clarity of the EAL/D phase statements and indicators through the online teacher survey.

Due to the small sample size of students in some demographic groups (e.g., Aboriginal or international students), the generalisability of results for these groups of students will need to be confirmed by a larger trial, preferably by a national trial involving students of more diverse cultural and linguistic background than those of NSW EAL/D students.

A number of factors may have influenced the final inter-rater reliability statistics. In the double marking process, pairs of teachers from the same school judged the *EAL/D Learning Progression* phase of a common group of students. This double marking process was used in preference to a more anonymous collection process because of the logistics of manually collecting and securely redistributing student work samples on all modes, as well as relevant assessment tasks and student information, within a limited timeframe. To minimise potential bias teachers were directed to assess students independently. However, it was difficult to ensure teachers did not confer about phase decisions or student performance particularly where they were working at the same school and in close proximity.

2. Trial design and data collections

This section provides details of the six stages of the trial, which include considerations of the trial design, sampling strategy and data collection processes.

Stage 1: Establishing sample cohort parameters

Various considerations influenced decisions on the numbers of teachers, schools, and students participating in the trial. These included the need for a large enough sample to ensure the validity of inferences drawn from the statistical estimation processes, the required number of training workshops as a result of the number of participating teachers and the mode of training delivery, and the approximate teacher time spent on student selection and assessment, work sample collection and data entry. Ultimately, due to practical time and resource constraints, the trial included around 100 teachers, each assessing 10 students.

Three key principles guided the further selection of teachers, schools and students to ensure the ability of the trial to address the research questions and the generalisability of the trial results.

The first principle was to include teachers with a range of prior ESL teaching and assessment background in order to achieve a representative sample of teachers to maximise the reliability of findings. Review of relevant literature suggests teacher background (e.g.,

teachers' familiarity of similar constructs such as the *ESL scales*[8] used in NSW, prior teaching and assessment experience, perceptions of language proficiency and assumptions of language acquisition) can impact on the way they use a tool such as *EAL/D Learning Progression* (e.g., Eckes, 2008; Erdosy, 2004; Knoch, Read & von Randow, 2007; Leckie & Baird, 2011). Representation of schools from different areas (such as metropolitan and regional locations) and across different types (such as mainstream primary, secondary and central schools and Intensive English Centres) in the sample helped ensure inclusion of teachers with different levels of exposures to ESL professional learning and experience assessing EAL/D student English language proficiencies.

A second principle was to include students from key scholastic grades –Kindergarten, Years 3, 5, 7 and 9. There were a number of reasons for the inclusion of these grades in the sample. First, as the *EAL/D Learning Progression* contained separate descriptors designed for assessing students at different learning stages (i.e., Kindergarten to Year 2, Year 3 to 6 and Year 7 to 10, respectively), the trial needed to include students from all three stages so that validity and reliability evidence for the EAL/D instrument could be investigated across all three stages. Secondly, the sample needed to include Years 3, 5, 7 and 9 so that the relationship between the assessed EAL/D reading and writing phases and the NAPLAN reading and writing scores for the same students could be examined. Kindergarten was included as an additional target grade because of the anecdotal evidence from NSW pointing to the difficulty in assessing the English language proficiency of Kindergarten students. This difficulty was partly due to the current ESL phase tool not including age appropriate descriptors to support teachers to make assessments.

The third principle was to get as wide a spread of demographic sub-groups as possible within the target sample. Efforts were made to ensure that the selection included an approximately equal distribution of students across gender groups, inclusion of students in the sample from diverse demographic backgrounds (i.e., Aboriginal, international students, refugees, students of different cultural and linguistic profiles), and an adequate representation of students from different English proficiency levels based on teachers' assessments using the current ESL phase tool. There was an intention to slightly over sample students from low to medium English proficiency levels (i.e., those at ESL Phases 1 and 2) as this would enable greater focus on whether *EAL/D Learning Progression* could be used consistently and appropriately for discriminating this group of students, who were considered to experience significant educational disadvantage.

[8] ESL scales is an assessment tool used by ESL teachers in NSW government schools to assess, monitor and report on students' learning of English as a second language. They were developed by the Australian Education Council and Australian Council for Educational Research as part of a national curriculum initiative in 1992. They describe the English language development of ESL students and provide a set of benchmarks against which the full range of ESL learners' achievements in English may be set.

Stage 2: Training and support for participating teachers

Stage 2 of the trial involved developing and conducting a program of professional learning to ensure participating teachers had clear and consistent understandings of the trial process and requirements and the *EAL/D Learning Progression.* In the initial training session teachers were introduced to the trial purpose and timeframe, data collection processes and requirements, and familiarised with the *EAL/D Learning Progression* and the data collection website template.

A period of one week between workshops enabled teachers to do a preliminary identification of the possible student cohort for the trial considering the scholastic grade, the range of ESL phases, as well as gender and special needs.

At the second workshop teachers developed consistent practice in using the *EAL/D Learning Progression* to decide on a phase. A number of resources were developed for the trial teachers to assist them in recording assessment task conditions, student background data and *EAL/D Learning Progression* phases. In order to collect inter-rater reliability data (research question 1), teachers were asked to nominate themselves for participation in double-marking. Samples of training and supporting materials developed to support teachers are provided in Appendix B. While the mode of training delivery was mainly face-to-face, a video conference option was available for teachers from regional schools.

Stage 3: Student selection and ESL phase data entry

In order to ensure a balanced representation of students from various demographic groups and across different English proficiency levels, a website was developed to support teachers to select an appropriate range of students for the trial. It showed the background information of all prospective students available for selection in a school, including language background, country of birth, visa subclass, residency status and date of arrival in Australia. Teachers then selected their students (and entered the current ESL phase data for the selected students) through the website, using the sampling principles communicated to them at the workshop. A sample student selection screen is shown at Appendix C.

At the end of the student selection stage, summarised reports were run to show the proportions of students in different groups (e.g., proficiency levels, grades, backgrounds) in the initial sample. Where underrepresentation of some groups was detected, requests were made to selected schools to increase a certain type of students to build a representative sample.

Stage 4: Data collection

This stage of the trial involved teachers collecting observations and work samples for each student and making assessments. Teachers then submitted their EAL/D phase judgements for each student on each language mode on a purpose-built website, and provided annotated work samples as evidence of the range and type of assessment tasks used.

Teachers used an *EAL/D Learning Progression* assessment record proforma (See Appendix D) to record student background information, assessment tasks, and observations to assist in deriving a final phase for each mode to be entered on the website. The EAL/D phase was entered via a lookup list showing four phases – Beginning (with the option to discriminate between Some Print Literacy and Limited Literacy Background), Emerging, Developing or Consolidating. Appendix E shows the website screen where teachers entered students' assessment.

Teachers participating in double marking were required to collect work samples for each of their students and provide assessment task and student background details using the assessment record shown in Appendix D. This information was then forwarded to their partner teacher in the same school who used it to make independent EAL/D phase judgements. An example of an annotated work sample and assessment record is included in Appendix F.

All participating teachers were required to submit to the website one work sample from one student in one mode (listening, speaking, reading or writing) in conjunction with the student and task details. In total over 100 work samples were collected, which provided data about the type and range of assessment practices used in the trial.

Stage 5: Teacher survey

An online teacher survey was developed primarily to collect information to inform a response to the third research question - *What is teachers' feedback on the trial process, the instrument and its usability?* The teacher survey provided data on teacher background information including current teaching position, level of ESL qualification and experience using similar ESL assessment tools. It provided quantitative and qualitative feedback on the *EAL/D Learning Progression* trial, on professional learning and support, on the content and clarity of the *EAL/D Learning Progression* statements and indicators and on its potential usability. A copy of the survey is provided in Appendix G.

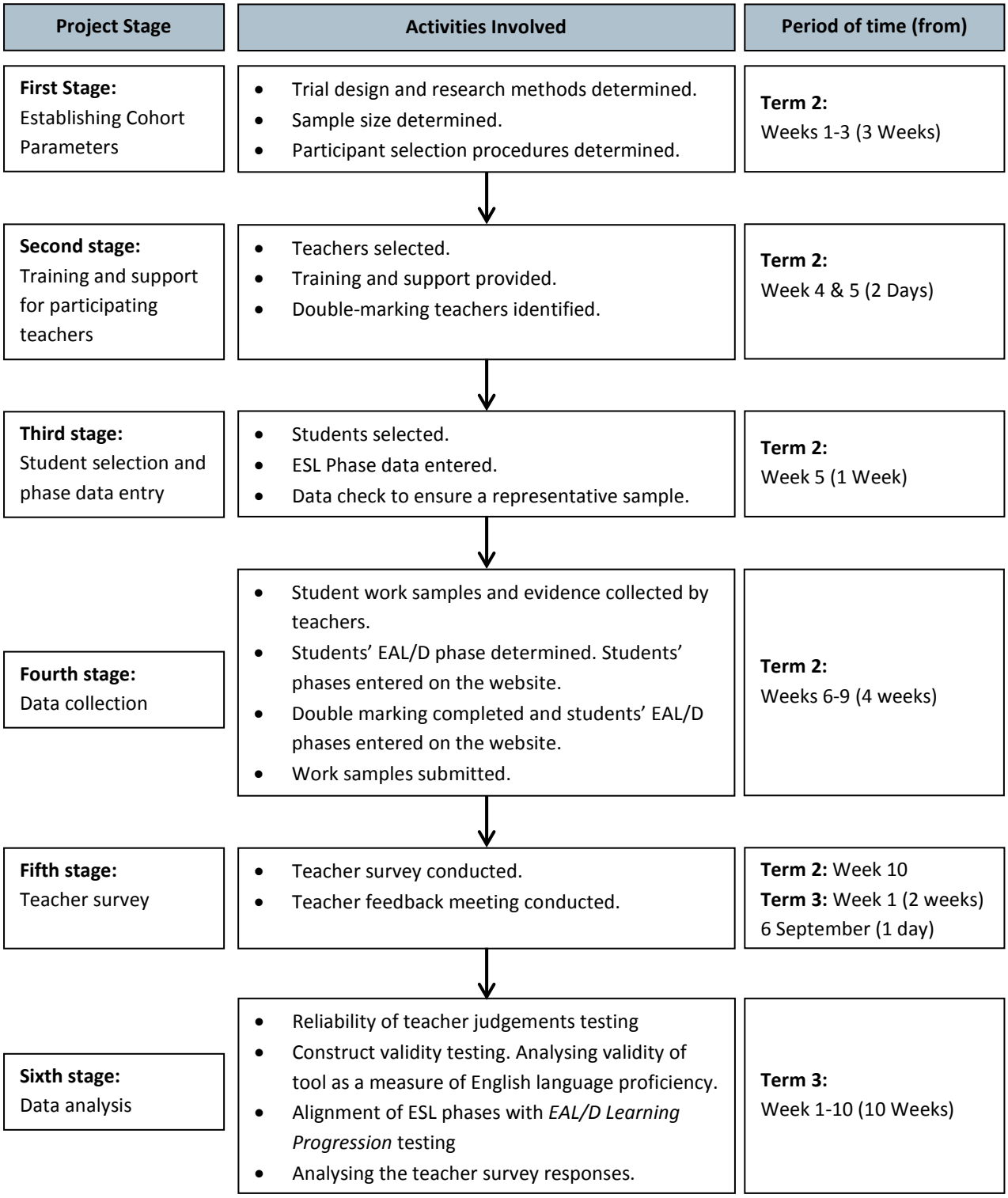
Participating teachers were invited to attend a focus group to provide a further opportunity to discuss the trial process, the potential for using the *EAL/D Learning Progression* in place of the current ESL phases and how the *EAL/D Learning Progression* could be used to derive a single phase level for each student. 29 teachers attended this focus group.

Stage 6: Data analysis

During this stage the data was analysed in relation to the three research questions. Methods and results of the analyses are discussed in detail in Section 4, with further interpretations of the trial findings presented in Section 5.

Figure 1 provides a schematic overview of the six trial stages.

Figure 1 A schematic representation of the trial stages



3. Descriptions of trial participants

3.1 Characteristics of participating teachers

In total, 97 teachers from 56 schools participated in this trial. Table 1 provides information on the teaching, training and qualifications background of the teachers, based on the responses provided by 74 of the 97 teachers to the evaluation survey.

Table 1 Background of the teachers

	Background category	No.	%
All Teachers		74	
Current position in the school	ESL teacher	46	62%
	Class teacher	16	22%
	Executive teaching	9	12%
	Other specialist teacher	2	3%
	Executive non-teaching	1	1%
Training and qualifications	TESOL* or ESL pre-service training	49	66%
	TESOL* or ESL postgraduate qualification	38	51%

Note: Based on the responses provided by 74 of the 97 teachers to the evaluation survey
TESOL: Teachers of English to Speakers of Other Languages

As illustrated in Table 1, around 60% of the participating teachers who responded to the survey were ESL specialist teachers and around one fifth were classroom teachers. Correspondingly, two thirds (66%) of the teachers reported to have had specialist TESOL or ESL pre-service training, and half of them reported to have TESOL or ESL postgraduate qualifications.

Of the total of 97 teachers, 66 participated in the double-marking process, which meant each of them not only marked around 10 of their own students, but also around 10 of another teacher's students.

3.2 Characteristics of schools included in the trial

56 schools were involved in the trial, with an approximate even distribution across primary and secondary contexts. 26 were primary schools, 17 were secondary schools, 1 was a central/combined school (K-12) and 12 were Intensive English Centres affiliated with a high school. The majority of schools (53) were in metropolitan areas, with 3 from regional areas.

3.3 Characteristics of students included in the trial sample

A total of 944 students were included in the trial, of which more of than half (639) were marked by two teachers to estimate inter-rater reliability.

Table 2 provides information on the background of all students in the sample. The demographic characteristics of the students included in the double-marking are broadly similar to those of the whole sample (see Appendix J for the distribution of the students included in the double-marking process across different demographic groups).

Table 2 Demographic statistics of the students included in the sample

		No of students	% of all students
All Students*		944	
Gender	Girls	452	47.9%
	Boys	492	52.1%
ESL Phases	Phase 1	345	36.5%
	Phase 2	369	39.1%
	Phase 3	201	21.3%
Grade	Kindergarten	150	15.9%
	Year 3	161	17.1%
	Year 5	161	17.1%
	Year 7	229	24.3%
	Year 9	233	24.7%
	Other Grades (Year 8, 10, 11)*	10	1.1%
Aboriginal Students		29	3.1%
Intensive English Centre		168	17.8%
Refugee		197	20.9%
International Student		28	3.0%

Note: The total number of students excludes 17 students who were not marked on any of the four language modes.
Though teachers were asked to select students from key target grades (Kindergarten, Years 3, 5, 7 and 9), a very small number of students from other grades were also included in the sample by teachers.

Overall, there were slightly more boys than girls (492 boys -v- 452 girls) included in the trial. The student selection process undertaken by the teachers seemed to work well as the final sample achieved a good balance of students from different English proficiency levels. Around 37% of students were from the low end of the proficiency scale (i.e., ESL Phase 1), 39% of them were Phase 2 students and 21% were Phase 3 students. The distribution reflected the sampling intention to have an over representation of students from the low to medium proficiency levels.

There was also an even distribution of students across primary and secondary levels of education, with 16% of them from Kindergarten, 17% each from Years 3 and 5, and 24% and 25% from Year 7 and Year 9 respectively. Students from Intensive English Centres made up

of 18% of the total sample cohort. Most of these students are newly arrived secondary students who required intensive English support. In addition, 21% of the students were from a refugee background. Due to practical constraints and the demographic characteristics of NSW EAL/D students, only 29 Aboriginal students and 28 International students were able to be included in the trial.

4. Data analysis and results

4.1 Reliability of teacher judgements using the *EAL/D Learning Progression*

Reliability refers to the consistency in teachers' judgments from one assessment scenario to another. It is an important criterion when assessing the suitability of an assessment instrument for high-stakes uses and is often considered to be a pre-requisite of validity (Weigle, 2002).

In this section, the following common types of inter-rater reliability indicators were investigated, based on comparisons of two teachers' judgements for the same student on the same mode, obtained from the double-marking process:

- Exact agreement rate – the proportion of the time the two teachers agreed exactly
- Adjacent agreement rate – the proportion of the time the two teachers agreed within one phase
- Kappa rates – Agreement rates adjusted by chance agreement
- Dependability Index – score reliability coefficient and
- Correlational statistics – inter-correlations between two teachers' judgements (in the next section 4.2.1)

Since agreement rates are sensitive to the number of scored categories used, teacher judgements on four broad EAL/D phases (Beginning, Emerging, Developing and Consolidating) are used in the calculations for each language mode, with the two Beginning categories (Beginning Some Print Literacy, Beginning Limited Literacy Background) collected for the reading and writing modes combined into one Beginning phase. This data treatment ensures that the estimated teacher consistency measures can be compared across the modes.

4.1.1 Exact agreement rates

Table 3 reports the exact agreement rates, calculated for each mode, across all students and for different groups of students separately.

Table 3 Exact agreement rates across modes and groups of students

Student Groups	Exact Agreement			
	Listening	Speaking	Reading	Writing
All Students	80.5%	78.7%	81.8%	82.4%
Girls	81.5%	80.5%	81.5%	83.6%
Boys	79.5%	77.0%	82.2%	81.3%
Aboriginal	100.0%	83.3%	88.9%	100.0%
ESL Phase 1	88.1%	84.4%	90.4%	87.2%
ESL Phase 2	78.7%	75.1%	81.5%	80.3%
ESL Phase 3	69.4%	75.5%	68.5%	76.4%
Kindergarten	84.4%	77.1%	86.3%	84.4%
Year 3	87.4%	72.7%	80.2%	88.2%
Year 5	85.2%	80.0%	80.9%	83.5%
Year 7	78.7%	81.3%	82.0%	82.7%
Year 9	71.5%	80.3%	80.9%	75.9%
Refugee Students	82.1%	78.8%	85.4%	88.1%
International Students	57.1%	71.4%	92.9%	71.4%
Intensive English Centre Students	89.7%	85.0%	96.3%	88.8%

Note: Total number of students included in the double-marking process: 639. Agreement rates are calculated based on teachers' judgements on four broad EAL/D phases for each mode.

Table 3 shows that, on average, 80 per cent of the time two teachers' judgements on a mode of language for the same student match exactly.

While the overall agreement rate appears to be satisfactory, teacher consistency does differ across groups of students, or across the modes. It would seem that teachers can assess reading and writing modes more reliably than they do listening and speaking, with speaking being identified consistently as the mode most difficult to be assessed reliably, across most of the student groups. The data also demonstrates that the comparatively lower reliability associated with assessing speaking is more of a problem for assessing boys than for girls. While teachers achieved broadly similar levels of consistency in their judgements across four modes when assessing girls, the exact agreement rate associated with assessing boys' speaking proficiency was 5.2 and 4.3 percentage points lower than those associated with assessing reading and listening respectively. There is no clear explanation for this difference between boys and girls results. Further analysis of data through a national trial with a larger cohort might shed light on the source of difference and/or confirm the generalisablity of these results.

In addition to the above, there is evidence that teachers were able to assess students of lower English proficiency levels more consistently than students of higher proficiency levels.

For example, the rate of exact match associated with assessing the reading levels of students of the lowest English proficiency level (ESL phase 1 students) was 21.9 percentage points higher than the equivalent rate associated with students of the highest English proficiency level (ESL phase 3 students). The corresponding differences in the agreement rates when assessing other modes were also significant, ranging from 10.8 percentage points for writing to 8.9 percentage points for speaking.

Greater consistency in teachers' judgements when assessing students of limited English proficiency might have also contributed to the higher than average agreement rates observed for assessments concerning refugee students, since refugee students are more likely to be of low English proficiency than the rest of the student population in the sample.

Table 3 also shows that teachers at the Intensive English Centres could assess their students more consistently than those at mainstream schools. Compared to average rates, the agreement rates for the Intensive English centre students were 14.5 percentage points higher on the reading mode, 9.2 percentage points higher on the listening, 6.4 and 6.3 percentage points on the writing and speaking modes respectively. This could also be due to the teacher marking pattern that is already observed above – that is, teachers are much more likely to agree with each other when they are assessing students who have very limited English proficiency.

4.1.2 Adjacent agreement rates

Table 4 shows the average adjacent agreement rates, across all students and separately for different groups of students. These rates indicate the frequencies of the occurrences of significant variations in assessments, which are an important criterion for evaluating the suitability of any instrument for high-stakes uses (such as for resource allocation purposes).

Table 4 Adjacent agreement rates across modes and student groups

Student Groups	Adjacent Agreement			
	Listening	Speaking	Reading	Writing
All Students	97.8%	97.9%	98.4%	99.0%
Girls	98.7%	98.0%	99.0%	100.0%
Boys	97.0%	97.9%	97.9%	98.2%
Aboriginal	100.0%	100.0%	100.0%	100.0%
ESL Phase 1	99.1%	98.6%	99.5%	100.0%
ESL Phase 2	97.2%	98.0%	98.0%	98.4%
ESL Phase 3	96.5%	96.5%	97.2%	98.6%
Kindergarten	97.9%	95.8%	98.9%	97.9%
Year 3	99.1%	98.2%	97.3%	99.1%
Year 5	94.8%	97.4%	96.5%	100.0%
Year 7	98.0%	98.0%	99.3%	98.7%
Year 9	98.7%	99.4%	99.4%	99.4%
Refugee Students	99.3%	98.7%	100.0%	99.3%
International Students	100.0%	100.0%	100.0%	100.0%
Intensive English Centre Students	100.0%	100.0%	100.0%	100.0%

Note: Total number of students included in the double-marking process: 639. The agreement rates are based on teachers' judgements using four broad EAL/D phases.

The table indicates that, on average, the proportion of the time when two teachers' EAL/D phase judgements for the same student on the same mode differing by more than 1 phase varied from 1.0 per cent, when assessing writing, to 2.2 per cent, when assessing listening. The level of significant discrepancies in the assessments seems to be at an acceptable level, given the amount of training provided through the trial. However, further analysis is required to explore whether comparatively higher than average occurrences of significant variations in teachers' judgements for the Kindergarten students' speaking and Year 5 students' listening phases are attributable to random factors or any systematic issues in the teachers' assessment processes.

4.1.3 Agreement rates based on five EAL/D phases

During the data collection stage, for the reading and writing modes, students at the Beginning EAL/D phase were further assessed by teachers in two sub-categories – Beginning with no or limited prior print literacy or Beginning with some prior print literacy, using the descriptors/criteria provided by ACARA. As it is of interest to know whether teachers can consistently discriminate the two Beginning sub-categories using the EAL/D descriptors and criteria, the exact and adjacent agreement rates for reading and writing are re-produced after data were recoded to treat the two subcategories as two separate phases. These rates are provided at Appendix H.

Appendix H demonstrates that, for any group of students with a reasonable sample size (i.e., greater than 100), the agreement rates based on five EAL/D phases were only marginally worse than those based on four phases, as reported in Tables 3 and 4.

4.1.4 Kappa rates

Acknowledging that two teachers can agree by chance alone, Kappa rates (Cohen, 1960), which adjust for chance agreement, were calculated for all students and selected groups of students and are provided in Table 5.

Table 5 Kappa rates across modes for selected student groups

Student Groups	Listening	Speaking	Reading	Writing
All Students	0.71	0.69	0.74	0.74
Kindergarten	0.75	0.64	0.76	0.71
Year 3	0.81	0.58	0.70	0.80
Year 5	0.78	0.68	0.71	0.75
Year 7	0.68	0.72	0.72	0.71
Year 9	0.54	0.70	0.70	0.61

Note: Total number of students included in the double-marking process: 639. The agreement rates are based on teachers' judgements using four EAL/D phases (Beginning, Emerging, Developing and Consolidating).

Kappa is a measure of the difference between the observed agreement and the expected agreement by chance alone, standardised to be on a -1 to +1 scale (Viera & Garrett, 2005, p. 361). Using the criteria suggested by Landis & Koch (1977), the overall agreement rates, which ranged from 0.69 for speaking to 0.74 for reading and writing, suggest that teachers achieve a substantial level of agreement between each other, when using the EAL/D instrument to make an assessment of a student's English language proficiency level.

4.1.5 Differences in owner teacher judgement and non-owner teacher judgement

One question relevant to analysing teacher assessment behaviour using the EAL/D instrument is whether there would be significant differences in judgements made by owner teachers (i.e., those who are assigned the class of students and have greater opportunities for ongoing interaction with the student assessed) and non-owner teachers (those who have limited ongoing interaction with the students and who would rely primarily on collected work samples to make assessments). For example, would owner teachers be systematically more lenient or harsh in their judgements as compared to non-owner teachers? This question has implications on the choice of the EAL/D implementation model, which would also have a flow-on effect on the level and type of professional training that needs to be provided. If there is no evidence of significant difference between the two groups of teachers, the tool may be used by any teachers who could be required to make

phase judgements based on a body of student work samples collected over time across the modes.

To investigate this, non-parametric tests – Wilcoxon signed rank test – were employed to examine the significance of the median difference between the EAL/D phase judgements made by matched pairs of owner and non-owner teachers, for each mode separately.[9]

The tests show that, for each language mode, the median difference in the matched pairs of judgements is not statistically significant. Using Wilcoxon signed rank tests, the smallest *p* value is 0.12 for the reading mode, the largest *p* value is 0.55 for the writing mode, all greater than $\alpha = 0.05$. Appendix I presents visually the distributions of the differences in the matched pairs of teacher EAL/D phase judgements, across the four modes. These graphs demonstrate that, while there are small differences in owner teacher and non-owner teacher judgements, there are no apparent patterns in these differences – that is, there is no evidence of owner teachers tending to assign either a higher or lower EAL/D phase for the same student than the non owner teachers. Having said that, around 60% of participating teachers (both owner and non-owner) were ESL specialists with considerable experience using a similar assessment tool to assess EAL/D students. This would indicate that teachers can use the *EAL/D Learning Progression* consistently to assign a phase regardless of the opportunities for ongoing interaction with the student, so long as they have expertise in using the tool and follow the assessment guidelines using a number of work samples over time to inform their phase judgement.

4.1.6 Generalisability analysis

Extending earlier analysis on the descriptive types of inter-rater agreement indicators, Generalisability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; see also Brennan, 2001; Shavelson & Webb, 1991) is used in this section to estimate the reliability of teachers' judgements using the EAL/D instrument for each mode. Two types of reliability coefficients are available from the G-theory analysis – the Generalisability Coefficient, which is used for decisions concerning relative standing of the persons (norm-referenced testing) and the Dependability Index, which is used for decisions concerning the absolute level of performance (criterion-referenced testing). As the *EAL/D Learning Progression* is most likely to be used to understand the absolute English proficiency level of an EAL/D student, either for resource allocation or program planning purposes, the Dependability Index is likely to be of most interest to users. For this reason, this section reports the Dependability Index (phi).

GENOVA is the computer program (Crick & Brennan, 1983) used. A separate G-study with a single-facet crossed design (*p* x *r*') was employed to analyse the score reliability for each

[9] During the double-marking process, each student was essentially assessed by an owner teacher and a non-owner teacher. The owner teacher collected the work samples and tasks which were then passed on to the non-owner teacher for assessment (also Section 2 about descriptions of the data collection process). The flag of being an owner or non owner teacher for the student assessed was collected at the same time when the teacher submitted the assessment data about the student.

mode. In each analysis, students (*p*) were the objects of the measurement with ratings (*r'*) that each student received for a particular language mode as random facets. Although reliability coefficients can be calculated for different rating scenarios (such as the average scores produced from a double-marking process), this section reports only the reliability of the scores produced from a single-marking process. This is because, if the EAL/D instrument is rolled out to schools for either resource or program planning, it is most likely that only one teacher (either a classroom teacher or an ESL teacher) will be employed to make judgements for one student (i.e., a single-marking scenario).

Table 6 Dependability Index for teachers' judgements using the *EAL/D Learning Progression*

	Listening	Speaking	Reading	Writing
Dependability Index	0.84	0.84	0.86	0.86

Note: The Dependability Index is estimated based on a single-marking scenario

Table 6 shows that the reliability of teachers' judgements across all modes reaches the conventionally desired level of score reliability (i.e., 0.8) for high-stakes tests (Schoonen, 2005; Shavelson & Webb, 1991). The Dependability Index for reading and writing aspects of language proficiency is around 0.86, slightly better than that for the conversational aspects of language (i.e., listening and speaking).

In summary, results in this section have collectively demonstrated that teachers can achieve a desirable level of consistency when using the EAL/D instrument to make judgements on a student's language proficiency levels on the four modes. Any occurrences of significant variations in teacher assessments are shown to be rare. Furthermore, there is no evidence of systematic differences in judgements made by teachers who have a direct teaching relationship with the student being assessed and those who don't. This finding suggests that the use of the EAL/D instrument in a school may not need to be limited to a particular group of teachers, so long as adequate training is provided. Notwithstanding this, evidence from this section has also revealed certain situations where comparatively lower inter-rater reliability was observed, for example, when teachers were assessing the speaking mode, in particular boys' speaking proficiencies. Variations in reliability may result from a number of factors, which are further discussed in Section 5.1.

4.2 Construct validity

4.2.1 Convergent and discriminant validity evidence

To help establish construct validity, this section uses the multitrait-multimethod analysis method proposed by Campbell & Fiske (1959), to investigate the convergent and discriminant validity evidence.

To facilitate the investigation, Table 7 provides a matrix of inter-correlations between two teachers' judgements, on the same mode and on different modes, for the same student.

Table 7 Inter-correlations between two teachers' judgements on the same or different modes

		Teacher 2 judgement			
	MODE	**LISTENING**	**READING**	**SPEAKING**	**WRITING**
Teacher 1 Judgement	**LISTENING**	.85**	.69**	.69**	.68**
	READING	.75**	.87**	.70**	.75**
	SPEAKING	.70**	.70**	.84**	.67**
	WRITING	.68**	.75**	.67**	.87**

Note: **Correlation is significant at the 0.01 level (2-tailed).
Nonparametric correlations were performed (Spearman's rho), as there was evidence rejecting the hypothesis that the eight sets of teacher assessment scores were normally distributed. Although 639 students were included in the double-marking process, 12 of them were not double-marked on all four modes. This results in a sample size varying from 627 to 630 for the correlation coefficients reported.

Convergent validity requires that measures of the same trait of performance, for the same student, using two independent methods (such as two teachers' judgements on the same mode produced independently) show sufficiently large and positive correlations (Campbell & Fiske, 1959, p. 2). This is because the measures are supposed to be assessing the same aspect of the latent proficiency. Evidence that EAL/D phase assessments meet this requirement was clearly demonstrated in section 4.1, as well as through the highlighted diagonal values in Table 7. The table shows that there are strong correlations between two teachers' EAL/D phase judgements on the same mode, ranging from 0.84 for speaking, to 0.87 for reading and writing. This provides convergent evidence necessary to support the validity of the EAL/D construct.

Discriminant validity requires evidence that one teacher's judgement on one mode correlates more highly with another teacher's judgement of the same mode than it does with the alternate teacher's judgements on any other modes, for the same student. This is because different modes are intended to measure different aspects of the language

proficiency, and there needs to be empirical evidence that teachers can use the EAL/D instrument to effectively discriminate the various conceptually distinct traits of the latent proficiency in an appropriate manner (Campbell & Fiske, 1959, p. 83; Kane, 2006, p. 40).

Table 7 shows there is a satisfactory level of discriminant validity evidence for the EAL/D instrument. This can be verified from the table, as each diagonal value is higher than any other values lying in its row or column. For example, Teacher 2's judgements on the listening mode correlate with Teacher 1's judgments on the same mode for the same students at 0.85. This correlation is higher than the correlations they have with Teacher 1's judgements on any other modes, for the same students (0.75, 0.70 and 0.68 for reading, speaking and writing, respectively).

4.2.2 Concurrent validity evidence

This section focuses on investigating the relationship between teachers' judgements on the reading and writing using the EAL/D instrument and the NAPLAN reading and writing scores, for the same students, to collect evidence either supporting or challenging the external aspect of construct validity (Messick, 1996).

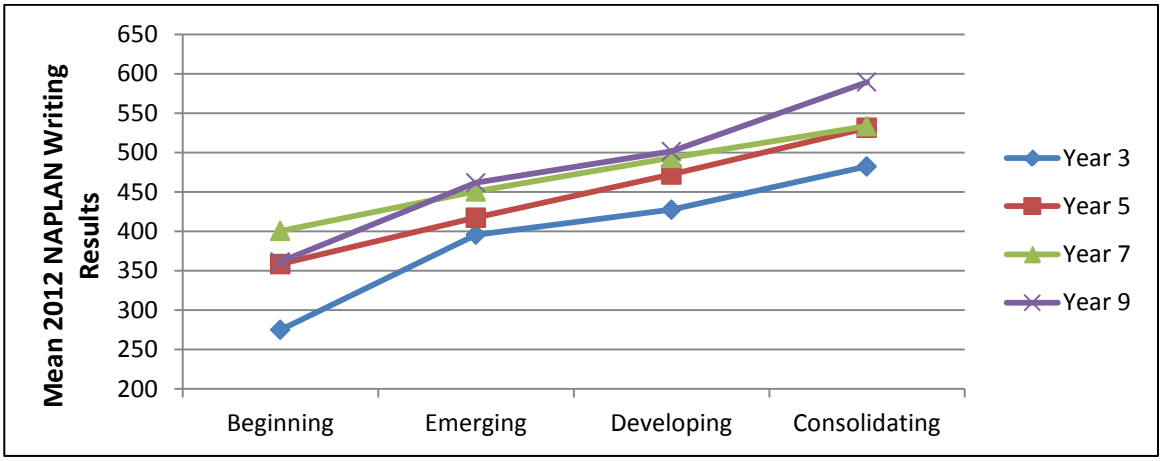
This aspect of validity refers to "the extent that the assessment scores' relationship with other measures and non-assessment behaviours reflect the expected high, low and interactive relations implicit in the specified construct" (Brualdi, 2002, p. 12). By appraising the degree to which empirical relationships are consistent with score meaning, the score interpretation (or the validity of score) is substantiated (Brualdi, 2002; also see Messick, 1996).

It is noted that, while NAPLAN and the EAL/D instrument were designed for different purposes (hence different constructs), the expectation is that assessments from these two instruments on reading or writing aspects of language should exhibit a reasonable level of concordance. The two sets of assessments were undertaken at a similar time (the trial data was collected in May/June, while the NAPLAN tests were administered in May) and both sets of assessments attempted to capture a similar aspect of language proficiency (either reading or writing) for the same students.

For this analysis, 90.4% of Years 3, 5, 7 and 9 students from this trial were able to be matched to the 2012 DEC NAPLAN dataset (i.e., 708 matched students for writing and 706 for reading). Of the students matched, 88 were exempted, 16 were absent and 3 were withdrawn from the NAPLAN reading tests. Similarly, 88 were exempted, 11 were absent and 3 were withdrawn from the NAPLAN writing tests.

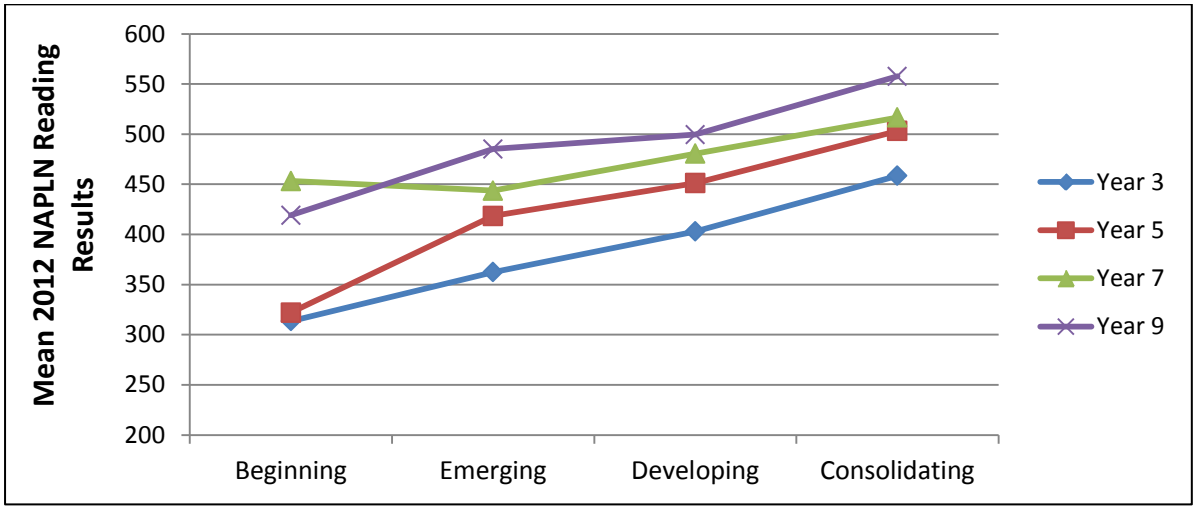
Figures 2 and 3 depict the mean NAPLAN scores of students at each assessed EAL/D phases, for writing and reading separately.

Figure 2 Relationship between NAPLAN writing results and EAL/D writing phases



Note: A total of 606 matched students were included in the analysis. For those students who were double-marked, the EAL/D phases used were those determined by the owner-teachers.

Figure 3 Relationship between NAPLAN reading results and EAL/D reading phases



Note: A total of 599 matched students were included in the analysis. For those students who were double-marked, the EAL/D phases used were those determined by the owner-teachers.

Before the two graphs are interpreted, it needs to be noted that students who were exempted, absent or withdrawn from the NAPLAN tests were not included in the calculation of the average NAPLAN scores, since there were no NAPLAN scores for them. As the majority of these students were exempt students who were most likely assessed at the lowest proficiency level – i.e., the Beginning phase for both reading and writing, the mean NAPLAN score represented an inflated proficiency estimate for those students at the Beginning EAL/D phase. In addition, patterns of relationships demonstrated in Figures 2 and 3 should be interpreted with caution due to the relatively small number of students of certain scholastic grades being assessed at some EAL/D phases, particularly in the Beginning and Consolidating phases.

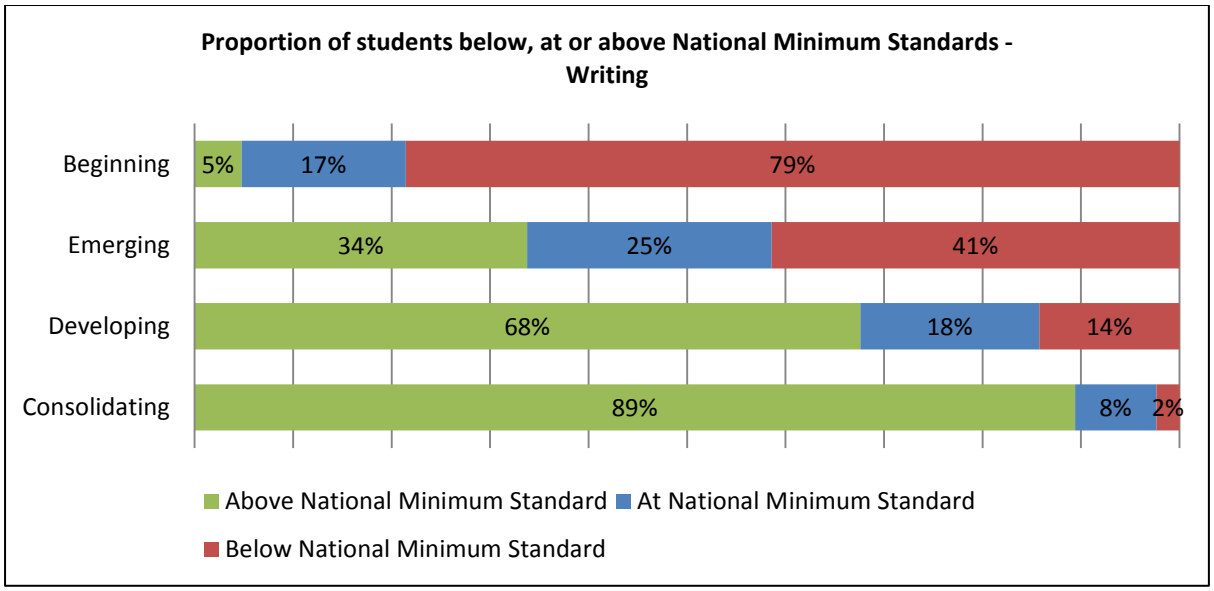
Bearing the limitations in mind, Figure 2 nonetheless shows that the expected relationship between the EAL/D writing phases and NAPLAN writing scores are empirically confirmed. For each grade cohort, higher EAL/D phases are associated with higher mean NAPLAN scores, and lower EAL/D phases with lower mean NAPLAN scores, as expected.

With regard to the EAL/D reading phases (Figure 3), the relationship is as expected for the cohort of Years 3, 5 and 9 students. However, for Year 7, the average NAPLAN score of students at the Emerging phase is slightly lower than that of students at the Beginning phase. This anomaly is most likely due to the biased average NAPLAN score for the Beginning students as a result of the exclusion of exempt students in the analysis.

To correct for this bias, the relationship between the EAL/D phases and NAPLAN was re-examined using NAPLAN results as referenced to the National Minimum Standards. Consistent with national reporting rules, exempt students were coded as having achieved below National Minimum Standards.

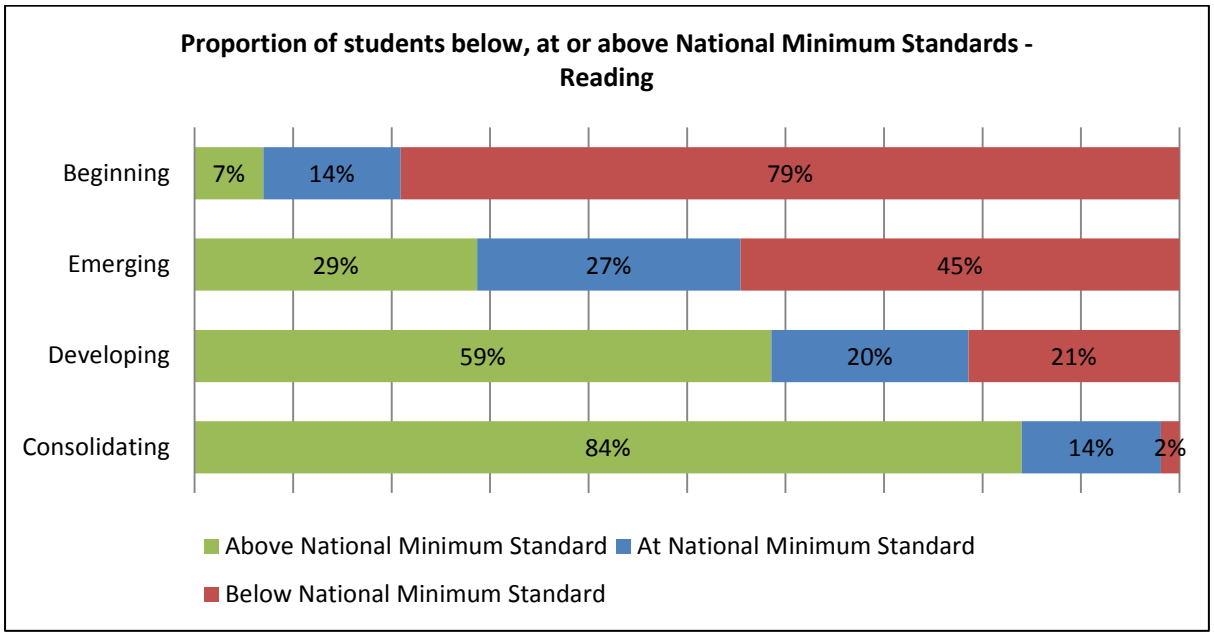
Figures 4 and 5 report the proportions of matched Years 3, 5, 7 and 9 students at each EAL/D phase who are either below, at or above National Minimum Standards, for writing and reading separately.

Figure 4 **Proportions of students at each EAL/D writing phase who are either below, at or above the National Minimum Standards**



Note: Exempt students included, and absent and withdrawn students are excluded from the analysis.

Figure 5 **Proportions of students at each EAL/D reading phase who are either below, at or above the National Minimum Standards**



Note: Exempt students included, and absent and withdrawn students are excluded from the analysis.

Figures 4 and 5 demonstrate that, as expected, students who were assessed at the Beginning EAL/D phase are most likely to achieve below the National Minimum Standards for their grades, and students who were at the Consolidating levels were mostly likely to achieve above the National Minimum Standards. For example, while nearly 80% of the students assessed at the Beginning phase on the writing mode were below the National Minimum Standards in the NAPLAN writing tests, only 14% of the students at Developing and 2% of the students at the Consolidating phase were below the National Minimum Standards.

The extent of concordance between the NAPLAN scores and the EAL/D phases shown in Figures 2 to 5 provides evidence to support the intended interpretations of the four ordered EAL/D phases.

As the same students were also assessed by teachers using the current NSW ESL phase tool, the relationship between the current NSW ESL phases and NAPLAN results were also examined (see Figures 6 and 7) and compared to those demonstrated in Figures 4 and 5.

Figure 6 **Proportions of students at each ESL phase who were either below, at or above the National Minimum Standards - Writing**

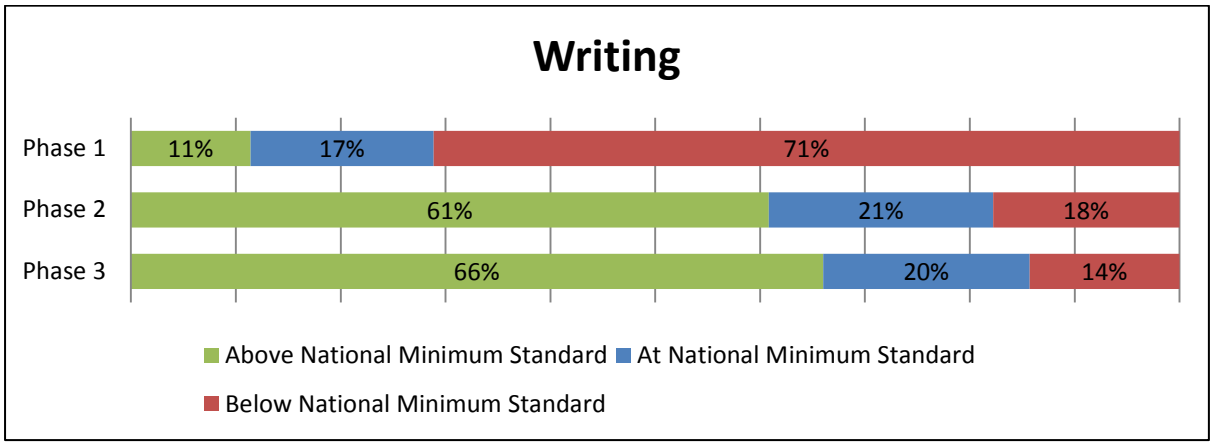
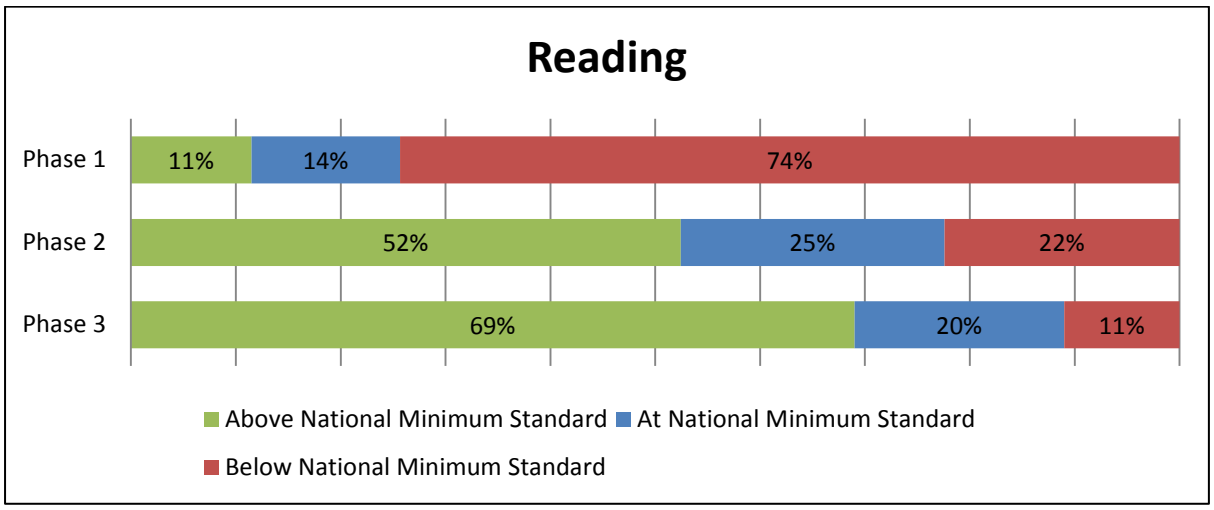


Figure 7 **Proportions of students at each ESL phase who were either below, at or above the National Minimum Standards - Reading**



When comparing Figures 6 and 7 to Figures 4 and 5, it is clear that the EAL/D instrument is more refined and discriminating than the current ESL phase tool, particularly in terms of its ability to capture those students at low English proficiency levels. For example, while 40% to 45% of the students at the second EAL/D level (Emerging) achieved test results below National Minimum Standards, only around one fifth (i.e., 18% to 22%) of the students at the current ESL Phase 2 level were below the National Minimum Standards. This provides evidence to support the argument that, for the purpose of determining ESL funding distribution in NSW government schools, the *EAL/D Learning Progression* is a better instrument than the current ESL Phase tool to identify and discriminate the educational needs of EAL/D students, particularly for those with very limited English proficiency.

4.2.3 Measurement aspect of validity

This section focuses on the following two research questions to examine the measurement qualities of the assessments made by teachers using the EAL/D instrument:

1) *Are the four language modes measuring a single underlying ability?* and
2) *Are the four rating scales used to assess the four modes functioning as intended?*

The *EAL/D Learning Progression* can be conceptualised as consisting of four component scales, each measuring an important aspect of the latent ability – EAL/D student's English language proficiency. As such there is a need to investigate the uni-dimensionality of the EAL/D construct, i.e., whether there is supporting evidence of the scores produced from the four scales measuring one single underlying ability. A modern approach to assess uni-dimensionality involves analysing the data according to a uni-dimensional measurement model in order to determine the extent to which the data conforms to the requirement of the model. A useful model for this type of analysis is the Rasch Model (see Rasch, 1960, 1980, for more detailed descriptions of this model).

The four component scales themselves are rating scales, since students' performances on each of the four modes are judged in ordered phase categories—Beginning, Emerging, Developing, Consolidating— in accordance with pre-defined criteria. An important requirement of a functional scale is that a higher score category on the scale, in general, should imply more of the underlying ability and vice versa. This requirement, referred to as the "inferential property" of a rating scale by Linacre (1999), is consistent with the scale definition and with the intended use of the scale. When this requirement is not met, doubts are cast on the meaning of the scale and on the validity of the measurement outcomes (Eckes, 2009).

This study uses a special form of the Rasch model — Rasch Partial Credit Model (Wright & Masters, 1982) to investigate the above-mentioned research questions. The Partial Credit Model incorporates the concept of order within a framework of uni-dimensionality. As visually demonstrated in Figure 8, the model conceptualises a functional rating scale as dividing the latent ability continuum into ordered categories, which qualitatively advance along this continuum (Linacre, 2010).

If there is sufficient evidence of the four modes measuring a single ability, and of the rating categories on the scales being used meaningfully and as intended, the EAL/D assessments across the four modes can be summarised to a single score as an indicator of the student's overall English language proficiency level. This has important implications for the utility of the EAL/D instrument as it is desirable to have one single assessment for each EAL/D student for the purpose of resource allocation.

Figure 8 **A functional EAL/D rating scale with ordered categories**

Latent Ability Continuum – increasing English proficiency

In this trial, the Winsteps computer program (Linacre, 2010) was used to provide the psychometric analysis of the EAL/D mode assessments using the Partial Credit Model. As the same EAL/D phase may correspond to a different ability level for students of different scholastic years, the Rasch analysis is conducted for each grade cohort (Kindergarten, Years 3, 5, 7 and 9) separately. [10] Another reason for conducting the analysis separately for each grade cohort is that the EAL/D instrument contains different descriptors and criteria for assessing students at three different learning stages (i.e., K-2, 3-6 and 7-10). Consequently, it is necessary to check that the psychometric qualities of the EAL/D assessments hold across student populations at different stages of learning.

A range of statistical and graphical tests of fit indicators are inspected in the following sections, in order to reach an overall conclusion of the degree of fit between data and the uni-dimensional model and the effectiveness of the rating scales used to assess the modes.

4.2.3.1 Global model fit

Distribution of standardised residuals (Table 8) is first examined to inspect the data-model fit at the global level. If data fits the Rasch model sufficiently well, the standardised residuals should be close to a normal distribution (i.e., N(0,1)) (Linacre, 2010). Satisfactory model fit may also be indicated when about 5% or fewer of all the responses (i.e., EAL/D phase assessments) have (absolute) standardised residuals ≥ 2 and about 1% or fewer have (absolute) standardised residuals ≥ 3 (Linacre, 2008).

Table 8 **Distribution of standardised residuals and occurrences of unexpected responses**

		Kindergarten	Year 3	Year 5	Year 7	Year 9
Standardised Residuals	**mean**	0.01	0.01	0.01	0.00	0.00
	standard deviation	1.05	1.07	0.94	0.96	1.01
% of the total observations having (absolute) standardised residuals ≥ 2		4.5%	4.1%	3.3%	4.5%	4.4%
% of the total observations having (absolute) standardised residuals ≥ 3		2.2%	2.5%	1.4%	2.7%	1.1%

[10] Where students were assessed by two teachers, only the owner teachers' judgements were used in the Rasch analysis.

Table 8 shows that, for each grade cohort data, the mean and standard deviation of the standardised residuals are very close to the expected values of 0 and 1. Furthermore, the percentages of the unexpected EAL/D mode assessments associated with absolute standardised residuals ≥ 2 are contained within the recommended limit for reasonable model fit (i.e., 5%).

With regard to the mode assessments that are most unexpected by the Rasch model (i.e., those that have absolute standardised residuals ≥ 3), the proportion of these assessments exceeds the recommended limit for all five datasets. Across all grades, the Year 7 data has the highest percentage of assessments deemed as the most unexpected by the Rasch model (2.7%).

The above results indicate that, while there is a satisfactory level of fit between data and model on a global level, there is also a larger than expected number of responses being identified as deviating substantially from model expectations. The latter could be attributable to few students having very atypical or uneven score profiles across the four modes (i.e., they have developed very different levels of proficiency across different language modes), which is not uncommon for EAL/D students, particularly those with disrupted education and limited literacy in their first language (e.g., Collier, 1995; Garcia, 2000). However, it could also be due to phase categories not being used consistently by teachers, thus introducing noise to the data that is unrelated to the underlying ability being measured. To investigate the genesis of misfit, the next section examines fit statistics at the mode and phase category level.

4.2.3.2 Fit at the mode level

Two types of fit statistics can be used to examine item level fit – INFIT and OUTFIT mean squares (Wright & Stone, 1999). Both are chi-square ratios based on the standardised residuals. While the OUTFIT statistic is an unweighted statistic which is heavily influenced by outlying, off-target, unexpected responses, the INFIT is sensitive to irregular inlying patterns with relatively more impact being given to unexpected responses close to a person's or item's measure (Wright & Masters, 1982; Wright & Stone, 1999). Both mean-square statistics have an expected value of 1.0, and a range from 0 to positive infinity. Values less than 1.0 indicate over-fit; that is, data is too predictable with respect to model expectations, causing summary statistics such as reliability indices, to report inflated results. Values greater than 1.0 indicate under fit; that is, there is more un-modelled noise in the data than expected. High mean-squares are considered a much greater threat to the validity than low mean-square values, because they suggest other sources of variance in the data, which means that summarised scores across four modes are likely to provide a distorted pictured of the data.

Mean-square values are also reported in various standardised forms, such as the INFIT and OUTFIT z–standardised t–statistics reported by the Winsteps Rasch computer program. The

statistical convention is that when the absolute value of a standardised t–statistic is greater than 2 (i.e., $p<0.05$), the null hypothesis that the data fits the Rasch model (perfectly), should be rejected.

For this trial, a desirable range defined by a lower-control limit of 0.5 and a higher-control limit of 1.5 (Linacre, 2010) is used to examine the fit at the mode level.

Table 9 reports the locations of the modes on the underlying ability continuum (i.e., the difficulty of the modes), the fit mean square values and the associated standardised values, for each grade cohort data separately. Within the table for each grade cohort, the modes are presented in the ascending order of their difficulty measures (i.e., from easiest to the most difficult).

Table 9 Fit statistics and estimated difficulty of each mode, for each scholastic year cohort

	NAME	MEASURE	ERROR	IN.MSQ	IN.ZSTD	OUT.MSQ	OUT.ZSTD	Difficulty
Kindergarten	SPEAKING	-0.49	0.24	0.89	-0.76	1.09	0.50	Easiest
	READING	-0.45	0.27	1.11	0.75	1.55	1.43	
	LISTENING	-0.12	0.28	0.90	-0.51	1.00	0.10	
	WRITING	1.06	0.29	0.75	-1.49	0.81	-0.46	Most Difficult
Year 3	SPEAKING	-1.06	0.27	1.03	0.23	0.94	-0.08	Easiest
	LISTENING	-0.91	0.26	1.05	0.38	1.17	0.63	
	READING	-0.32	0.26	0.74	-1.91	1.31	0.75	
	WRITING	2.29	0.28	0.85	-0.75	1.18	0.58	Most Difficult
Year 5	LISTENING	-1.69	0.24	0.84	-1.29	0.71	-0.84	Easiest
	SPEAKING	-1.61	0.25	0.81	-1.62	0.87	-0.21	
	WRITING	-0.37	0.25	1.07	0.55	1.02	0.18	
	READING	3.68	0.25	1.16	1.17	0.93	0.00	Most Difficult
Year 7	SPEAKING	-1.27	0.18	1.04	0.40	1.17	0.96	Easiest
	LISTENING	-0.75	0.18	1.04	0.4	0.79	-1.18	
	READING	0.79	0.18	0.65	-3.71	0.56	-3.12	
	WRITING	1.23	0.19	1.1	0.94	1.19	1.11	Most Difficult
Year 9	LISTENING	-0.98	0.18	0.84	-1.66	1.21	0.64	Easiest
	SPEAKING	-0.9	0.18	1.05	0.51	1.18	0.69	
	WRITING	0.76	0.18	1.14	1.52	0.96	0.03	
	READING	1.11	0.18	0.83	-1.84	0.77	-0.98	Most Difficult

Notation: IN.MSQ – INFIT mean square; IN.ZSTD – z-Standardised t-statistics for INFIT mean square value; OUT.MSQ – OUTFIT mean square; OUT.ZSTD – z-Standardised t-statistics for OUTFIT mean square value

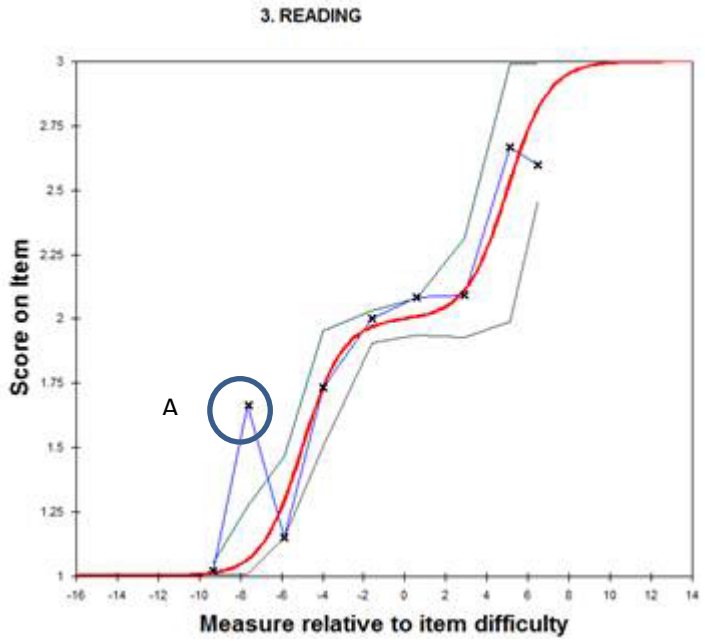
Attention is first paid to any large mean square values as they indicate under-fit of the data to the model. Table 9 shows that, except for one statistic (OUTFIT mean square value of 1.55 for the reading mode estimated from the Kindergarten data), all OUTFIT and INFIT mean-square values are below the upper-control limit of 1.5, for all of the modes across all grades. With regard to the one exception, its respective standardised t-statistics is less than 2, indicating that the under-fit in this case is not statistically significant.

All INFIT and OUTFIT mean square values are also above the lower control limit of 0.5, suggesting no cause for concern regarding model overfit. If there was evidence of overfit, it would have indicated a lack of local independence – that is, the modes were not working independently of each other.

4.2.3.3 Item characteristic curves

Item Characteristics Curves (ICC) are next examined for all modes, for each grade dataset separately (total no. of ICCs: 4 modes x 5 grades = 20). The ICC graph can provide clues as to which person ability group on the ability continuum did not contribute to the development of useful measurements (e.g., which group has the greatest difference between actual and expected scores). The first ICC inspected is the one for the reading mode, obtained from the Kindergarten dataset (Figure 9), since the previous section suggested that segments of the assessments made on the reading mode for this grade may not support useful measurement.

Figure 9 ICC for reading from the Kindergarten dataset



Note: a score of 1 corresponds to the Beginning phase, 2 to Emerging, 3 to Developing, and 4 corresponds to Consolidating.

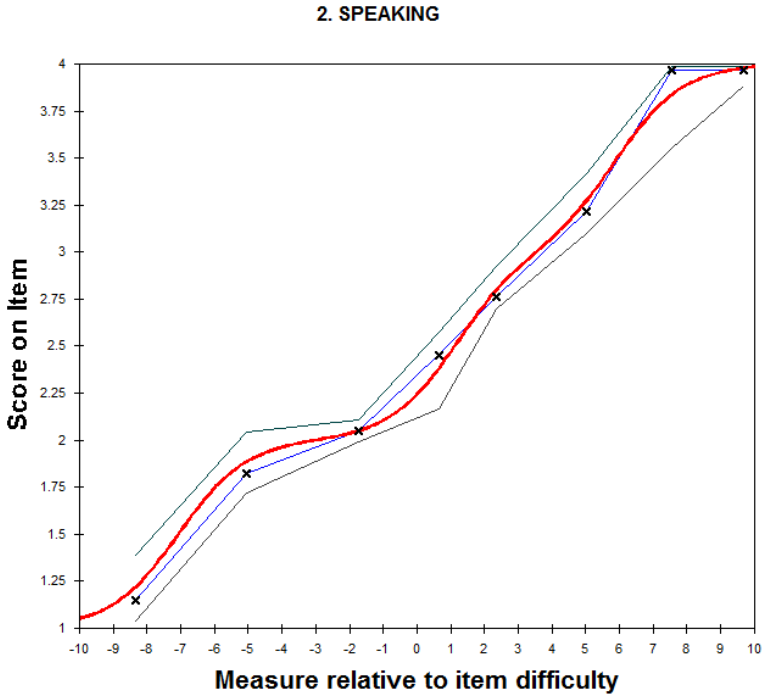
If data conforms to the model requirements, a lower ability group (each group represented by a cross on the graph), on average, should receive a lower score (i.e., a lower phase category) on the mode, than an adjacent higher ability group.

While the majority of the groups behaved in an expected manner (i.e., they are contained within the 95% confidence bands, indicated by the two green/grey lines), one anomalous group is identified on the figure (circled and labelled as A). For the students included in this group, the average score they received on reading was significantly higher than the model predicts, given the low average ability estimate of these students.

Although the misfit could be entirely due to chance because of the small number of students included in this group, assessment records of those students included in this group were nonetheless inspected to check if there was any evidence of teachers' EAL/D judgements not appropriately reflecting the reading ability being measured. The inspection identified four Kindergarten students in this group that contributed to the observed misfit. All four (three of them were girls) were assessed at Emerging phase on reading, but Beginning on all other modes – speaking, listening and writing. Two of these students were assessed by the same teacher, who provided additional comments from her observations of the students. Based on the teacher comments received for three of the four students, it would appear that the misfit originated from these students displaying disparate levels of proficiencies across different modes.

All other ICCs, an example of which is included below, show no evidence of significant misfit.

Figure 10 ICC for Speaking from the Year 7 dataset



Note: a score of 1 corresponds to the Beginning phase, 2 to Emerging, 3 to Developing, and 4 corresponds to Consolidating.

4.2.3.4 Effectiveness of the EAL/D rating scales

The expectation of a functional rating scale is that in general, students with higher overall proficiency produce observed results in higher EAL/D phase categories, and vice versa. One useful way to check whether the empirical data adheres to this expectation is to examine the average ability measure estimated for each phase. The average ability measure is defined as the average of the Rasch ability estimates for all students who were rated in this particular phase category for this mode.

Table 10 reports the average ability measure for each EAL/D phase as well as the Rasch-Andrich thresholds, across the four modes, for each grade dataset separately.

It is noted that, for each grade, across all modes, the average person ability measures advance with the sequential categories on all rating scales, which is consistent with the intention that a higher category on a rating scale should indicate more of the underlying ability.

Another indicator of the effectiveness of a rating scale is the ordering of category thresholds, which can be inspected visually through the Category Probability Curves (CPC). These graphs simplify inferences about which category is most likely to be observed at any point along the ability continuum by visually presenting the category boundaries (Linacre, 2010). It is noted that, in this study, numbers of students in some EAL/D phase categories for some scholastic years can be small (i.e., less than 10). In these cases, the estimates of category boundaries and Rasch-Andrich thresholds should be interpreted with caution. Appendix K provides the counts of students in each EAL/D phase for each of the grade cohort included in the sample.

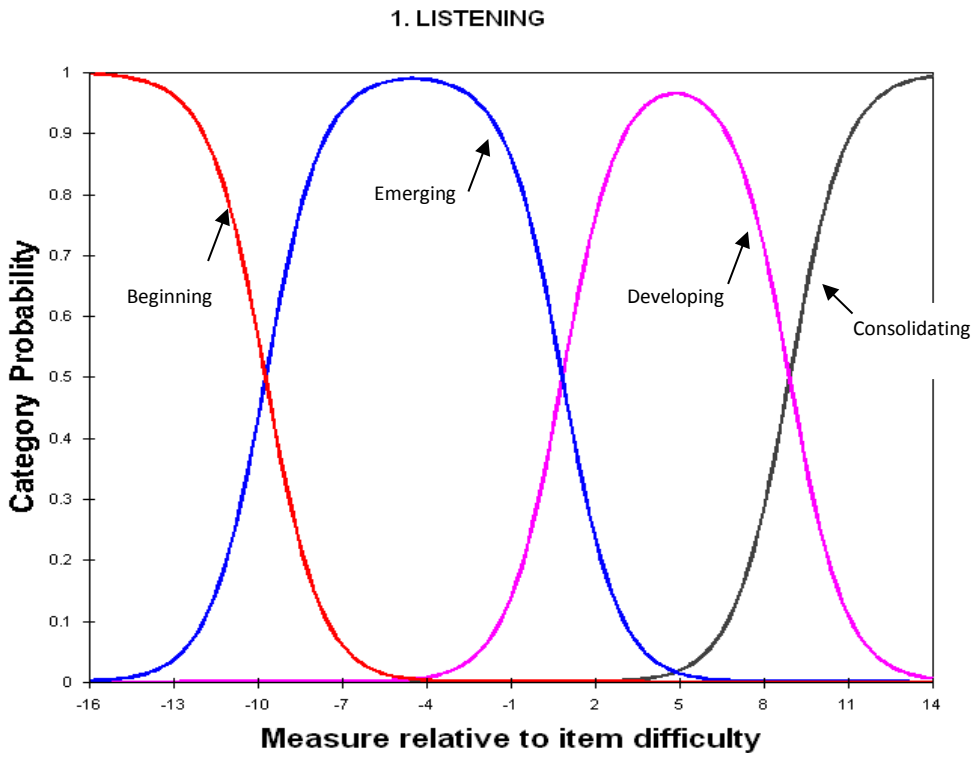
The examination of all of the CPC graphs (total 4 modes x 5 grades = 20) confirm that the EAL/D phase categories are working in accordance with the expectations.

Table 10 Average ability measures and Rasch-Andrich thresholds

Mode	EAL/D Level	Kindergarten		Year 3		Year 5		Year 7		Year 9	
		Average Ability Measure	Thresholds	Average Ability Measure	Thresholds	Average Ability Measure	Thresholds	Average Ability Measure	Thresholds	Average Ability Measure	Thresholds
LISTENING	1	-9.2	na	-10.28	na	-12.86	na	-8.82	na	-11.88	na
	2	-3.98	-8.19	-4.29	-8.21	-4.07	-9.73	-2.46	-6.76	-2.7	-10.35
	3	2.28	0.43	2.73	-0.85	3.6	0.83	1.86	0.87	3.56	2.42
	4	7.94	7.76	10.63	9.06	9.4	8.91	6.44	5.89	8.67	7.93
SPEAKING	1	-8.96	na	-11.38	na	-12.86	na	-9.01	na	-11.01	na
	2	-3.91	-7.06	-4.39	-8.61	-4.49	-9.81	-2.69	-7.06	-2.84	-9.09
	3	1.71	0.37	2.75	-1.06	3.68	0.11	1.49	1.15	3.14	1.76
	4	6.9	6.69	10.62	9.67	9.78	9.7	6.32	5.91	8.33	7.33
READING	1	-7.91	na	-10.9	na	-13.61	na	-8.12	na	-10.38	na
	2	-1.78	-4.99	-3.95	-10.04	-3.86	na	-1.94	-6.56	-0.99	-8.38
	3	4.28	4.99	4.03	-0.11	3.8	-3.97	2.66	-0.13	4.35	1.23
	4	8.9	na	11.25	10.16	9.56	3.97	8.52	6.69	9.39	7.15
WRITING	1	-7.24	na	-10.49	na	-12.63	na	-8.08	na	-10.97	na
	2	-0.56	-4.67	-1.99	-10.57	-2.37	-11.06	-1.23	-7.01	-1.03	-9.47
	3	6.6	4.67	6.34	0.23	5.22	1.58	2.9	0.45	4.51	1.84
	4	8.82	na	11.67	10.35	9.97	9.49	8.04	6.56	9.02	7.62

Note: Levels 1, 2, 3, 4 correspond to four ordered EAL/D phases: Beginning, Emerging, Developing and Consolidating.

Figure 11 Category Probability Curve for the listening mode, from the Year 5 dataset



For example, Figure 11 (obtained from Winsteps for the listening mode, from the Year 5 dataset) shows that all categories on the rating scale are "modal"; that is, each category is the most probable response category for some portion of the latent construct. The Rasch-Andrich thresholds (located at the intersections of adjacent probability curves, also reported in Table 10) are spread across the latent continuum; they are neither too close nor too far apart. Collectively, all the phase categories help in defining distinct points on the latent construct being measured. No disordered thresholds (i.e., where a higher threshold such as the 2–3 threshold has a lower measure on the latent continuum than a lower threshold such as the 1–2 threshold) are observed. Higher ability persons are more likely to score in a higher category than lower ability persons, across the continuum, as expected.

To summarise, fit statistics presented in the above sections confirm that teachers' mode judgements obtained from using the EAL/D instrument contribute to the development of a single ability continuum. Furthermore, there is sufficient evidence to indicate that the four rating scales are functioning properly, with a higher phase on a rating scale corresponding to a higher overall ability level, and vice versa. Taken together, it means that, for each student, the EAL/D assessments made by teachers across the four modes can be summarised into a single score, for the practical purpose of comparing a single ability.

4.2.4 Structural aspect of validity

In this section, the internal structural patterns in the teacher EAL/D mode judgements is compared to the expected interrelations among the different aspects of language proficiency of EAL/D students, derived from the theory of second language acquisition or prior empirical observations. The fundamental idea is that if the EAL/D instrument is developed based on the domain theory and is used to measure the underlying ability in an appropriate manner, the internal structure of the assessment scores should be consistent with what is known about the structural relations inherent in behavioural manifestations of the underlying construct (Messick, 1996; Loevinger, 1957).

In order to investigate the structural patterns in the EAL/D judgements, the same Rasch analysis was re-run, first using the combined data of Year 3 and 5 (referred to hereunder as the 'primary student data') and then using the combined data of Year 7 and 9 (referred to hereunder as the 'secondary student data'). The data is combined to yield more reliable structural patterns for interpretation.[11] For each dataset, after the influence of the first factor (i.e., the Rasch measures) has been removed from the raw data, a second factor, which explains the most variance in the residual data, is extracted from the residual data using the Principal Component Analysis (PCA).[12]

Figures 12 and 13 show the factor loadings from the PCA derived from the primary and secondary datasets respectively. In both figures, the x-axis represents the underlying ability continuum (in logits). Mode difficulties and person abilities are both represented along this continuum. The modes are located on both figures based on their estimated difficulties – easier modes with negative logits are on the left and the more difficult modes with positive logits are on the right. The y-axis represents the second factor within the residual data. The numbers on the y-axis indicate the factor loadings; that is, the correlations the modes have with this second factor. The sign of the loadings itself is arbitrarily set by the Rasch Program – Winsteps.

If the data fits the uni-dimensional model perfectly, the distribution of the standardised residuals should resemble a normal distribution, and the expectation is that the second factor would have a large factor loading on one mode and small loadings on other modes (Linacre, 2010). If there are clusters of modes which have significant loadings (i.e., those situated at the top or the bottom of the plots), then the meaning of the second factor is interpreted by contrasting the modes with opposite signs of loadings (Linacre, 2010). This trial uses a factor loading of 0.5 (Daftaripard & Lange, 2009) as the cut-off for identifying

[11] In order to yield more reliable structural patterns for interpretation, teacher judgements for Years 3 and 5, and those for Years 7 and 9, were combined for the analysis carried out in this section. The fact that the same set of descriptors are used in the *EAL/D Learning Progression* for assessing students from Years 3 to 6, and for assessing students from Years 7 to 10 provides justifications for combing the datasets in the way outlined. Rasch analysis for both sets of data indicates satisfactory fit of data to the Rasch model and that all rating scales functioning as desired.

[12] For a description of the PCA extraction method being used to detect secondary factors in the data, see Linacre (2010, p. 319).

modes with substantial loadings on the second factor. In both figures, modes with loadings that are of the same sign and are equal to or greater than 0.5 are grouped together in boxes, for easy interpretation of the meaning of the second factor.

Figure 12 Factor loading of the second factor in EAL/D assessment data – primary students

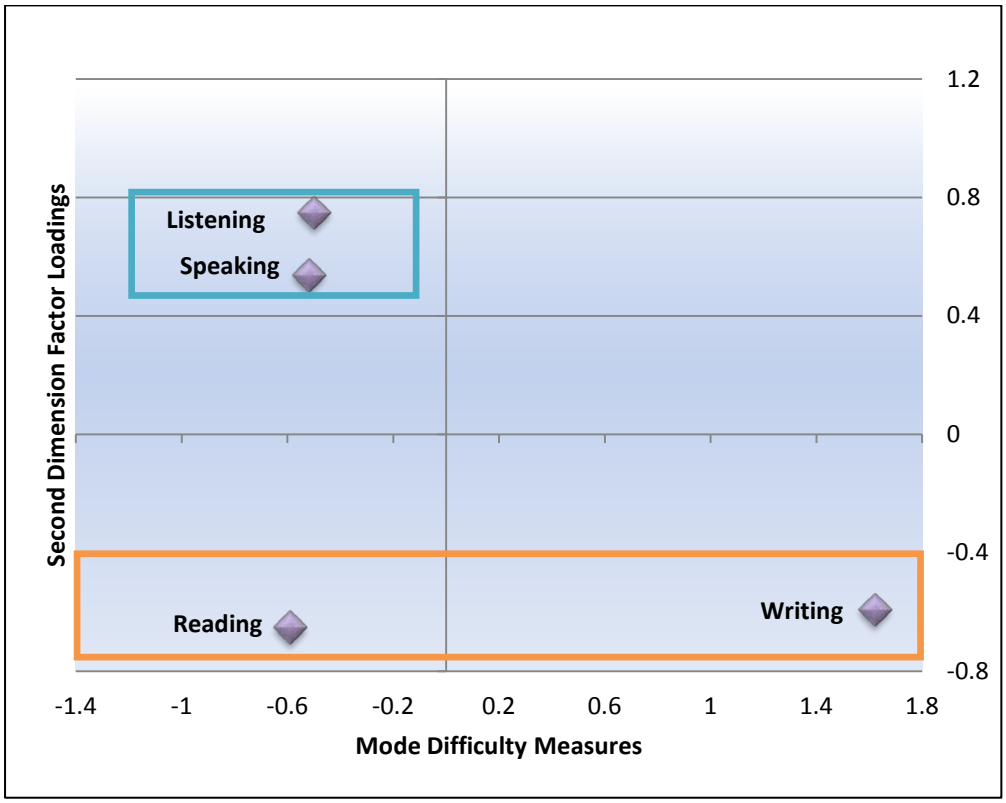
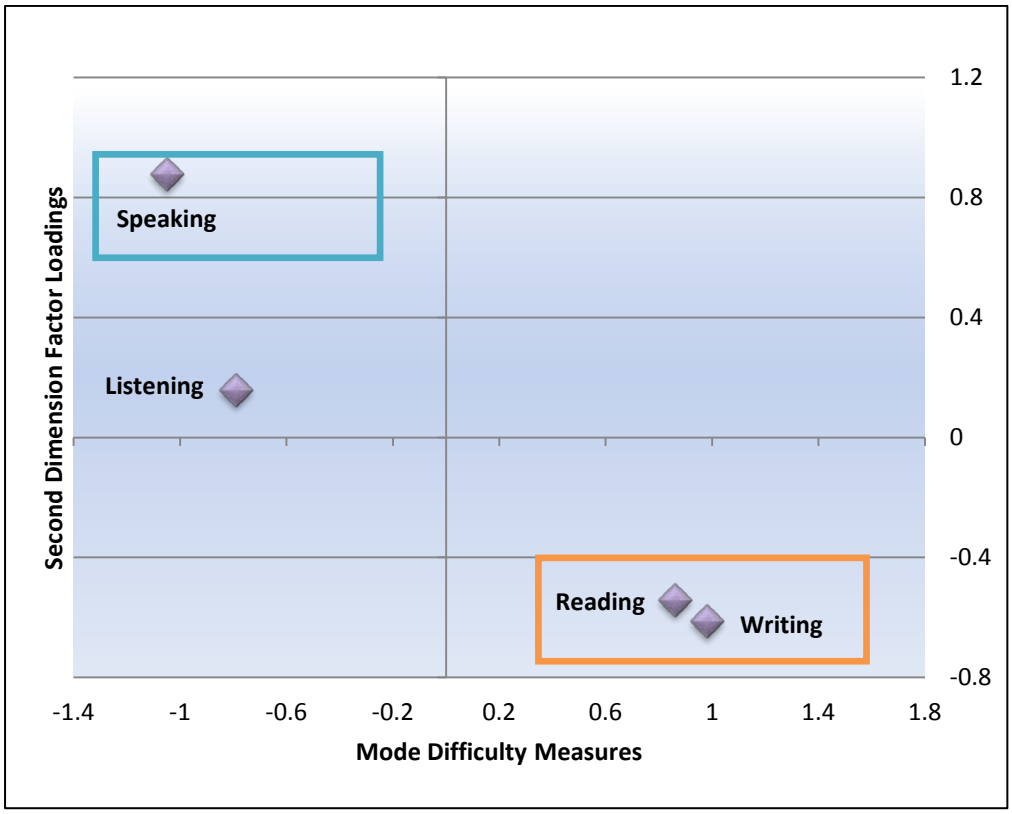


Figure 13 Factor loading of the second factor in EAL/D assessment data – secondary students



Results from the Winsteps program indicate that the majority of the variance in the EAL/D mode judgements is explainable by the uni-dimensional model (78% for the primary data and 72% for the secondary data). However, around 10% of total raw variance in both data sets is attributable to a second factor in the data. This second factor is characterised by the contrast between the academic aspects of the language (reading and writing) and the conversational aspects of language (listening and speaking for primary students, or speaking for the secondary students).[13] Furthermore, both figures also showed that, across the primary and secondary contexts, writing is consistently identified as the most difficult mode to develop or achieve, as compared to other modes.

The discovery of the second factor, indicating a tendency for reading and writing to develop at a different rate to listening and speaking, is not completely surprising, given findings from other studies relating to EAL/D students' progression along different dimensions of the English language.

[13] For the primary (Years 3 and 5 combined), the second factor explains 9% of variance in the raw data, while the Rasch measures (the first factor) explain 77.8% of the variance. The Eigenvalue of the second factor is 1.6. For the secondary data, the second factor explains 10.3% of the variance in the raw data, while the Rasch measures explain 71.8% of the variance. The Eigenvalue of the second factor is 1.5.

For instance, Cummins (1981, 1984) study suggested that academic English proficiency is more difficult to develop than conversational aspects of language for EAL/D students. His study demonstrated that, while it could take students up to 3 years to develop conversational competency, it could take between 7 to 10 years for them to develop academic language competency.

A reverse pattern though has also been observed with some students from Asian countries. Nunan study (2003) found that EAL/D students, particularly older Asian students who have had schooling in their home country, have higher proficiency in written than spoken English. While more countries within Asia are introducing English as a compulsory subject in the early years of school, the quality of English teaching and the qualifications required to teach are variable. As a result many students learn English in their home countries required for formal academic contexts rather than English interactional skills, thus developing a higher level of academic language skills than the conversational language skills.

4.2.5 Further evidence of concurrent validity

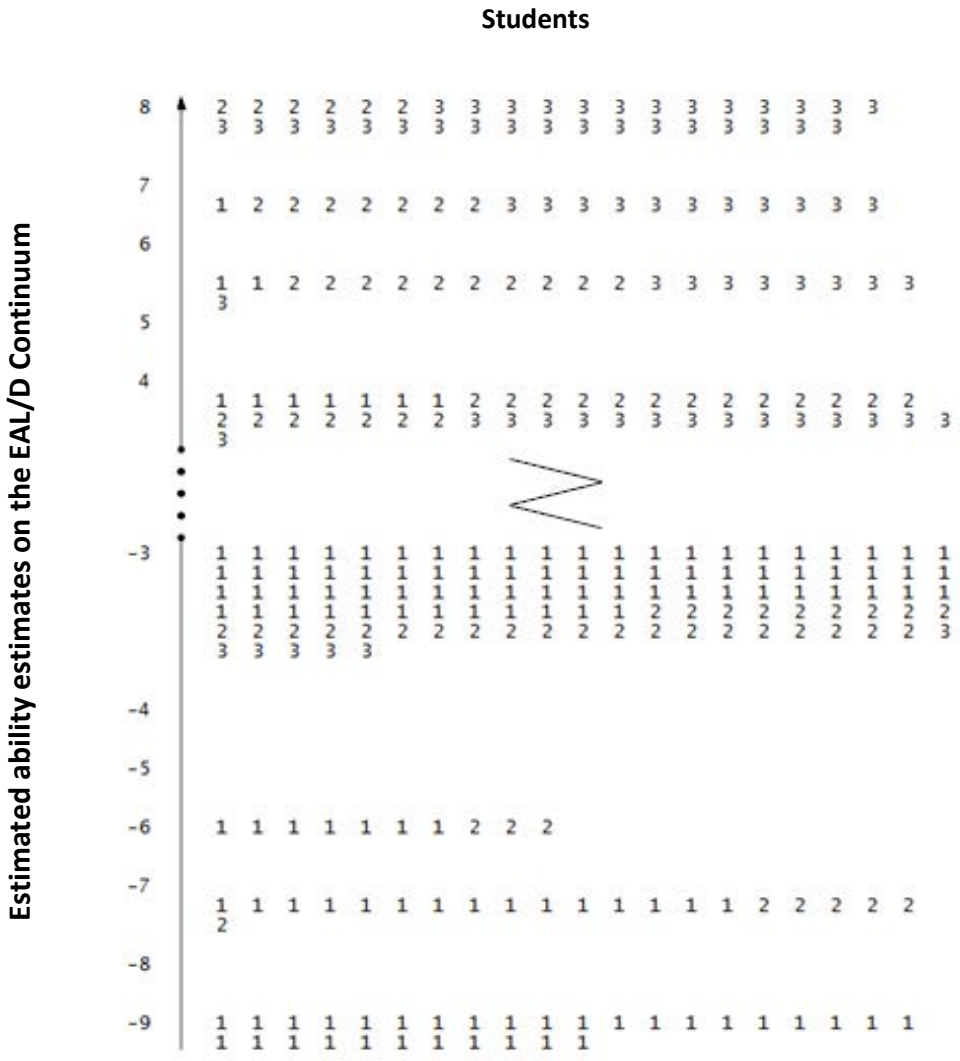
4.2.5.1 Alignment between the current ESL phase and EAL/D Learning Progression

This section presents further concurrent validity evidence on the relationship between the current tool used in NSW government schools to assess an EAL/D student's English language proficiency – the ESL phase tool and the EAL/D instrument. A reasonable level of congruence between the two instruments is expected, since both are theoretically similar constructs attempting to measure the same underlying ability. If the expected relationship is not realised in the data, questions might be raised on the validity of either instrument.

In order to inspect the concordance between the two instruments, two variable maps were obtained from the Winsteps program, one each from the Rasch analysis using the primary student data set and the secondary student data set respectively. Figure 14 is the variable map obtained from the secondary dataset. As the equivalent map derived from the primary student data is similar to Figure 14, it is not included here.

Figure 14 demonstrates that there is a reasonable level of concordance between teacher judgements using the current ESL phase tool and the EAL/D instrument. Values on the vertical axis in the figure indicate the Rasch ability estimates of students based on teachers' judgements using the EAL/D instrument. Each student is also represented on the map by the ESL phase assigned by the student's owner teacher using the ESL phase tool. It is observed that students who have high ability estimates from the EAL/D instrument (i.e. those located at the upper end of the continuum) were mostly assessed as ESL phase 3 students. Conversely, students with limited English proficiency (i.e., those located at the low end of the ability continuum), were mostly assessed as ESL phase 1 students.

Figure 14 **Locations of students on the EAL/D Proression continuum with their ESL phases**



Note: Variable map obtained from the secondary student dataset (Years 7 and 9 combined).

Another method of demonstrating the concurrent validity for the EAL/D instrument is to map the average ability estimates of students in the three ESL phases on the EAL/D Progression continuum. This analysis could also be useful if any future funding formula based on the EAL/D phase categories may need to be aligned with the current funding arrangements using the three ESL phases, as this analysis provides additional information on the alignment between the two instruments.

Figure 15 illustrates the results when the primary student data is used in the analysis and Figure 16 when the secondary student data is used.

On both figures, the Rasch average ability measures, estimated from the EAL/D assessments, of the Phase 1, 2, and 3 students are identified on the continuum. Also identified on the same continuum are the ability estimates of students with various typical EAL/D proficiency profiles (e.g., from students who were assessed Beginning across all four

modes, to those who were assessed Consolidating across all the four modes). This makes it easier to understand the language developmental profiles of an average ESL Phase 1, 2 or 3 student, relative to major developmental milestones on the *EAL/D Learning Progression* continuum.

Figures 15 and 16 demonstrate that higher ESL phases, on average, are associated with higher overall language proficiency levels on the *EAL/D Learning Progression* continuum, another indication that the EAL/D instrument is operating as intended.

For both primary (Years 3 and 5) and secondary (Years 7 and 9) students, the average ability estimate of a Phase 1 student corresponds to that of a student achieving Emerging over all four language modes.

Figure 15 Average locations of the Phase 1, 2 and 3 students – primary student data

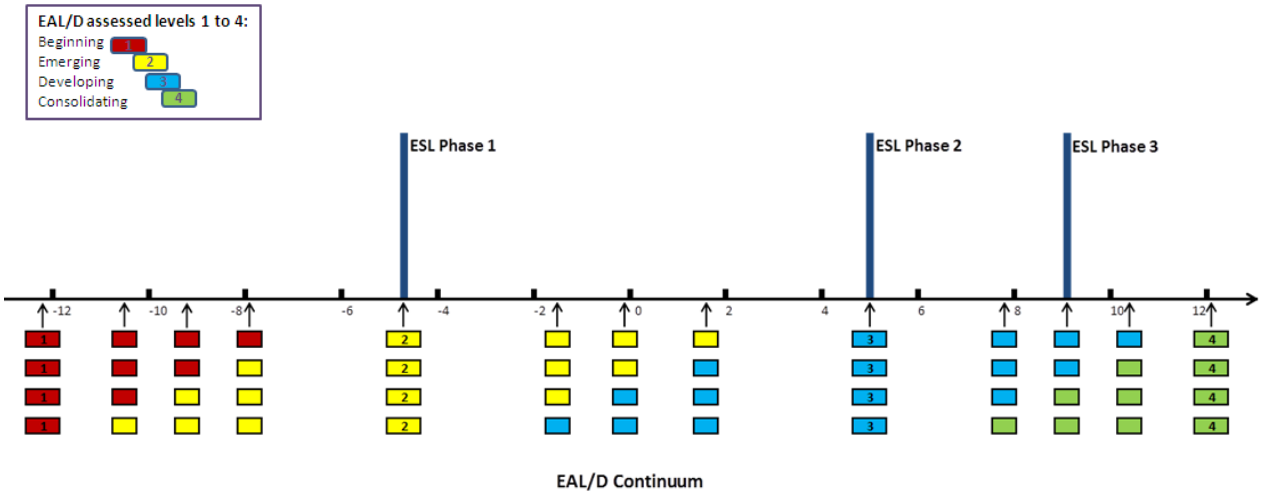
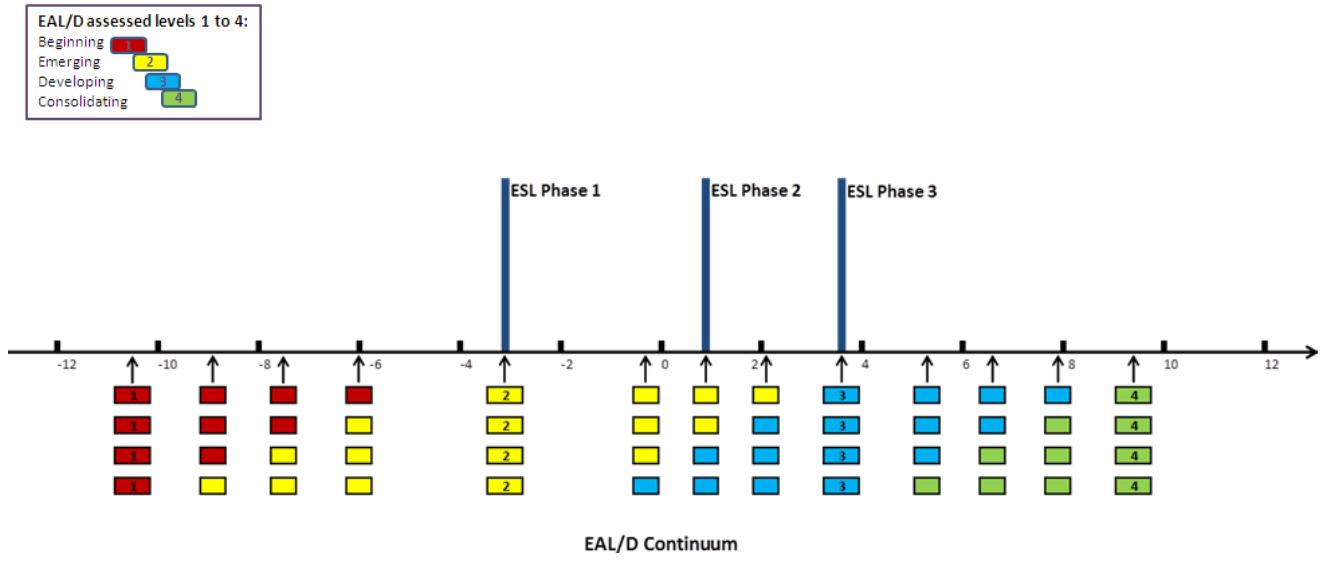


Figure 16 Average locations of the Phase 1, 2 and 3 students – secondary student data

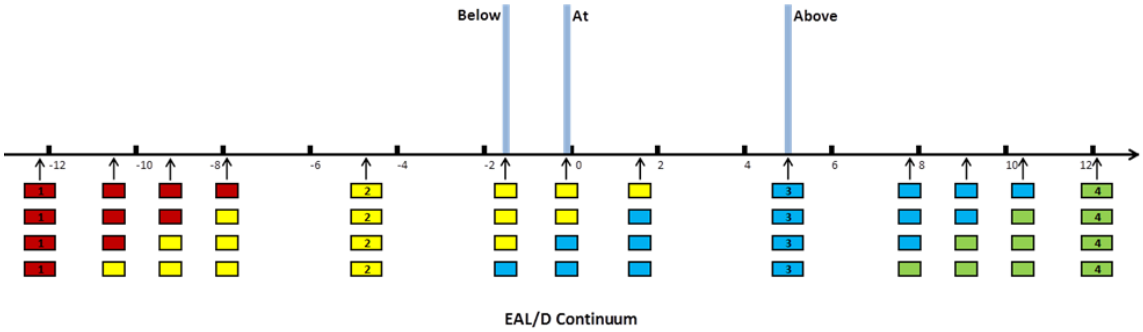


For a Phase 2 or Phase 3 student, however, the relative location of these students on the continuum is different for primary than for secondary students. While the average ability estimate of a primary Phase 2 student is the same as that of a primary student assessed at the Developing phase across all four modes, the average ability estimate of a secondary Phase 2 student is the same as that of a secondary student assessed with two modes at the Emerging phase, and two modes at the Developing phase. This needs to be further investigated as it could be further evidence of the current ESL Phase tool not taking into account the different language proficiency characteristics in the EAL/D learners across different age cohorts.

4.2.5.2 Alignment between NAPLAN and EAL/D Learning Progression

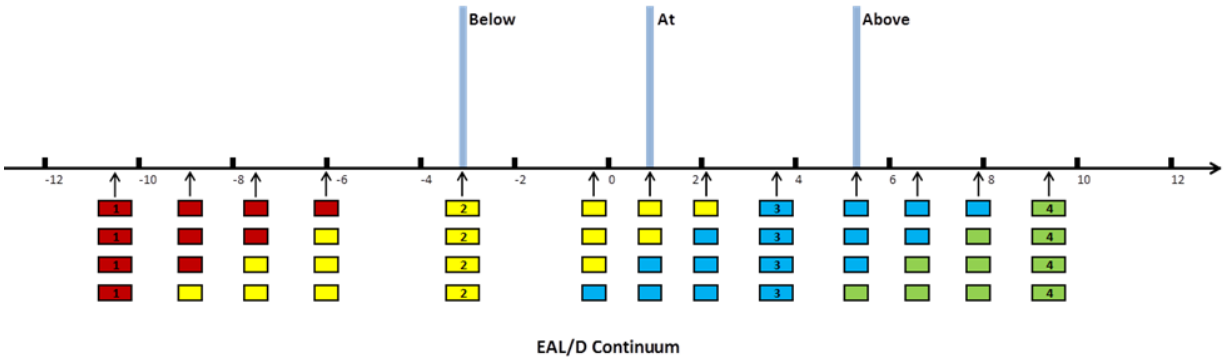
Similar analysis was also carried out to further examine the relationship between the EAL/D instrument and NAPLAN. A student matched to the NAPLAN test cohort is first coded in reference to the National Minimum Standard for his/her grade, based on the student's 2012 NAPLAN results on the reading and writing tests.[14] The average ability estimates of those students who achieved below, at or above National Minimum Standard (NMS) were then calculated and mapped onto the EAL/D continuum (see Figures 17 and 18 for the locations of these estimates, for primary students and secondary students separately).

Figure 17 Locations of ability estimates of students at, below or above NMS - primary



[14] A student is coded in reference to the National Minimum Standard using the following method: If achieving below National Minimum Standard for reading or writing – code as 'below'; otherwise if achieving 'at' National Minimum Standard for reading or writing – code as 'at'. If a student achieves above National Minimum Standard for both reading and writing, code as 'above'.

Figure 18 Locations of ability estimates of students at, below or above NMS - secondary



It is observed that the predominant relationship as demonstrated through these two figures is as expected. That is, higher language proficiency as assessed by teachers using the EAL/D instrument corresponds to higher levels of achievement in NAPLAN reading and writing tests. The relative locations of different groups of students are similar across the EAL/D continuums, constructed from using the primary and secondary student data separately.

4.3 Teacher feedback survey

The teacher feedback survey (Appendix G) and focus group meeting provided opportunities for teachers to provide quantitative and qualitative feedback on the utility of the *EAL/D Learning Progression* across a range of teaching contexts when compared with the existing ESL assessment frameworks; the trial process and adequacy of support for participating teachers; and the wording and layout of the *EAL/D Learning Progression* statements and indicators. A total of 74 teachers (77% of the trial participants) responded to the survey.

4.3.1 Feedback on the *EAL/D Learning Progression* utility

The survey showed overall a positive response to the *EAL/D Learning Progression* and support for its implementation in NSW government schools. Teachers generally felt confident to use the *EAL/D Learning Progression* to identify the broad level of English language proficiency of a range of students.

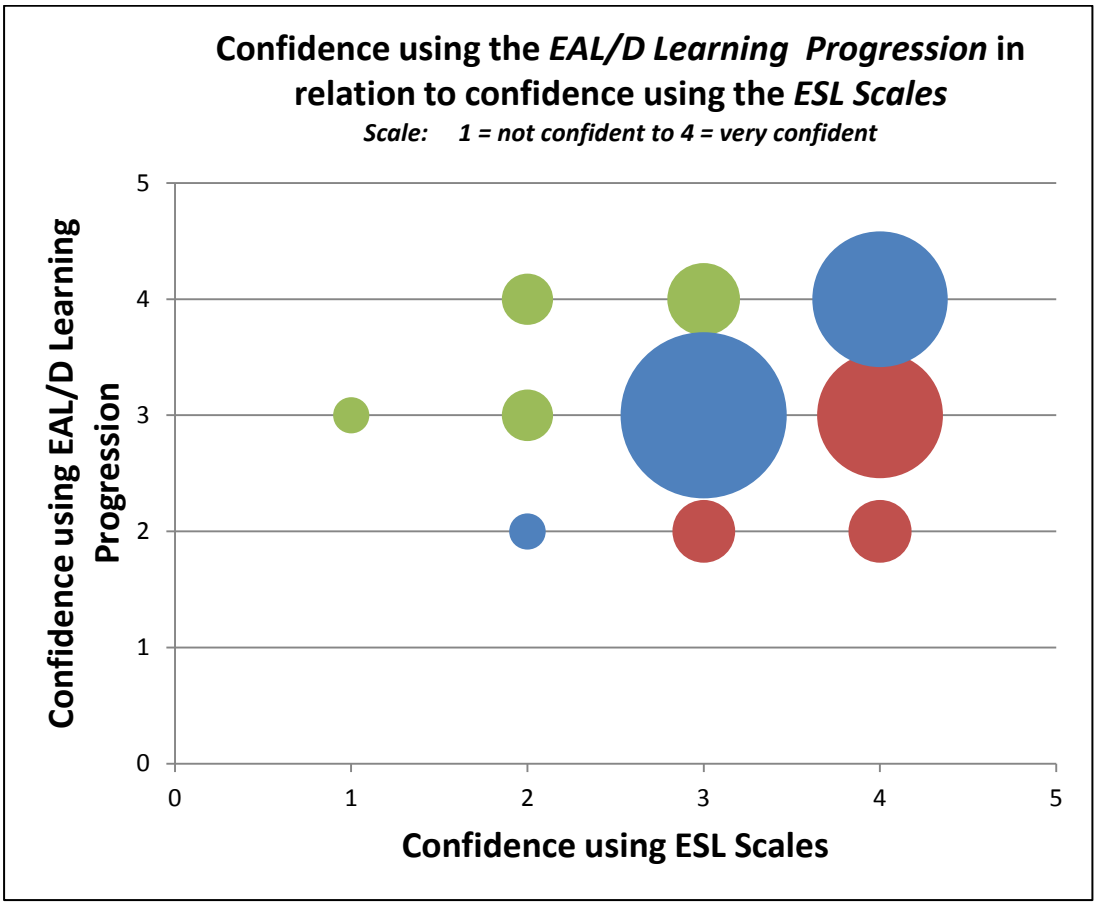
While the overall response to the tool and its usability was positive, there were a significant number of responses indicating that the *EAL/D Learning Progression* was too broad (16 comments).

It is likely that these negative responses to the *EAL/D Learning Progression* stem from some misconceptions about the tool's purpose and confusion with the existing *ESL Scales* document. With 74% of respondents having prior experience using the *ESL Scales* this is not surprising. The *ESL Scales* is an ESL assessment framework used by ESL teachers to describe the language learning needs of EAL/D learners. It is significantly different from the *EAL/D Learning Progression* in its level of detail, number of levels and presumed ESL specialist knowledge. Feedback showed that some teachers expected the progression to be a diagnostic tool like the *ESL Scales*. For example a number of teachers suggested the *EAL/D*

Learning Progression would be of limited use in an Intensive English Centre context where currently the *ESL Scales* are used to make detailed on-arrival assessment for class placement, to show English language development in the short term of stay in the Intensive English Program and to report students' English language proficiency to ESL teachers on transition to high school. The *EAL/D Learning Progression* would not be appropriate for this purpose.

Predicting some possible interference because of teachers' familiarity with the *ESL Scales*, the trial compared teacher confidence using the two tools. The results (shown in a bubble chart in Figure 19) indicate that most teachers had similar levels of confidence in using both the *ESL Scales* and the *EAL/D Learning Progression*. The confidence in using the *ESL Scales* is to be expected as it has been used in NSW since 1996. The similar level of confidence in using the *EAL/D Learning Progression* indicates how accessible this tool is after only two days professional learning.

Figure 19 **Confidence in using the EAL/D Learning Progression in relation to confidence using the ESL Scales**



Note: The size of the bubbles is proportional to the number of teachers in each category. Red bubbles indicate those teachers who reported higher levels of confidence in using ESL scales, while green bubbles indicate those who reported higher levels of confidence in using the EAL/D instrument. Blue bubbles indicate same levels of confidence.
Only teachers who have used ESL scales are included in this analysis.

Teachers specifically supported the use of the *EAL/D Learning Progression* to replace the current ESL Phases with 66% of respondents indicating their support. A number of teachers (6) commented in detail on the reasons for replacing the existing phase statements with the *EAL/D Learning Progression*. *The EAL/D Learning Progression* was seen as the more favourable tool because it showed the impact of the age and stage of schooling on phase characteristics. It does not link English language proficiency to length of time learning English, it is more detailed and therefore supports more accurate teacher judgement.

The survey results also indicated teacher interest in using the *EAL/D Learning Progression* more broadly. Other uses identified by teachers included:

- reporting EAL/D student need on the *ESL Annual Survey*
- informing programming and planning for teaching
- raising classroom teacher awareness of student needs
- broad assessment of student needs by class teachers
- monitoring and tracking EAL/D students' progress over time and at key transition points (e.g. IEC to high school and primary to high school)
- identifying students requiring targeted or intensive support

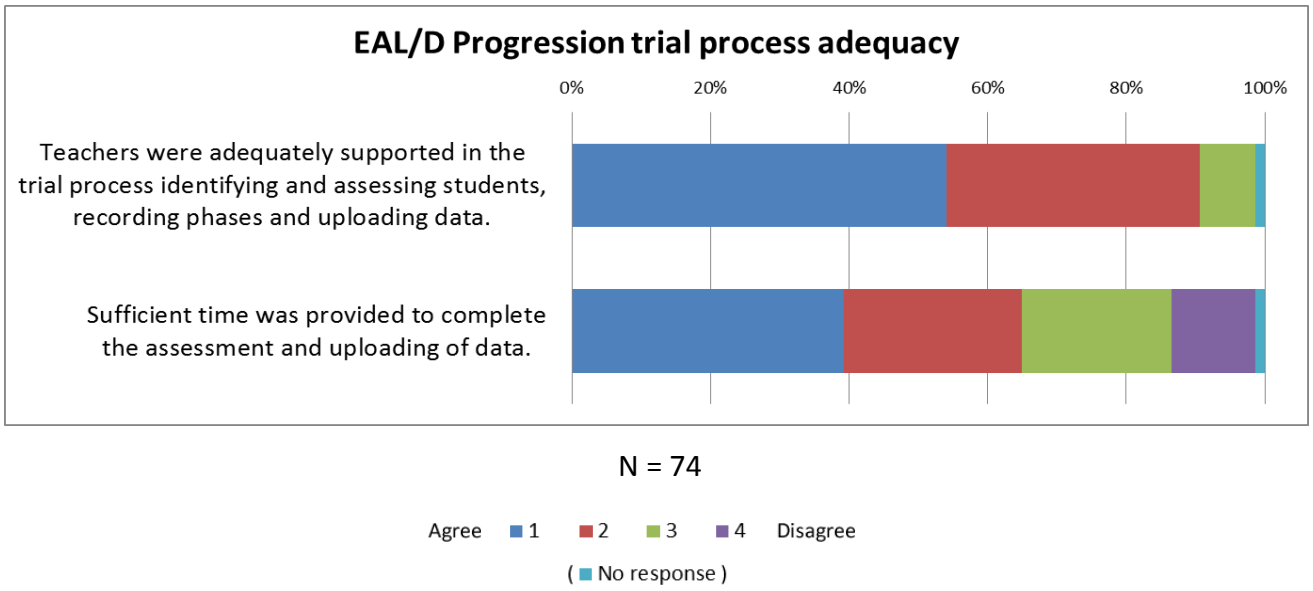
The trial included a small number of Aboriginal students (29 in total). The majority of the teachers surveyed (65%) indicated that they were not sure whether the *EAL/D Learning Progression* was appropriate for use with Aboriginal students. Of those who did provide a response, 14 thought the tool was appropriate, and 5 thought it was not appropriate for use with Aboriginal students. It is noted that, of the 4 teachers who did assess Aboriginal students, 3 responded that the *EAL/D Learning Progression* was appropriate and 1 was unsure.

The implications from teacher responses are that the *EAL/D Learning Progression* should be implemented with clear information about the purpose of the tool, its potential use in ESL, IEC and mainstream class contexts and how it relates to other tools including the current ESL phases, the ESL Scales and the literacy continuum. Furthermore, the issue of the appropriateness and relevance of the *EAL/D Learning Progression* for use with some demographic groups, in particular the Aboriginal students, needs to be further examined with a larger study.

4.3.2 Feedback on *EAL/D Learning Progression* trial and teacher support

Generally teachers found the training program comprehensive and well-organised. The survey data shows that around 90% of teachers claimed they were adequately prepared for identifying and assessing students, recording phases and uploading data (Figure 20). Comments from teachers indicated the training provided clear, concise and easy to follow instructions. Through the collegial deconstruction of and practice using the *EAL/D Learning Progression*, the training provided a close understanding of the tool that was necessary for its independent use in the school context.

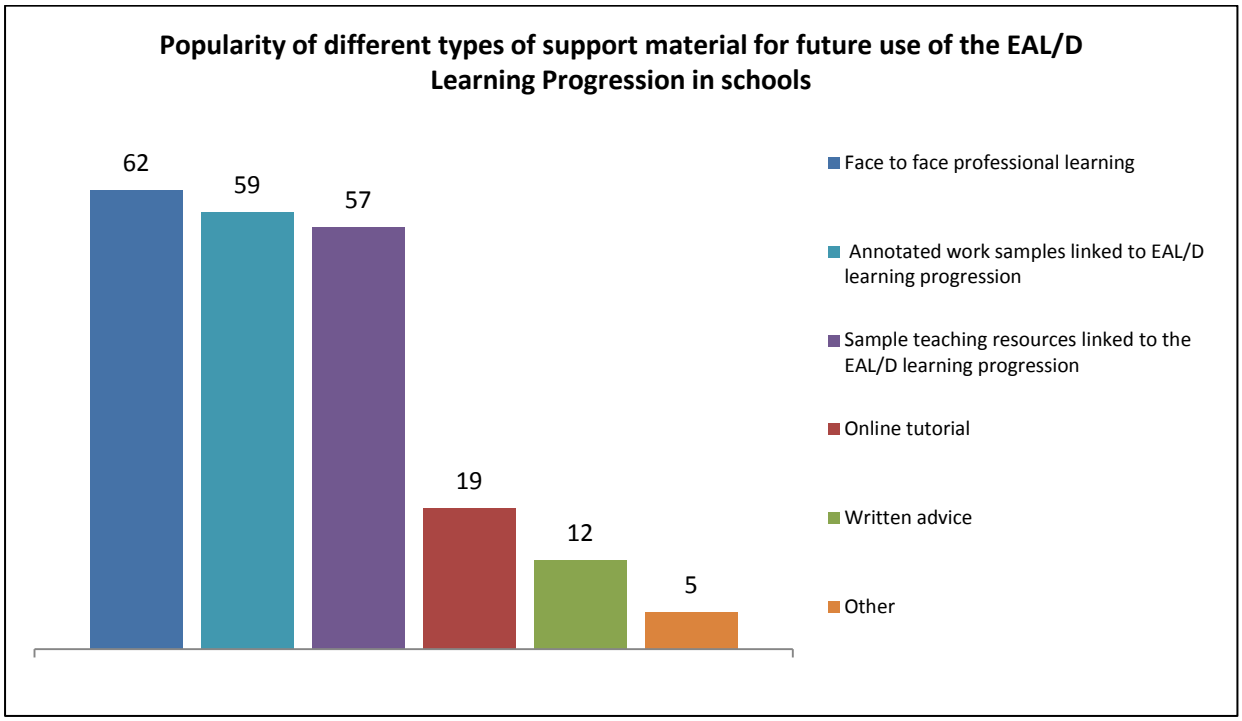
Figure 20 Teacher feedback on the *EAL/D Learning Progression* trial process



Survey data also indicates significant concern by teachers about the sufficiency of time provided to complete the assessment and uploading of student data (Figure 20). This was corroborated and exemplified by teacher feedback (see below).

The survey data (Figure 21) indicated that the preferred model of training delivery was face to face although those who participated in a combined model of face-to-face delivery and video conferencing with online tutorial support also found this mode effective. Teacher comments clearly indicated the value of collaborative and interactive learning. It is also clear from the survey and feedback that teachers valued highly the additional teaching materials including annotated work samples to support consistency of teacher judgement, recording and monitoring of student English language proficiency.

Figure 21 **Popularity of different types of support material for future use of the *EAL/D Learning Progression***



In the teacher feedback some specific issues were raised:

1. The participants had difficulty accessing from home the *EAL/D Learning Progression* site to enter data and complete the evaluation. While this technical difficulty was quickly resolved it indicates a deeper concern from teachers about the limited time to enter data and complete the survey at school.
2. Teachers reported timetabling issues which made it difficult to access students in the particular scholastic grade groups required for the trial. As a result some teachers reported insufficient access to target students and inadequate time to make informed EAL/D phase judgements based on ongoing observations and assessments.
3. Teachers reported insufficient time to complete all trial requirements including planning for assessment, collecting work samples, double marking and the online evaluation. The end of term 2 was a particularly busy time of the year to conduct the trial as teachers were marking final assessment tasks, writing reports and completing the *ESL Annual Survey*.
4. Teachers indicated difficulty collecting sufficient assessment evidence within the timeframe to confidently assess students' EAL/D phase level in each mode. Teachers' comments indicated that a range of assessment evidence both formal and informal was required to confidently make an EAL/D phase judgement in each mode. Where teachers had limited time to collect assessment data, they felt less confident about their judgements.

The feedback and survey data indicate the need for a program of mixed flexible and face to face delivery of teacher professional learning that provides opportunity for professional

dialogue and collaboration. While the level of teacher support was adequate additional annotated work samples now available through ACARA would enhance the consistency of teacher judgement using the progression.

To ensure teachers are able to make confident phase judgements teachers would need ongoing access to targeted EAL/D students and considerably more time to plan for assessment, gather assessment evidence over time and enter phase data for each mode.

4.3.3 Feedback on the *EAL/D Learning Progression* wording

A number of teachers provided some very detailed comments and suggested edits to the learning progression in their feedback to this question.

A few teachers suggested the indicators be organised so that they are in a consistent sequence or organised under sub headings to allow teachers to track progression across phases. It was suggested that phasing students would be easier if indicators were 'lined up'.

A number of comments were made from the perspective of non ESL specialists who were new to some of the terminology of this specialist field and suggested some criteria may be difficult to interpret. For example, 'Their first language influence is evident in the way they organise texts'. Similarly, teachers with limited experience teaching EAL/D students may find some gradation words difficult to interpret e.g.: 'limited' and 'varying'.

A specific expression used in the *Characteristic of the learner* statements, '*Learning a language requires constant focus and attention, and students will tire easily and may experience a high level of frustration*' (Beginning English: Limited Literacy Background), created a great deal of discussion in training workshops with contrasting responses from ESL specialists and mainstream teachers. While mainstream teachers welcomed this type of information, some ESL specialists felt it could lead teachers to overlook physical causes for these behaviours. The behaviours described may be health related – particularly for refugee students. For example, poor concentration in the classroom may have a health related, treatable cause.

A small number of teachers commented on the misalignment between the level of English indicated in the *Characteristics of the Learner* statement and that expressed in the indicators. For example, in the Years 3-6 EMERGING Speaking mode, the indicators for the 'Beginner Phase' seems to imply a beginner in their first days /weeks in Australia, whereas the statement in the 'Characteristics of the Learner Group', seemed to imply more ability.

This feedback will be forwarded to ACARA as authors of the *EAL/D Learning Progression* for consideration or possible amendments of the text.

5. Further discussions of the trial results

5.1 Reliability of teachers' judgements using *EAL/D Learning Progression*

Analysis of agreement rates, kappa rates and inter-rater correlations showed that the *EAL/D Learning Progression* enabled teachers to make consistent judgements of English language proficiency across the modes. While teacher judgements were less consistent with some student cohorts (e.g., boys and students of relatively higher language proficiency) and in speaking and listening modes, these results do not necessarily render the *EAL/D Learning Progression* instrument unreliable. A number of other factors unrelated to the EAL/D could have contributed to the variations in the reliability of teacher judgements.

For example, the data indicated lower rate of consistency in teachers' judgements when assessing certain groups of students such as ESL Phase 3 students. This could be related to limited teacher experience assessing phase 3 students. ESL teachers usually target phase 1 and 2 students for specialist face-to-face teaching support with limited time available for phase 3 students. For phase 1 and 2 students EAL/D phase judgements would be based on ongoing observation and assessment. In contrast, EAL/D phase judgement of phase 3 students would be based on a limited number of observations or a single assessment task, whatever was possible in the timeframe allowed for this trial.

Conversely the data showed a high consistency rate amongst IEC teachers. Again this could be attributable to teacher capacity to develop a close knowledge of their students' English language proficiency. IEC teachers have a detailed knowledge of their students because they work with smaller groups of students for extended periods of the school day; teachers have access to a wealth of student background details and to specialist bilingual and counsellor support. IEC teachers are required to engage in continual professional learning to ensure consistency of teacher judgement against the IEC curriculum framework. Teachers are required to assess and report student progress each term to determine class placement or high school readiness. An additional factor contributing to the observed higher consistency in teachers' assessments of IEC students' English proficiencies is that most of the students at the IEC are of low English proficiency levels, and generally speaking, English language proficiency is easier to assess at lower levels where the language characteristics are better understood by most teachers.

It is noted that the trial was conducted before ACARA released annotated work samples for each mode, which would have supported teachers to make more consistent judgements. If a national trial is conducted, it will be interesting to see if these additional materials will make a difference to the inter-rater reliability.

Results also pointed to greater difficulty in assessing listening and speaking modes consistently, as compared to reading and writing. This may have resulted from limited teacher experience in assessing informal student interactions which dominate the listening

and speaking mode indicators of the EAL/D instrument. In addition, teacher responses to the survey indicated that time constraints of the trial meant teacher judgement on oral language skills were sometimes based on insufficient assessment evidence. While teachers routinely collect formal spoken task assessment data in response to curriculum demands they rarely routinely collect and track observational data of student interactions.

In terms of evaluating the suitability of the EAL/D instrument for identifying English language proficiency to inform important decisions such as resource allocation for schools, results from the Generalisability analysis indicated that the level of consistency in teachers' phase judgements, made using the EAL/D instrument, on any one language mode, exceeded the industry standard required for score reliability for assessments to be used in high-stakes situations. Furthermore, corroborating evidence from adjacent agreement rates analysis confirmed that any chance of significant variations in teacher judgements on a language mode is contained within an acceptable level.

Additional results also showed that asking teachers to further discriminate between the two Beginning levels (*Some print literacy* and *Limited literacy background*) does not have a significant adverse impact on the observed rate of consistency in the resultant teacher judgements. This provides evidence to support the use and collection of all five levels from the *EAL/D Learning Progression* when identifying English language proficiency, if the EAL/D instrument is implemented in NSW government schools, and if additional time required for assessment and collection is not an issue.

In the comparison of judgements by owner teachers (those who have greater opportunities for ongoing interaction with the students assessed) and non-owner teachers (those who have limited ongoing interaction with the students and who would rely primarily on collected work samples to make assessments), the tests showed that there was no significant difference between student phase judgements. These results have implications for the implementation of the EAL/D in schools. They suggest that teachers can use the *EAL/D Learning Progression* consistently to assign a phase regardless of the opportunities for ongoing interaction with the student, so long as they have expertise in using the tool and follow the assessment guidelines in collecting a number of work samples over time to inform phase judgements.

5.2 Construct validity of the *EAL/D Learning Progression*

The analysis of **discriminant validity** showed that teachers not only understood the conceptual distinction among the various language modes, but they could also use the *EAL/D Learning Progression* to empirically discriminate between the modes in the appropriate way. This means that teachers can decide on a phase on a mode without judgement being clouded by student performance on another mode.

Concurrent validity evidence indicated that the relationship between the assessed EAL/D phases and NAPLAN performances on a similar trait of language proficiency for the same

students was as expected. In general, higher EAL/D phases were indeed associated with higher mean NAPLAN scores, and vice versa. Similarly, there is a reasonable level of congruence between phase judgements made by teachers using the current ESL tool and the EAL/D instrument, which helps strengthen the argument for the validity of the EAL/D construct.

Further analysis comparing the current ESL phases and NAPLAN results showed that *EAL/D Learning Progression* is a more refined instrument than the current phase tool in terms of identifying EAL/D students' needs for the purpose of allocating resources.

The psychometric analysis indicated strong **measurement evidence** substantiating the claim that the four language modes were measuring the same underlying ability. There is no evidence of modes under-fitting the uni-dimensional measurement model, which would have suggested other significant sources for explaining variations in assessed performances across the modes. Nor was there any evidence of modes over-fitting the measurement model, which would have indicated that they were operating too similarly to each other, even though they were meant to measure different aspects of the latent ability.

The four rating scales were also shown to be operating satisfactorily, with strong evidence indicating that the four phase categories were being used consistently by the teachers and as intended. As expected, a higher phase implied a higher level of overall ability, and vice versa. When the above evidence is taken together, it means that, for each student, a single score, derived from summarising the EAL/D phase levels across the four language modes, is sufficient to characterise the entire profile of the student's performances.

The range of the measurement evidence collected not only improves users' confidence in the interpretations of or the inferences that may be drawn from the assessments produced from using the EAL/D instrument, but it also broadens the utility of this instrument. This is because for the purpose of resource allocation a single score for a student is much more practical.

The analysis of the internal **structural** patterns of teacher judgements using the instrument shows the expected contrast between academic aspects of the language (reading and writing) and conversational aspects of the language (listening and speaking). It also shows that writing is consistently being identified as the most difficult mode to achieve as compared with other modes. This is consistent with second language acquisition research (Cummins, 1981) indicating that spoken language is acquired before more academic written language. The fact that the structural patterns in the EAL/D assessments are consistent with findings from other studies on EAL/D students' progression along the different aspects of English language ability provides evidence that the *EAL/D Learning Progression* is operating as intended.

5.3 Implications of the trial results on the EAL/D implementation in NSW

The survey data analysis showed that the *EAL/D Learning Progression* is able to be used confidently by both ESL specialist and mainstream teachers to identify a broad phase of English language proficiency. However, implementation of the instrument needs to take into consideration the current ESL teaching context in NSW schools and articulate the purpose of the *EAL/D Learning Progression* in relation to existing assessment frameworks that exist – particularly the *ESL Scales* and the *ESL Phases*. The *ESL Scales* remain a valuable tool to the ESL specialist teacher for assessment, programming and planning as it describes in far greater detail than the *EAL/D Learning Progression* the English language development of EAL/D learners across the modes. The *EAL/D Learning Progression* is an appropriate replacement of the current ESL phase framework as it supports teacher professional judgement with age and stage appropriate phase indicators.

The implementation of the NSW syllabuses and literacy continuum K-10 from 2013 has implications for the process of implementation of the *EAL/D Learning Progression* in NSW. In addition to its role as a resource allocation measure, *EAL/D Learning Progression* is seen as a more accessible tool than the *ESL Scales* for assisting mainstream teachers to identify and plan support for EAL/D students. To support mainstream teachers to use the *EAL/D Learning Progression,* it should be linked to the syllabuses and literacy continuum. Additional resources would need to be developed to support mainstream teachers to use the instrument for this purpose.

The implementation in NSW from 2014 of the new resource allocation model, including a separate loading for EAL/D students, requires the immediate design and implementation of a professional learning program primarily to introduce the *EAL/D Learning Progression* as the ESL resource allocation measure. The professional learning program should be offered to all teachers of EAL/D learners across all schools through a mixed mode of face-to-face and online delivery that provides ample opportunity for professional dialogue and collaboration. Teachers would have access to a range of teacher support including additional annotated work samples now available through ACARA.

The timeframe for professional learning and implementation would need to be carefully staged to ensure teachers had ample time to plan for assessment, gather assessment evidence over time and enter phase level data. In some schools where almost the entire student population are EAL/D students, considerable time will be needed to assess and record phases for all students. For this reason, professional learning would need to begin early in 2013 with the aim to have all schools reporting in the *ESL Annual Survey* in 2014 using the *EAL/D Learning Progression*.

Schools should be informed of the implementation of the *EAL/D Learning Progression* at the start of 2013. Professional learning should occur during terms one and two. The new school enrolment system (SALM), which is currently being developed, should build in the capacity

for teachers to enter EAL/D phase data. For the 2013 *ESL Annual Survey* process the current ESL phase data will be required. From the beginning of term 3, 2013 teachers should start recording *EAL/D Learning Progression* phase data. The *ESL Annual Survey* in 2014 will use *EAL/D Learning Progression* phase data to allocate ESL resources to schools in 2015. Teachers will have at least 4 weeks to enter *EAL/D Learning Progression* data on the data collection site. The first year of full implementation will be a steep learning curve for school teams. Support for schools should be focused in the following areas:

- provide information early in the school year as well as prior to the data collection.
- complete face-to-face training sessions prior to the data collection in the form of local area workshops.
- provide online self-access support material to enhance and supplement the training.
- engage Regional/State Office representatives to promote widespread understanding of the model.
- develop professional learning modules specifically devoted to the implementation of the *EAL/D Learning Progression.*

6. Conclusions

The NSW trial of the *EAL/D Learning Progression* has provided sufficient reliability and validity evidence for it to be used in NSW government schools as a broad measure of English language proficiency for resource allocation. In addition, the NSW trial has national implications in terms of informing the design and the associated cost-benefit analysis of a prospective national trial. Such a national trial would provide recommendations about the potential use of the progression to report English language proficiency across jurisdictions.

6.1 Can teachers assess each of the four language modes consistently using the *EAL/D Learning Progression*? Does this consistency vary across different types of students?

In general the research showed that the *EAL/D Learning Progression* enabled teachers to make consistent judgements of English language proficiency across the modes. While teacher judgements were less consistent for some student cohorts (e.g., boys and students of relatively higher language proficiency) and in speaking and listening modes, these results do not necessarily suggest that the instrument itself is unreliable. A number of other factors could have influenced the reliability of teacher judgements. These factors include teachers' prior exposure to assessment tasks for certain modes, understanding of the language characteristics of certain groups of students, and time required and available to collect work samples and make judgements. Targeted training programs could be further developed based on the empirical evidence collected in this trial to further improve the reliability of teacher judgements in certain areas. The findings from this trial indicate that teachers make

more consistent judgements when they have time to develop a detailed profile of the student and when the assessment tasks provide sufficient information that correlates with the *EAL/D Learning Progression* indicators.

Implications for future development:

(a) Professional learning needs to stress the importance of judging students' phases based on ongoing student assessment. Resources will need to be developed to support teacher assessment using the *EAL/D Learning Progression* instrument. This would include samples of assessment tasks and annotated work samples demonstrating the evidence required to judge students against the *EAL/D Learning Progression* indicators. For the reading mode the assessment resources should demonstrate how to collect evidence of constrained and unconstrained reading skills. For the writing mode assessment resources should include a range of text types across school subjects. For speaking and listening modes the assessment resources should exemplify assessment tasks and student work in both formal presentation tasks and informal procedural contexts.

(b) A national trial is needed to test the generalisability of reliability evidence from this trial. A national trial would need to include a broader range of demographics and larger cohorts of specific student groups, in particular Aboriginal students. Such a trial should include recommendations about the potential use of the *EAL/D Learning Progression* to report English language proficiency across jurisdictions. Furthermore, the national trial, with a larger student cohort and guided by sample assessment tasks and annotated work samples, should also examine whether the observed differences in the consistency of teachers' judgements across the four modes are replicated in the national data. Evidence of differential reliability across groups of schools, students and modes will also be useful to inform future programs of teacher training and professional development.

6.2 Is there evidence of construct validity to support the intended interpretation of the assessed levels of the EAL/D students?

In order to construct a comprehensive network of evidence to support or challenge the intended interpretations and use of EAL/D assessments, the trial collected evidence pertinent to four aspects of validity – concurrent, discriminant, measurement and structural aspects of validity. Findings from the trial suggest that there is sufficient evidence to support the claim that the *EAL/D Learning Progression* provides a balanced and accurate reflection of English language development, and it can be used to derive a single measure of English language proficiency for the purpose of allocating ESL funding to schools.

The research also found that the *EAL/D Learning Progression* is a more refined tool than the current ESL phases used in NSW. Given the extent and the range of validity and reliability

evidence collected, the *EAL/D Learning Progression* is a viable tool to replace the ESL phases as a broad measure of English language proficiency and as the basis for the formula used to allocate ESL funding to schools.

Implications for future development:

(a) The *EAL/D Learning Progression* could replace the NSW ESL phase tool as the broad measure of English language proficiency and as the basis for the formula used to allocate ESL funding to NSW government schools. The current ESL resource allocation model needs to be reviewed, and modelling conducted, using the *EAL/D Learning Progression* in place of the current three ESL phases.

(b) A program of professional learning needs to be implemented in NSW government schools throughout 2013. This will ensure adequate time to prepare teachers to report using the *EAL/D Learning Progression* phases in the ESL NAP and ESL Annual Surveys in 2014, which will inform the ESL resource allocation in 2015.

(c) Guidelines need to be developed describing the process of deriving an overall phase judgement from the four language mode phase judgements for each EAL/D student.

(d) The new school administration and management system (SALM) currently being developed would need to build in the capacity for teachers to enter EAL/D mode and phase data. Teachers should be able to update at any time the EAL/D phase judgements for individual students in SALM by 2014.

(e) As an interim measure, during 2013, data collection tools should be revised so that *EAL/D Learning Progression* phase data can be collected in each mode for every ESL student. This would replace the current ESL phase data collection tools in schools.

6.3 Considering the importance of training on teachers' approach to using the *EAL/D Learning Progression*, what types of training, how much training, and what additional support should be provided to teachers prior to the use of the instrument?

The survey data analysis showed that while teachers felt they had adequate support to understand and confidently use the instrument, additional time is required for teachers to plan for assessment, to collect assessment data over time across the modes and to report student phases. Participants were positive about the content and type of professional learning indicating that future professional learning should be predominantly face-to-face, supported by additional teaching materials including annotated work samples and recording and monitoring proformas.

The detailed feedback provided by teachers may be useful for ACARA and its related working group to consider future enhancements to the *EAL/D Learning Progression* and its supporting packages.

The survey indicated a range of views about the purpose and possible uses of the *EAL/D Learning Progression* in schools. It is apparent from the trial that the *EAL/D Learning Progression* has a clear purpose as a broad measure of English language proficiency for resource allocation. While the instrument has been designed as a tool to support classroom teachers to identify their EAL/D students' needs and to monitor their linguistic progression, initially the focus of professional learning should be on supporting teachers to make consistent phase judgements. However, many teachers in the trial were interested in using the *EAL/D Learning Progression* to inform mainstream program planning and monitoring and tracking student progress. With the new NSW syllabuses and literacy continuum being implemented in 2014, it is timely that links be explored and appropriate support material developed if appropriate.

Implications for future development:

(a) Clear guidelines as to the purpose and potential uses of the *EAL/D Learning Progression* should be provided to participants in professional learning programs about the *EAL/D Learning Progression.*

(b) An additional project needs to be established to map the *EAL/D Learning Progression* against the new NSW syllabuses and the literacy continuum. Teaching resources should be developed to support class teachers to program and plan for EAL/D learners.

(c) The detailed feedback provided by teachers related to the instrument itself may be useful to ACARA as authors of the *EAL/D Learning Progression*.

7. Summary of Recommendations

Recommendations for NSW government schools

1 Implementation for resource allocation

1.1 Implementation of the *EAL/D Learning Progression* is recommended as a replacement for the ESL Phase tool currently used in NSW government schools. If adopted, the *EAL/D Learning Progression* will become the broad measure of English language proficiency used for EAL/D students, and will become the basis of the allocation formula used for the ESL funding component of the new Resource Allocation Model in NSW government schools.

1.2 Full implementation of the *EAL/D Learning Progression* is recommended for 2014 so that data can be collected and used in the resource allocation process for 2015.

1.3 An implementation plan is recommended that builds on the learning from the trial, and includes:

a) developing and conducting a program of professional learning in government schools during 2013 to prepare teachers to use the *EAL/D Learning Progression* phase assessments for all EAL/D students from the beginning of 2014

b) ensuring professional learning programs emphasise the importance of judging students' phases based on ongoing student assessment

c) developing resources to support teacher assessment using the *EAL/D Learning Progression* instrument, including samples of assessment tasks and annotated work samples demonstrating the evidence required to judge students against the *EAL/D Learning Progression* indicators

d) providing participants in professional learning programs with clear guidelines as to the purpose and potential uses of the *EAL/D Learning Progression*

e) developing and providing participants in professional learning programs with clear guidelines as to the process of deriving an overall phase judgement from the four language mode phase judgements for each EAL/D student

f) building capacity in the new student administration and learning management system (SALM) to allow teachers to enter EAL/D mode and phase data. Teachers should be able to update at any time the EAL/D phase judgements for individual students in SALM by 2014

g) revising school data collection tools to collect *EAL/D Learning Progression* phases in each mode for every EAL/D student. This would replace the current ESL phase data collection tools.

2 Classroom support

The *EAL/D Learning Progression* should be mapped against the new NSW syllabuses and the literacy continuum. On the basis of this mapping, teaching resources should be developed where appropriate to support classroom teachers to program and plan for EAL/D learners.

National recommendations

3 National trial of the *EAL/D Learning Progression*

A national trial is recommended to test the generalisability of the evidence from the NSW trial. This national trial should include:

- a broader range of student and teacher demographic groups, including larger cohorts of specific student groups, in particular Aboriginal students
- examination of differential reliability in teachers' judgements across different types of schools, students and language modes to inform future programs of teacher training and professional development
- recommendations about the potential use of the *EAL/D Learning Progression* to report English language proficiency across jurisdictions.

4 Detailed information to ACARA

It is recommended that ACARA is provided with this report of the NSW trial and with access to the detailed feedback provided by teachers related to the instrument itself. This may be useful to ACARA as authors of current and possible future versions of the *EAL/D Learning Progression*.

Final note

In conclusion, this trial has provided sufficient reliability and validity evidence for the *EAL/D Learning Progression* to be used in NSW government schools as a broad measure of English language proficiency for resource allocation. In addition, the NSW trial has national implications in terms of informing the design and the associated cost-benefit analysis of a prospective national trial. Such a national trial would provide recommendations about the potential use of the progression to report English language proficiency across jurisdictions.

References

Andrich, D. A. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Australian Curriculum, Assessment and Reporting Authority (2011). Guide to understanding ICSEA, ACARA, Sydney.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Brualdi, A. (2002). Traditional concepts of validity. **In** Rudner, L. and W. Schafer (2002). *What Teachers Need to Know About Assessment*. Washington, DC: National Education Association.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105.

Cohen, R. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Collier, V. P. (1995). Acquiring a second language for school. *Directions in Language and Education, 1 (4).* Washington DC. The National Clearing House for Bilingual Education.

Crick, G. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalised analysis of variance system* (ACT Technical Bulletin No 43). Iowa City, IA: American College Testing Program.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioural measurements: Theory of generalisability of scores and profiles*. New York, NY: John Wiley.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.*),* Schooling and language minority students: *A theoretical framework.* Evaluation, Dissemination and Assessment Center, California State University, Los Angeles.

Cummins, J. (1984). *Bilingualism and Special Education: Issues in Assessment and Pedagogy*. Clevedon: Multilingual Matters.

Daftaripard, P., & Lange, R. (2009). Theoretical complexity vs. Rasch item difficulty in reading tests. *Rasch Measurement Transactions, 2009, 23*(2), 1212–1213.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155–185.

Eckes, T. (2009). Many-facet Rasch measurement. In *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg, France: Language Policy Division, Council of Europe.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions* (TOEFL Research Rep. No. 70). Princeton, NJ: ETS.

Garcia, G (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?* NCBE Issue Brief No 5. national clearinghouse for bilingual education, September 2000. http://www.ncela.gwu.edu/pubs/issuebriefs/ib5.htm

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.), (pp.17–64). Westport, CT: American Council on Education.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26–43.

Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement. 48*(4), 399–418.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*(2), 103–122.

Linacre, J. M. (2008). *Facets Rasch model computer program* [Software program manual]. Chicago, IL: Winsteps.com.

Linacre, J. M. (2010). *A user's guide to Winsteps Ministep Rasch-Model computer programs* [Software program manual 3.70.0]. Chicago, IL:Winsteps.com

Messick, S. (1996). Validity of performance assessments. In G. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp.1–18). Washington, DC: National Centre for Educational Statistics.

Nunan, D. (2003). The impact of English as a global language on educational policies and practices in the Asia-Pacific region*. TESOL Quarterly 2003, 37*(4), 589–613.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago,

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modelling. *Language Testing*, *22*(1), 1–30.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: a primer*. Newbury Park, CA: Sage.

Thomas, W.P., & Collier, V. P. (2002). *A National Trial of School Effectiveness for Language Minority Students' Long-Term Academic Achievement.* UC Berkeley: Center for Research on Education, Diversity and Excellence.

Viera, A., & Garrett, J. M. (2005). Understanding interobserver agreement: The Kappa statistic. *Family Medicine, 37*(5), 360–363*.*

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Wright, B. D., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.

Appendices

Appendix A ESL LEARNER PHASES

In NSW government schools, ESL students are identified as being in one of three phases of acquiring English as a second language. This assessment is made by the ESL teacher on enrolment and on an ongoing basis. The phase descriptions are used to report on ESL student needs in the *ESL Annual Survey* and for determining priorities for allocating available ESL teacher support within the school.

First phase
First phase learners are students whose understanding and production of spoken or written English is obviously limited in all social and educational situations. These learners are acquiring *basic* English language proficiency and demonstrate *elementary functioning* in an English speaking classroom.

First phase learners range from complete beginners with minimal or no English to students who can communicate in English with limited fluency about familiar events, themes and topics related to their immediate personal experiences.

Some first phase learners may have studied English in their country of origin and have developed reading and writing skills but have negligible oral skills in English. First phase learners may also include students from language backgrounds other than English who were born in Australia and have had limited or no exposure to English prior to entering Kindergarten. Students at the end of first phase will have acquired various levels of literacy in English depending on factors such as age on entry and literacy in their first language.

Second phase
Second phase learners are those students whose understanding and production of spoken and written English is progressing, but is still limited to a range of familiar social and educational situations. These learners have *transitional* English language proficiency and demonstrate *partial and variable functioning* across the school curriculum.

Second phase learners range from students who have acquired a basic communicative repertoire in English which enables them to participate in some class activities to students who can communicate with some degree of confidence and coherence about subject matter appropriate to their age group but removed from their immediate personal experiences.

Students at the end of second phase will have made significant progress in their oral English language skills and will have been applying their English language and literacy skills to both formal and informal situations.

As a general guide, an ESL student will have moved beyond the second phase after a total period of three years instruction with ESL support.

Third phase

Third phase learners are students who generally function fluently and competently in English, but who occasionally need assistance in meeting the particular language and literacy demands of English in specific social and educational situations. These learners may demonstrate *apparent functioning* throughout the school curriculum.

Third phase learners range from students who have developed a transitional communicative repertoire in English which enables them to function in most language and literacy activities; to students who can communicate in English with confidence and clarity to a level approaching that of first language speakers about subject matter appropriate to their age group but unrelated to their direct personal experience.

Third phase learners may exhibit effective oral English communication skills but experience persistent barriers to successful completion of literacy tasks. Students at the end of third phase will normally have extended their English language and literacy skills in both formal and informal situations and be able to learn and participate effectively in the mainstream curriculum.

As a general guide, an ESL student will have moved beyond the third phase of ESL learning after a total period of seven years instruction with ESL support.

Appendix B Samples of training and support material

Identifying the EAL/D student group

Participants used this form during the workshop to list and consider students for possible inclusion in the trial prior to final student identification on the data collection website.

	Student name	M/F	Year group	ESL Phase	Refugee / International student	Aboriginal student
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

Overview of assessment practices

Participants referred to this document as a guide to the range of both formal and informal assessment tasks that could provide evidence of student work.

	Assessment by the teacher Formal ◄----------► Informal			Assessment by the student Formal ◄----------► Informal	
Ways of assessing	**Testing**	**Oracy & literacy tasks**	**Observation**	**Guided reflection**	**Personal reflection**
Examples of skills/tasks	Listening/reading comprehension Prepared oral presentation Dictation Reading unseen passage aloud Standardized tests such as Neale analysis End of unit tests Supervised writing Cloze activities Multiple choice items System-wide skills tests High stakes exams, eg. HSC,NAPLAN	Initial interview Debates Tasks within unit of work Research assignments Folios of student work Audio tapes of student interviews Video recording of role play Error analysis Dictagloss Read and retell Cloze passages	Discussions with specialist staff Chatting with students/staff Monitoring group participation Writing conferences Completing homework Attendance Student interaction	Student questionnaire Student self-evaluation Self-checking materials (eg. computer programs) Guided reflection Reviewing self/peers' activity on video Journal writing	Evaluating peers' writing Peer/teacher interaction Peer assessment
Forms of documentation	Individual reports Reports comparing schools and states	Student progress reports Outcomes based profiles School needs analysis Criterion rating scales	Task completion records Rolls Observational diary Anecdotal record/checklist	Results of questionnaires Contracts Collecting data	Personal journals Self-evaluation sheet for a unit of work

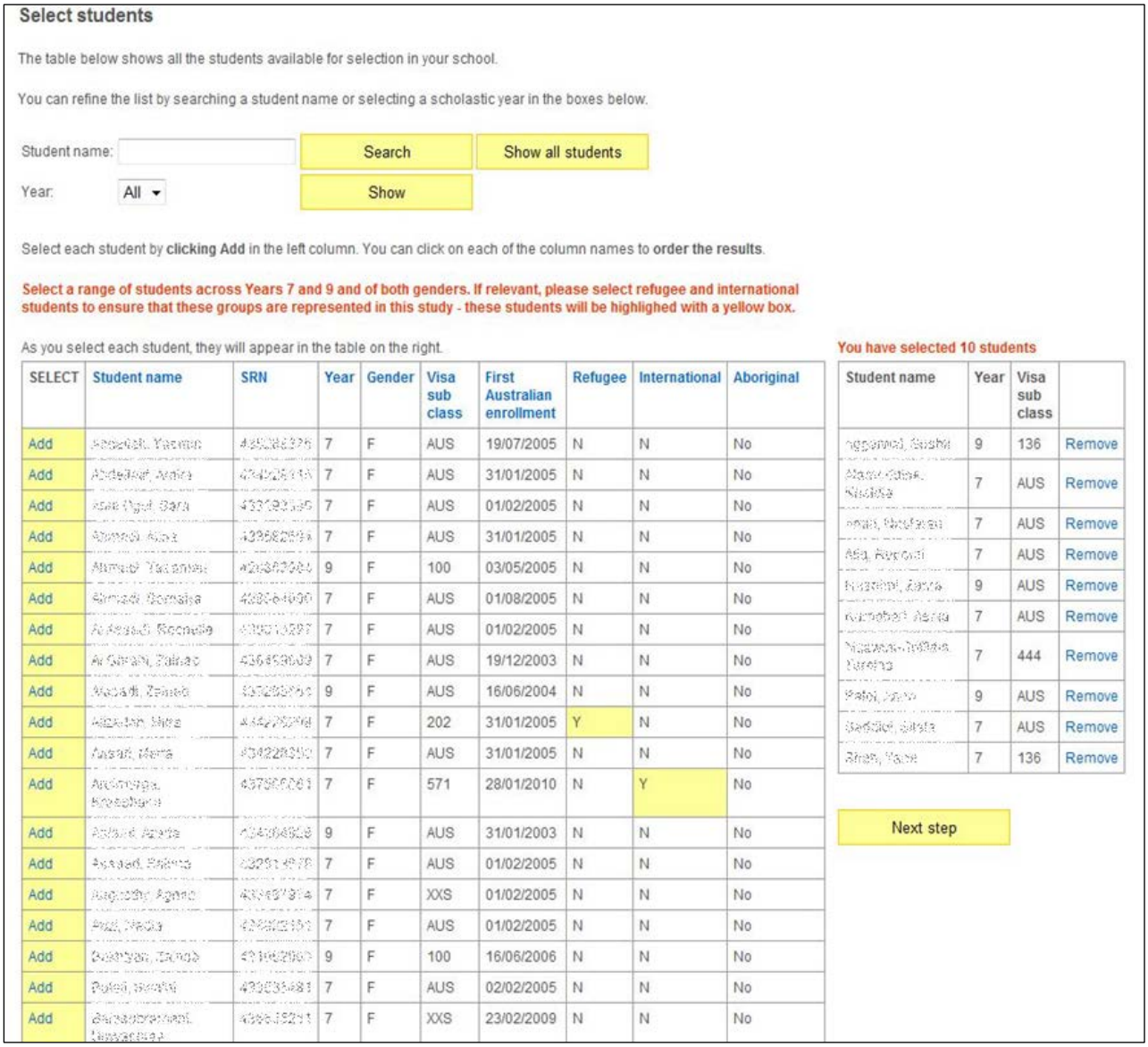
Adapted from NSW Department of School Education, (1994). School Based Training Course for Teachers of ESL students K-12. Student module, p.53.

Assessment matrix

Participants used this form to identify assessment tasks that were part of their current teaching program and could be used to collect evidence of student phases.

Task	KLA	What scaffolding was provided?	What was being assessed?	How could evidence be collected?

Appendix C Student selection screen



Select students

The table below shows all the students available for selection in your school.

You can refine the list by searching a student name or selecting a scholastic year in the boxes below.

Student name: Search Show all students

Year: All Show

Select each student by **clicking Add** in the left column. You can click on each of the column names to **order the results**.

Select a range of students across Years 7 and 9 and of both genders. If relevant, please select refugee and international students to ensure that these groups are represented in this study - these students will be highlighed with a yellow box.

As you select each student, they will appear in the table on the right.

SELECT	Student name	SRN	Year	Gender	Visa sub class	First Australian enrollment	Refugee	International	Aboriginal
Add	[redacted]	[redacted]	7	F	AUS	19/07/2005	N	N	No
Add	[redacted]	[redacted]	7	F	AUS	31/01/2005	N	N	No
Add	[redacted]	[redacted]	7	F	AUS	01/02/2005	N	N	No
Add	[redacted]	[redacted]	7	F	AUS	31/01/2005	N	N	No
Add	[redacted]	[redacted]	9	F	100	03/05/2005	N	N	No
Add	[redacted]	[redacted]	7	F	AUS	01/08/2005	N	N	No
Add	[redacted]	[redacted]	7	F	AUS	01/02/2005	N	N	No
Add	[redacted]	[redacted]	7	F	AUS	19/12/2003	N	N	No
Add	[redacted]	[redacted]	9	F	AUS	16/06/2004	N	N	No
Add	[redacted]	[redacted]	7	F	202	31/01/2005	Y	N	No
Add	[redacted]	[redacted]	7	F	AUS	31/01/2005	N	N	No
Add	[redacted]	[redacted]	7	F	571	28/01/2010	N	Y	No
Add	[redacted]	[redacted]	9	F	AUS	31/01/2003	N	N	No
Add	[redacted]	[redacted]	7	F	AUS	01/02/2005	N	N	No
Add	[redacted]	[redacted]	7	F	XXS	01/02/2005	N	N	No
Add	[redacted]	[redacted]	7	F	AUS	01/02/2005	N	N	No
Add	[redacted]	[redacted]	9	F	100	16/06/2006	N	N	No
Add	[redacted]	[redacted]	7	F	AUS	02/02/2005	N	N	No
Add	[redacted]	[redacted]	7	F	XXS	23/02/2009	N	N	No

You have selected 10 students

Student name	Year	Visa sub class	
[redacted]	9	136	Remove
[redacted]	7	AUS	Remove
[redacted]	7	AUS	Remove
[redacted]	7	AUS	Remove
[redacted]	9	AUS	Remove
[redacted]	7	AUS	Remove
[redacted]	7	444	Remove
[redacted]	9	AUS	Remove
[redacted]	7	AUS	Remove
[redacted]	7	136	Remove

Next step

Appendix D *EAL/D Learning Progression* Assessment Record

EAL/D Learning Progression Assessment Record

Teacher:_____ _____ School: _____

Student details

Family name		**Given name/s**				**Year**	
Date of birth		**Age**		**Gender**		**Country of birth/origin**	

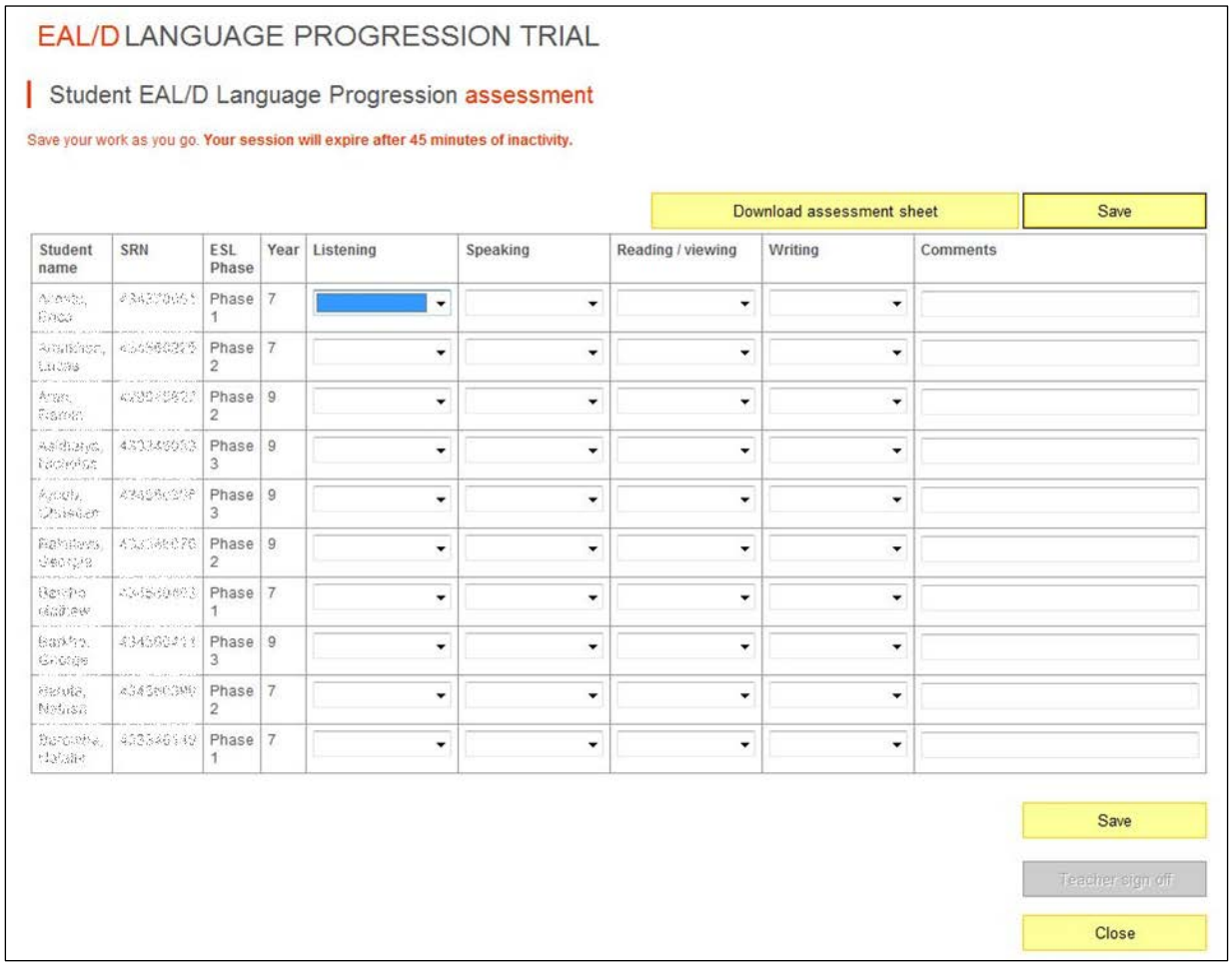
Languages other than English spoken					
Level of literacy in languages other than English (able to speak, read and/or write?)					
Australian citizenship details or Visa class *(letters)* **and sub-class** *(numbers)*	**Visa class**		**Visa sub-class**		**Date of arrival in Australia**	

Assessment task details

KLA:	**Language mode/s:** *Listening / Speaking / Reading / Writing*
Assessment task:	**Assessment task criteria:**
Assessment conditions: *Formal / Informal* *Individual / pair / group* *Planned / spontaneous language use* *Other:*	**Scaffolding provided / prior learning:** *Student access to resources (dictionary, text models)* *Language modelling and practice prior to task*
Student can do well:	**Student is working towards:**
Relevant EAL/D learning progression criteria:	**Relevant EAL/D learning progression criteria:**

EAL/D learning progression phase:				
Beginner Some print literacy	***Beginner Limited literacy background***	***Emerging***	***Developing***	***Consolidating***

Appendix E *EAL/D Learning Progression* assessment data collection web screen



Appendix F Annotated work sample used for double marking

EAL/D Learning Progression Assessment Record

Teacher: [redacted] School: [redacted]

Student details

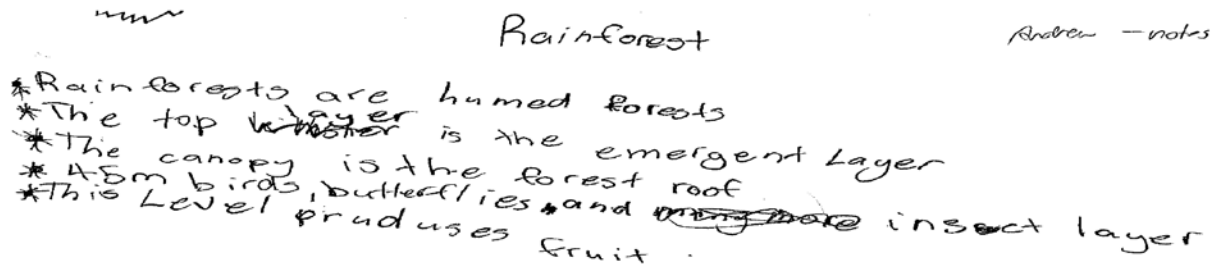
Family name [redacted]		Given name/s [redacted]		Year 5	
Date of birth 14/12/2001	Age 10yrs	Gender M	Country of birth/origin Australia		
Languages other than English spoken	Greek				
Level of literacy in languages other than English (able to speak, read and/or write?)	nil				
Australian citizenship details or Visa class *(letters)* and sub-class *(numbers)*	Visa class	Aust C – Aust Citizen.	Visa sub-class	Date of arrival in Australia	—

Assessment task details

KLA: English / Science	**Language mode/s:** Listening / Speaking / Reading / ~~Writing~~
Assessment task: dictogloss + text reconstruction	**Assessment task criteria:** • listen to text read aloud • 2nd reading – take notes whilst listening • 3rd reading – add detail to notes • reconstruct text using notes.
Assessment conditions: ~~Formal~~ / Informal Individual / pair / group ~~Planned~~ / spontaneous language use Other:	**Scaffolding provided / prior learning:** Student access to resources (dictionary, text models) Language modelling and practice prior to task – discussion of complex /technical vocab in the text – related to class work on COGS unit 'Ecosystems' thus known/familiar content
Student can do well: • take basic notes • construct grammatically correct sentences that make sense • appropriate punctuation • identify some key ideas	**Student is working towards:** • more detailed notes using abbreviations rather than full sentences • complex sentences • extract more detail from a text.
Relevant EAL/D learning progression criteria: • follow simple instructions + recognise familiar words in spoken texts • demonstrate understanding of short spoken texts • respond to key words • demonstrate listening behaviours • use subject–verb agreement with accuracy • use present tense verbs	**Relevant EAL/D learning progression criteria:** • give relevant details • understand vocab beyond immediate experiences • take notes • use paragraphs, including topic sentences • use conjunctions + relative pronouns to make compound + complex sentences

EAL/D learning progression phase:

Beginner Some print literacy	Beginner Limited literacy background	Emerging Listening	Developing Writing	Consolidating

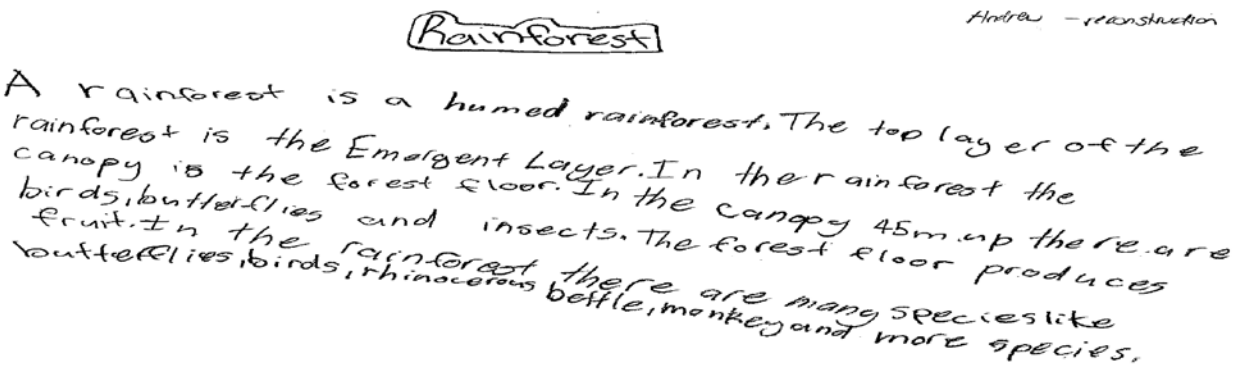


Rainforest

Andrew – notes

*Rainforests are humed forests
*The top layer is the emergent Layer
*The canopy is the forest roof
*45m birds, butterflies and insect layer
*This Level prud uses fruit.

Writing sample – first draft



Rainforest

Andrew – reconstruction

A rainforest is a humed rainforest. The top layer of the rainforest is the Emergent Layer. In the rainforest the canopy is the forest floor. In the canopy 45m up there are birds, butterflies and insects. The forest floor produces fruit. In the rainforest there are many species like butterflies, birds, rhinoceros beetle, monkey and more species.

Writing sample – final draft

Appendix G Online teacher survey

Teacher online survey – *EAL/D Learning Progression* evaluation

Section One: Participant details

1. Current position in school (select one):
    a. Class teacher
    b. ESL teacher
    c. Executive non-teaching
    d. Executive teaching
    e. Other specialist teacher

2. Did your pre-service teacher training include Teaching English to Speakers of Other Languages (TESOL) or English as a Second Language (ESL)? (Yes/No)

3. Do you have postgraduate qualifications in TESOL or ESL? (Yes/No)

4. Do you use the ESL Scales to assess ESL students? (yes/no/ sometimes)

5. If you have used the ESL Scales, how confident are you in using them?

    Not confident 1 2 3 4 **Very confident**

Section Two: The *EAL/D Learning Progression* trial process

6. How confident were you in using the *EAL/D Learning Progression*?

    Not confident 1 2 3 4 **Very confident**

7. The training provided sufficient information about the *EAL/D Learning Progression*.

    Agree 1 2 3 4 **Disagree**

8. The training provided sufficient and clear information about the trial process.

    Agree 1 2 3 4 **Disagree**

9. Teachers were adequately supported in the trial process identifying and assessing students, recording phases and uploading data.
    Agree 1 2 3 4 **Disagree**

10. Sufficient time was provided to complete the assessment and uploading of data.
    Agree 1 2 3 4 **Disagree**

11. Please provide any additional comments regarding the *EAL/D Learning Progression* trial process.
    (free text)

Section Three: The *EAL/D Learning Progression* usability

12. Overall, the *EAL/D Learning Progression* is a useful tool for assessing English language progression.

    Agree 1 2 3 4 **Disagree**

13. Which format of the *EAL/D Learning Progression* did you use to assess students' phase?
    a. complete document including all stages of schooling
    b. EAL/D learning Progression age appropriate stage only
    c. EAL/D learning progression by mode
    d. your own format

14. Did you use the Assessment Recording Sheet? (Yes/No)
    How useful was it?

    Not useful 1 2 3 4 **Very useful**

15. Did you use the Individual Student Summary Sheet? (Yes/No)
    How useful was it?

    Not useful 1 2 3 4 **Very useful**

16. The collection of work samples and recording of student observations were essential for making a phase judgement.

    Agree 1 2 3 4 **Disagree**

17. How many work samples / observations per student did you use to make an assessment for each mode? (free response)

    Listening 1 2 3 4

    Speaking 1 2 3 4

    Reading 1 2 3 4

    Writing 1 2 3 4

18. Approximately how did it take to assess each student per mode?

    Listening <15 mins 15 – 30 mins >30 mins

    Speaking <15 mins 15 – 30 mins >30 mins

    Reading <15 mins 15 – 30 mins >30 mins

    Writing <15 mins 15 – 30 mins >30 mins

19. Would you like to see the *EAL/D Learning Progression* replace the current ESL Phases? (yes / no)

20. What other uses do you see for the *EAL/D Learning Progression* in your school? (free response)

21. Which of the following resources or professional learning would support the implementation of the *EAL/D Learning Progression* in your school?
    e. Face to face professional learning
    f. Online tutorial
    g. Written advice
    h. Sample teaching resources linked to the EAL/D learning progression
    i. Annotated work samples linked to EAL/D learning progression
    j. Other

22. Please provide any additional comments regarding the *EAL/D Learning Progression* usability.
    (free text)

Section Four: The *EAL/D Learning Progression* feedback

23. The four phases are adequate for a broad classification of English Language Learning progression.

    Agree 1 2 3 4 **Disagree**

24. Would you be able to use the *EAL/D Learning Progression* to decide on an overall phase of an ESL learner? (Yes/ No)

    If yes, which of the following processes could be used to calculate an overall phase:
    k. Teacher judgement based on *Characteristic of the Learner* statements
    l. A formula (eg: the student's overall phase is calculated based on the average of the four separate phases submitted for each student)
    m. A teacher judgement based on specific guidelines (eg: the student's overall phase is the lowest phase recorded for that student)
    n. Other: …..

25. The phase labels (beginning, emerging, developing, consolidating) clearly describe four broad stages of English language progression.

    Agree 1 2 3 4 **Disagree**

26. There is a clear English language learning progression across the phases (within each stage).

    Agree 1 2 3 4 **Disagree**

27. There is alignment across modes (listening, speaking, reading and viewing and writing) within each phase.

    Agree 1 2 3 4 **Disagree**

28. The *Characteristic of the learner phase statements* are clearly expressed with an appropriate level of detail.

    Agree 1 2 3 4 **Disagree**

29. The *Mode statements* are clearly expressed with an appropriate level of detail.

Agree 1 2 3 4 **Disagree**

30. The *indicators* are clearly expressed with an appropriate level of detail.

Agree 1 2 3 4 **Disagree**

31. Do you think the *EAL/D Learning Progression* is appropriate for use with Aboriginal students?

(yes/no/unsure)

32. Do you think the *EAL/D Learning Progression* is appropriate for use with refugee students with limited literacy background and/or disrupted education?

(yes/no/unsure)

33. Please provide any additional comments regarding the *EAL/D Learning Progression*.

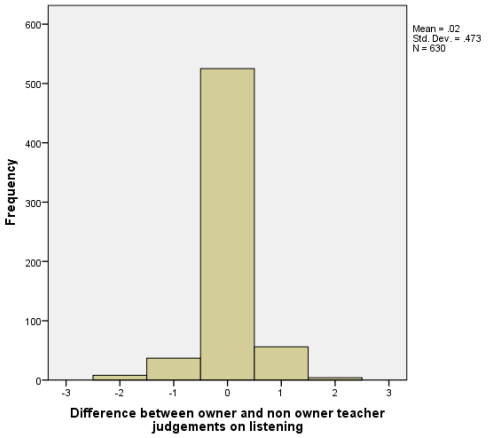
(free text)

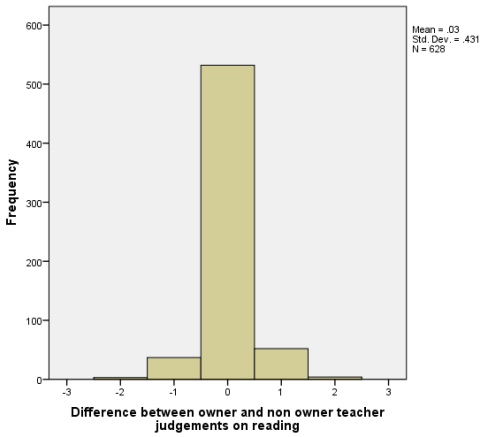
Appendix H Agreement rates (based on five EAL/D levels) for reading and writing

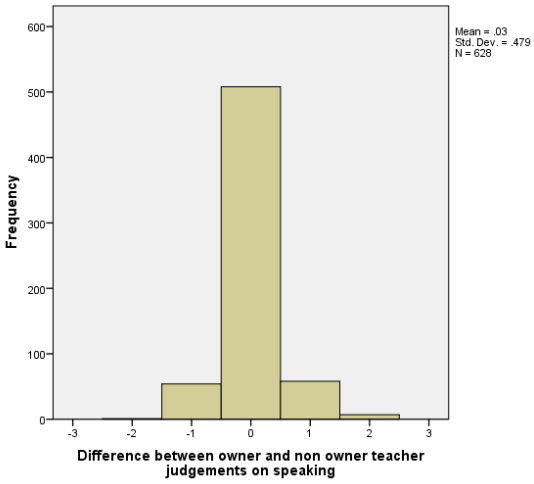
	Exact Agreement Rate		Adjacent Agreement	
Student Groups	**Reading**	**Writing**	**Reading**	**Writing**
All Students	81.2%	81.6%	97.6%	98.1%
Female	81.1%	83.2%	98.3%	99.0%
Male	81.3%	80.1%	97.0%	97.3%
Aboriginal	83.3%	94.4%	100.0%	100.0%
ESL Phase 1	89.5%	85.8%	98.2%	98.2%
ESL Phase 2	81.0%	79.9%	97.2%	97.6%
ESL Phase 3	68.5%	76.4%	97.2%	98.6%
Kindergarten	83.2%	81.3%	94.7%	94.8%
Year 3	79.3%	87.3%	97.3%	99.1%
Year 5	80.9%	83.5%	96.5%	99.1%
Year 7	82.0%	82.0%	99.3%	98.0%
Year 9	80.9%	75.9%	98.7%	98.7%
Refugee Students	85.4%	87.4%	99.3%	98.0%
International Students	92.9%	71.4%	100.0%	100.0%
Intensive English Centre Students	96.3%	87.9%	100.0%	100.0%

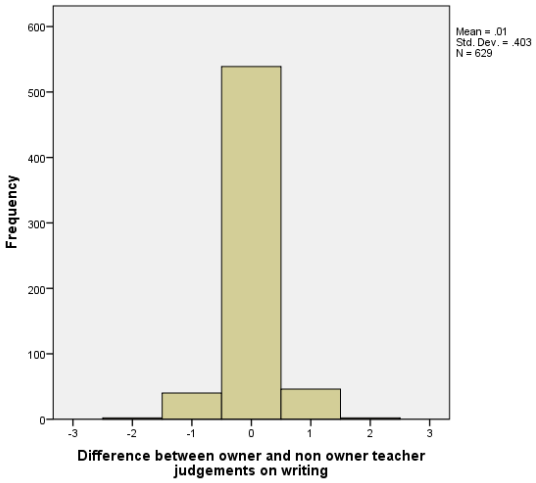
Note: Total number of students included in the double-marking process: 639.
The agreement rates are based on teachers' judgements using five EAL/D phases, with teachers' judgements on Beginning level students further separated into Beginning with limited or no prior print literacy (BLP) and Beginning with some prior print literacy (BSP).

Appendix I Histograms of the differences in the matched pairs of owner and non-owner teacher judgements









Note: For the above histograms, a -1 on the x-axis means that owner teacher judgement is one phase lower than the matched non owner teacher judgement. Conversely, a +1 on the x-axis means that owner teacher judgement is one phase higher than the matched non owner teacher judgement.

Appendix J Demographic statistics of the students included in the double-marking

		No of students	% of all students
All Students		639	
Gender	Girls	302	47.3%
	Boys	337	52.7%
ESL Phases	Phase 1	220	34.4%
	Phase 2	251	39.3%
	Phase 3	150	23.5%
Grade	Kindergarten	96	15.0%
	Year 3	112	17.5%
	Year 5	116	18.2%
	Year 7	154	24.1%
	Year 9	161	25.2%
Aboriginal Students		18	2.8%
Intensive English Centre		109	17.1%
Refugee		151	23.6%
International Student		14	2.2%

Appendix K Counts of students in each EAL/D phase for each grade cohort

Mode	EAL/D Phases	Kindergarten	Year 3	Year 5	Year 7	Year 9
LISTENING	Beginning	38	9	6	24	11
	Emerging	83	43	27	104	99
	Developing	25	84	71	77	81
	Consolidating	3	23	57	24	40
SPEAKING	Beginning	43	8	6	21	15
	Emerging	75	41	23	102	85
	Developing	26	89	83	78	86
	Consolidating	5	21	49	28	45
READING	Beginning	65	7	3	34	25
	Emerging	75	57	34	105	102
	Developing	7	79	71	78	75
	Consolidating	2	16	53	12	29
WRITING	Beginning	84	11	6	34	20
	Emerging	59	79	42	125	112
	Developing	5	63	73	59	71
	Consolidating	1	6	40	11	28

Note: The data used for the Rasch analysis only includes students who were assessed on all four language modes.

Further Information

For further information relating to the report please contact:

Dr. Lucy Lu
Statistics Unit
Centre for Education Statistics and Evaluation
Office of Education
NSW Department of Education and Communities

T 9561 8691
F 9561 8055
www.dec.nsw.gov.au/cese

Margaret Turnbull
Multicultural Programs Unit
NSW Department of Education and Communities

T 9244 5340
F 9244 5381

GPO Box 33
Sydney NSW 2001
T 9561 8000
www.det.nsw.edu.au

NSW Department of Education and
Communities