Language Proficiency Test Validity and Reliability Report

(Formerly Community Language Teacher Test)

School of Education and Social Work
Sydney Institute for Community Language Education

This report was published by The University of Sydney

NSW | 2006|sydney.edu.au

Contents

EXECUTIVE SUMMARY

The NSW Department of Education K-6 Community Languages program began in 1981 with the appointment of 30 teachers to NSW government primary schools, but increasing demand has seen this program increase to 244 full-time-equivalent positions in 136 schools teaching 30 different languages. Ensuring the levels of language proficiency of teachers in the program has always been important. The first tests were introduced in 1981 and teachers were assessed in speaking, listening, reading and writing by Department of Education officers. In 1996 the test was tendered to universities. Teachers in the K-6 Community Languages Program were able to gain permanent positions once they passed the test.

In 2020, Sydney Institute for Community Language Education (SICLE) was contracted by the NSW Department of Education to develop the **Community Languages Teacher Test (CLTT)** in up to 30 languages to assess the vocational proficiency of teachers. The task involved assembling a group of highly skilled and qualified languages experts to develop the tests and assess teachers. We worked with the examiners in locating and writing test items, preparing them for the tests and moderating marking. The lack of evidence for the reliability and validity of the CLTT was evident from the beginning. In 2023 SICLE was funded by the NSW Department of Education, Curriculum Early Years and Primary Unit for a study to collect reliability and validity evidence for the test. This report summarises the findings of the study.

The data set consisted of 149 test candidate results in the areas of Reading, Writing, Listening and Speaking and Reading Aloud plus the area of Cultural Competence. In addition, test rubrics and examination papers along with candidate test samples were examined. The final source of data involved three languages experts for the main test languages, Arabic, Chinese and Korean. Each responded to a set of 55 questions on the test validity (construct, content and criterion) and alignment with International Second Language Proficiency Rating (ISLPR) and NESA requirements. There was then a focus group discussion and the academics submitted any changes to their responses following the discussion. The data were subjected to a range of tests for reliability and validity. The mitigation to threats to reliability was also examined. The procedures, observations and analyses of data presented in this report provide strong support to the contention that the Community Language Teacher tests are both reliable and valid. The study has led to changes being made in the test to address some minor issues emerging from the study. We thank the NSW Department of Education for funding this study.

Background

The K-6 Community Languages program began in 1981 and now involves 244 full-time-equivalent positions in 136 schools teaching over 42,000 children in 30 different languages. Originally established to cater for background speakers, the program also encourages the inclusion of non-background speakers, thereby fostering and celebrating harmony and diversity. The main languages taught are Arabic, Chinese, Greek, Italian, Korean and Vietnamese, but there are also many new and emerging languages such as Bengali, Dari, Nepali and Samoan. Students generally receive two or more hours of language teaching each week.

The first proficiency testing was introduced in 1981 and teachers were assessed in speaking, listening, reading and writing by Department of Education officers. In 1996 the test was tendered to universities and between 1996 and 2015 the tests were developed and administered by UNSWIL at the University of NSW. Sydney University (SICLE) gained the tender for the test in 2020. SICLE's task involved assembling a group of highly skilled and qualified languages experts to develop the tests and assess teachers. We worked with the examiners in locating and writing test items, preparing them for the tests and moderating marking. The test was developed to align with Level 3 *Basic Vocational Proficiency* in the International Second Language Proficiency Rating Scale (ISLPR) (Ingram, 2005, 2009). The test also aligned with the NSW Education Standards Authority Subject Content Knowledge Requirements document (NESA July 2018): in particular, Standard 2 for Languages. For DET accreditation and also for NESA accreditation purposes, both *Superior* and *Acceptable* results meet the minimum discipline study as listed in the NESA Subject Content Guidelines (NESA 2018). There are now test materials and assessment criteria for Arabic, Chinese, French, German, Greek, Hindi, Indonesian, Italian, Japanese, Korean, Samoan, Spanish, Tamil, Turkish and Vietnamese. A total of 180 candidates have sat for the CLTT and 179 have gained acceptable standard. The reason for such a high pass rate is that candidates are generally background speakers of the language with high levels of proficiency.

When SICLE took over the testing in 2020, there was concern at the lack of documentation of the test records or development from previous iterations. It was stated in one document that the test was based on Level 3 Minimum Vocational Proficiency of the International Second Language Proficiency Rating Scale (ISLPR) but there was no evidence of this. We contacted ISLPR in Brisbane and they had no record of ever having been involved in the test development or administration. Contact was made with the former manager of the program in the DOE and through her we managed to gain copies of past CLTT tests.

Description of the test

As of January 2024 the Language Proficiency Test (formerly Community Language Teacher Test) is offered in the following languages:

Arabic
Greek
Japanese
Spanish
French
Hindi
Korean
Tamil
Chinese (Mandarin)
Indonesian
Samoan
Turkish
German
Italian
Vietnamese

Tests are currently being developed in Macedonian, Hebrew, Punjabi, Khmer, Persian and Auslan. Nine of the examiners are academics/ researchers at the University of Sydney and three at other universities. Four languages not taught at tertiary level; examiners are experienced teachers with tertiary education in the language.

Candidates register with the NSW Department of Education and their names and contact details are then forwarded to SICLE. Candidates are then sent the handbook of instructions and an online or face to face test is organised.

The writing component consists of three tasks to be completed within 60 minutes and 10 minutes reading time. The first task is a formal letter to parents about an issue of concern; the second task is an informal email to a colleague; and the third task is writing a picture story for students (pre-2022) or writing an explanation about a photo depicting cultural events (post 2022). The reading component consists of two texts with questions (to be answered in English or the community languages) asking about the possible audience, location (type of text) and purpose of the text. Candidates must then write a summary of the text in the community language or English (without using any of the original text or translation). Half an hour is given for this component. The speaking/ listening consists of an introductory conversation which is not assessed but which is intended to put candidates at ease. The original test consisted of two role plays for which candidates were given a role play card and time to prepare. This was changed in 2022 to candidates being given a topic for a formal talk to a group of teachers. They had time to prepare and then gave a talk as if to the group. The examiner then engaged them in conversation with several questions about the talk. The final assessment task is a read aloud text. The scoring rubric applied to each of the skill areas is as follows:

- Band A Superior. The level is higher than that of a completed major in the target language.
- Band B Acceptable (Major/ Minor). The level is equivalent to that of a University of Sydney student who has successfully completed at least 4 or 6 (post-beginner) language units in the target language.
- Band C Below Standard. The level is not equivalent to that of a completed major or minor in the target language.
- Band D Very Below Standard. Unacceptable, displays little expertise in the component.

Development of the test

The test was developed in 2018/2019 in Arabic, French, Chinese (Mandarin), German, Greek, Indonesian, Italian, Japanese, Korean, Spanish and Vietnamese by academics at NSW universities. The process was co-ordinated by Dr. Nerida Jarkey. Languages specialists were first asked about the levels of proficiency expected of tertiary students who had completed two or three years post-beginner study of the language. Work samples and examination questions were collected. There were then discussions where these levels were aligned with

the International Second Language Proficiency Rating Scale
(ISLPR) and the Common European Framework of Languages (CEFR). From the detailed descriptors of these assessment scales in English and relevant languages, the academics reached agreement on alignment with Level B2 on CEFR and Level 3 on ISLPR.

The second step was the development of the test specifications for writers. Drawing on available test information (Wylie & Ingram, 1999; North, 2005, 2007, 2016), a detailed set of specifications for listening, speaking, reading and writing was developed. Language academics made changes for each specific language where necessary. The length of response required and issues such as differences between formal/ informal registers in each language were to take into account. The final step was the examination of existing tests/ assessments that academics were using. Many of these assessments were deemed to be assessments of content knowledge contained in the university programs rather than proficiency per se. For example, there were many sets of questions or tasks relating to knowledge of grammatical features and vocabulary. It was decided that the assessment of grammatical features was best dealt with in the rubrics to be developed for listening, speaking, reading and writing. This was also the case for the assessment of cultural knowledge which was seen in the rubrics as cultural appropriateness of responses.

The list of test items was then agreed on. The test items reflected those used in the ISLPR and also used commonly in CEFR assessments. The final tests were then discussed in meetings between languages academics in order to reach consistency and coherence as far as possible across languages.

The present study

Description of data

Copies of all written components of the test and recordings of oral assessment are stored on a secure online portal. The first data set consisted of 95 tests covering 16 languages. Arabic (n = 19), Chinese Mandarin (n = 19), Korean (n = 17) and Vietnamese (n = 9) were languages which had the largest numbers of candidates. The assessment rubrics were changed in December 2022. The second data set consisted of 54 candidates who were assessed in four content domains: Reading Aloud, Reading Comprehension, Writing, and Speaking and Listening. Scores were recorded in the range 0 to 3 using the rubric for each component. The second data set involved candidates in one of the eight languages tested in that period. The majority of candidates participated in the tests of Arabic (n = 7), and Chinese (Mandarin n = 32), with Korean (n = 4).

Table 1: CLTT data from 2020 – May 2024

Year	Candidate number	Languages
2020	57	Arabic, French, Greek, Hindi, Indonesian, Italian, Japanese, Korean, Mandarin, Samoan, Spanish, Tamil, Turkish, Vietnamese
2021	22	Arabic, Greek, Hindi, Indonesian, Mandarin, Korean, Vietnamese
2022	15	Arabic, French, German, Italian, Korean, Mandarin, Tamil, Vietnamese
2023	53	Arabic, French, German, Hindi, Italian, Korean, Mandarin, Spanish, Tamil
2024 May	33	Arabic, Greek Mandarin, Spanish, Vietnamese
Total	180	

Table 2 Assessment Structure Post Dec 2022

Domain	Sub-domain	Max Score
Reading	Read - Audience and Purpose	3
	Read - Summarise main idea	3
Writing	Write - Convey meaning and purpose	3
	Write - Structure, Syntax and Vocabulary	3
Speaking	Speak - Read Aloud /Punctuation and intonation	3
	Speak - Participate in formal conversation	3
Listening	Listen - Recount key events	3
	Listen - Identify main idea and purpose	3
Cultural Understanding	Knowledge of authentic text	3
	Knowledge of history and culture	3

Three academics, language experts in Chinese, Korean and Arabic were involved in answering and discussing a range of questions relating to the test assessing construct, face and content aspects of validity. They were given a set of 55 questions covering listening/ speaking/ reading/ writing and cultural competence. The questions explored alignment with ISLPR 3 and issues of content, face and construct validity. Language experts were also given copies of the tests in their language along with marking rubrics and sample candidate completed tests (with marks and comments but name removed). In addition, they received a copy of the ISLPR descriptors. Academics returned their answers to these questions. A focus group interview was then conducted via zoom and recorded. This 90-minute discussion explored differences in responses and issues that emerged from the questions. Academics were then asked to revise and submit responses if they had changed their opinions because of the discussion.

Data Analysis - Reliability

Introduction

This section provides evidence regarding the aspects of reliability: **stability** reliability (how similar are results if students are assessed at different times?) and **equivalent forms** reliability (how similar are results if students are assessed with a different sample of equivalent tasks?). The third aspect of reliability, **inter-rater** reliability, is not included as each language (except for Hindi) only has one marker. Evidence is also provided using statistical calculations. The final section describes how reliability issues were addressed (Traub & Rowley,1991; Miller, Linn & Gronlund, 2009).

Stability Reliability

The data indicate that there is consistency in outcomes in the different tests administered at different times. The candidature that attempted these assessments is relatively homogeneous, being typically background speakers of the individual language being assessed. It can be assumed that the candidates of the Pre_Dec2022 tests are of the same general ability and background as those of the Post Dec2022 candidates. In addition, candidates who pass do not return to do similar assessments in future as happens in educational courses because the test is one of language proficiency. It can be assumed, therefore, that the level of consistency serves as evidence of stability reliability in the consistency of results of the test over time.

Equivalent Forms Reliability

Changes were made to the test and there are two different tests before and after December 2022. This is normal practice in many tests such as the NAPLAN. For example, each year a Year 3 mathematics test may have a different number of items, and a minor variation to the relative numbers of in different strands. There is no question, however, that the NAPLAN Year 3 Mathematics assessments over time represent equivalent test forms. Similarly, then, the two assessment structures of the CLTT Test can be considered to be equivalent test forms. The construct of the Post_Dec2022 forms may in fact be considered more robust in content than the Pre_Dec2022 tests. The similarity of results in two versions of the test show that there is no advantage nor disadvantage to candidates.

Statistical Evidence

Overall results from Cronbach's Alpha (α) and Rasch analyses provide strong statistical evidence for the reliability of the CLTT. Cronbach's alpha coefficient measures the internal consistency of a set of survey items. The statistic indicates the degree to which a collection of items consistently measure the same characteristic and these have been applied is evaluated on a standardized scale in the range 0 to 1. Values above 0.7 are considered in the acceptable range. The analyses conducted show that for both the test forms the Cronbach's α statistics are well into, and above, the acceptable range with values of 0.89 and 0.82 respectively.

The application of Rasch analysis to the data was appropriate as the Community Language Teacher Test conducted had a clear Model of Intent which was the estimation of candidates' proficiency in language. The assessment structure contains additional items that were considered consistent attributes to assess student proficiency in the acquisition of a particular language. The responses are consistent with the constructed modelling of a scale (language proficiency). Please see Appendix 1 for results from the Rasch analysis.

Addressing issues in reliability

Test reliability can be impacted by a number of factors which include physical and environmental aspects such as test length, test timing and consistency of test taking venues and procedures, issues with the difficulty of the items, the objectivity of the rating of candidate responses and the homogeneity of the candidates. All evidence is that these issues were addressed well in the conduct of the tests. There was no evidence, for example, in any of the data that students had not completed any task. Marks were awarded for each component for every student which would indicate an attempt had been made for that component. Tests were conducted at the Sydney Secondary College of Languages during Covid-19 to ensure that candidates could be properly distanced for reasons of safety. After that tests were conducted at Sydney University. Tests are now conducted online and face to face. The analyses conducted show that the scoring rubrics provide a range of possible scores and the full range has been applied in some instances. There is no evidence to suggest that the tasks are ambiguous or use unfamiliar language that may make items artificially difficult.

Markers were generally university researchers, graduates and researchers in their language. Where no academics were available, teachers with at least three years' tertiary study in the language and considerable experience in teaching the language were employed. Examiners were thoroughly briefed and moderation marking was conducted across languages. There was one case where some assessment issues were not clarified with the examiner of a specific language. In spoken communication in this language, it is normal for background speakers to mix English with the target language. In the test, the examiner had unrealistic expectations that candidates use only the target language and failed all candidates for using some English. These tests were then sent to two other markers whose results aligned with each other but differed from the original examiner. This issue was thus clarified. The conclusion is then that all aspects of test administration and marking addressed issues that may impact on reliability.

Data Analysis - Validity

Introduction

The evidence indicates that factors relating to validity have been satisfactorily addressed in the development of the assessments. There are no specific measures for validity (like Cronbach's α for reliability) but rather a collection of standards, statements and statistical analyses. Validity relates to the overall fairness of the assessments and the alignment between what the instruments are designed to assess and how well they achieve that aim. The commonly accepted requirement is that evidence must meets four types of validity: construct, content, face and criterion (Messick, 1989). **Construct** means that a language proficiency test must match current research knowledge of proficiency; **content** validity assesses whether the test assesses all aspects of the construct; **face** validity assesses how suitable the content is; and **criterion** evaluates how well the test can predict a concrete outcome or approximate the results of another test.

Construct Validity

The tests were within the conceptual framework of the International Second Language Proficiency Rating (ISLPR) which is supported by a strong tradition of research (Ingram, 2004, 2007). The CLTT was aligned with Level 3 (Minimum Vocational Proficiency) in the ISLPR. It was also aligned with B2 on the Common European Framework of Reference (CEFR). The structure of the CLTT also aligns directly with the Subject Content Knowledge Requirements document, released by the NSW Education Standards Authority (version July 2018) in particular, Standard 2 for Languages. In the development stages, university academics were asked to identify the linguistic skills/ knowledge and level of proficiency expected of students who had completed a major or minor in the target language. They were also asked to identify which level of the Common European Framework of Reference (CEFR) and Level in the ISLPR that students who be expected to gain. Researchers then collected and developed tasks suitable for students with a major or minor in the language. They attempted to identify which aspects of these tasks were assessing only program content and which could be seen to be assessing broader proficiency,

Table 3 ISLPR Level 3 descriptor

S:3, L:3, R:3, W:3	Basic 'Vocational' Proficiency	Able to perform effectively in a wide range of informal and formal situations pertinent to social and community life and everyday commerce and recreation, and in situations which are not linguistically demanding in own 'vocational' fields.	Some universities accept this as the minimum level for entry to undergraduate degree programmes.

Hence the CLTT has been designed to this construct, specifically, with this outcome in mind and the item writers have worked with an image of the kinds of skills a person would demonstrate to provide evidence of achieving the levels described in the ISLPR L3 standard, and, as mentioned previously.

Content Validity

The CLTT was benchmarked against other similar assessments such as the ISLPR, thus supporting the content validity. The CLTT assessments, drawing on the ILSPR Level 3, include:

- **Reading.** The candidate demonstrates the ability to comprehend texts in the target language including awareness of definitions and social context of language usage;
- **Writing.** Skills reflect appropriate syntax and use of words in order to communicate effectively both in formal and informal contexts. The writing indicates knowledge of the culture and the vocabulary appropriate to express this knowledge;
- **Listening and Speaking**. The candidate exhibits the ability to comprehend spoken text and language, as well as engage in receptive interactions and produce sustained oral text in a range of genres. The candidate also demonstrates a proficiency in using the relevant vocabulary, syntax and structures in the target language; and
- **Cultural Knowledge.** The candidate responds to authentic texts, both written and spoken, including, but not limited to, poetry, prose, drama, song, film or digital media. The candidate also demonstrates knowledge about culture and history, informing the social contexts of formal and informal language usage and communication.

This content structure aligns very closely with other similar assessments that are conducted nationally and internationally and the inclusion of Listening, Speaking and Cultural Knowledge domains is a feature that provides a higher level of validation.

Criterion Validity

Scoring rubrics were developed and standardisation workshops conducted to ensure that there is consistent application of the rubrics. Over time, exemplars of student work that is evidence of a particular level of achievement relative to the standard expressed in the rubric have been developed and distributed as reference materials for all examiners. Furthermore, these rubrics are in the public domain so that candidates are aware of the standard they are expected to achieve in order to 'pass' the CLTT in their language, and the various components of the test in which they will engage. These marking rubrics are effectively "standards statements" and as such do not vary over time. At the introduction of the Post Dec 2022 assessment structure new rubrics were developed to align with the new item tasks. These were developed in collaboration with experts across a number of languages. It can be confidently stated that criterion validity was addressed in that the assessment of candidate responses in each domain are consistently measured across time and between the languages.

Face Validity

The CLTT versions in each language were assessed for face validity by lecturers and experts in each language.

Expert Assessment of all Validity Types

In early 2024, a small-scale survey of language experts in the three most common languages (Arabic, Korean and Mandarin) was conducted to provide an external source of evidence of the components of Validity addressed above. The evaluations requested the participants to measure (a numeric value) and provide a free response comment upon a number of aspects of each domain within the CLTT assessments. There were specific questions relating to content, construct and criterion validity. A zoom focus group was then conducted where experts discussed their responses with researchers. A copy of the survey is attached as Appendix 2.

The survey measures were Likert-like with a code 3 indicating that the aspect was 'Very Appropriate', a code 2 – 'Somewhat Appropriate', and a 1 'Not Appropriate' as a contributor to the overall validity of the tests. Table 4 summarises the feedback on the components investigated. There was a general agreement that a Non-Compensatory score in excess of 80% of the maximum test mark was required for a candidate to achieve the required standard equivalent to ISLPR.L3. The non-Compensatory element was explicit that that minimum score applied to each component domain as well as the overall test.

There was a difference between responses to the survey and comments made in the focus group. This was because of confusion regarding the purpose of the test. The purposes stated in the briefing of experts were that the test should be equivalent to the levels of proficiency expected of tertiary graduates with a Major or Minor in the language and equivalent to Level 3 ISLPR. The experts (and one in particular) were also conscious that teachers passing this test could be teaching 'background' speakers of the language in Stage 6 (Year 11 and 12) study of the language where high levels of teacher language proficiency are required. In discussion all experts agreed that the proficiency level required for teaching Stage 6 was in fact beyond the level of Major/ Minor study required by accreditation authorities. This issue is explored in the following section which summarises responses from the focus group.

Reading

There was general agreement that the reading component was very appropriate to assess proficiency of teachers K-10. The Korean expert (KE) suggested that a wider range of genres to be selected for reading beyond narrative. All experts felt that candidates need to be tested on skimming and scanning skills and reading 'between the lines', that is for critical and inferential reading. The Arabic expert (AE) recommended poetry and literature in addition to informative texts so that candidates could demonstrate inferential reading.

Table 4 Summary of Language Expert Responses Sorted by Question Intent Domain

	Items	**Instructions**	**Time allowed**	**Assess expectations**	**How Well Assess Domain Levels**	**Marking Criteria Match with Standards Primary/Secondary Teacher**	**Alignment with Vocational Proficiency ISLPR L3**	**Pass equates to ISLPR L3**	**Assesses or Incorporates Cultural understandings**	**Minimum Score to Indicate ISLPR L3**
Score Rubric 1	Very Appropriate	Very Appropriate	Very Appropriate	Very Well	Very Well	Very Well	Very Well	Very Well	Very Well	Score
Score Rubric 2	Somewhat Appropriate	Somewhat Appropriate	Somewhat Appropriate	Somewhat	Somewhat	Somewhat	Somewhat	Somewhat	Somewhat	
Score Rubric 3	Not Appropriate	Not Appropriate	Not Appropriate	Not Well	Not Well	Not Well	Not Well	Not Well	Not Well	
Experts	Mean Scores	Mean Scores	Mean Scores	Mean Scores	Mean Scores	Mean Scores	Mean Scores	Mean Scores	Mean Scores	Mean Scores
Reading	2.7	2.7	not asked	2.3	2.0	2.3	2.3	2.7	2.0	>80% - 7/9
Writing _ email	2.3	2.3	2.7	2.7	3.0	2.3	3.0	3.0	Inc	>80% - 7/9
Writing _ Letter	2.3	2.3	2.7	2.3	2.7	2.3	2.7	3.0	Inc	>80% - 7/9
Writing _recount	2.3	2.3	2.7	2.3	2.3	2.3	2.0	Incomplete	Inc	>80% - 7/9
Listening	2.3	2.3	3.0	1.7	2.0	1.7	2.3	Incomplete	Inc.	>80% - 7/9
Speaking	2.3	3.0	2.3	2.3	2.0	2.3	2.0	2.3	Inc	>80% - 7/9

Writing

There was general agreement that the writing component was very appropriate to assess the proficiency of K-10 teachers. AE commented that the time for writing (700 words in 60 minutes) was too short. Researchers confirmed that the Arabic teachers had problems finishing the writing test in time. AE commented that 300 words of formal Arabic would be expected in that time. There are issues of diglossia in Arabic with great differences between spoken dialect and formal written standard Arabic. AE also commented that the writing tasks need to specify context more so that candidates would know if formal, semi-formal or more formal language was required: tasks 1 and 2 were overlapping requiring the same level of formality. Selecting the level of formality was also difficult in terms of the third writing task as the stimulus picture was not clear enough in terms of audience. KE liked the first two writing tasks for Korean because of the different levels of formality. For task 3 he recommended a choice of cultural event to write about because if candidates did not know the content of the specific cultural event they could not answer this task. The Chinese expert (CE) also liked writing tasks 1 and 2. She suggested a recount text type for Task 3. She also suggested the idea of a report for parents on a student misbehaving such as pushing over chairs in the classroom.

Listening

All experts felt the listening tasks were very appropriate for K-10 teachers. Again, the issue of listening to and comprehending more formal language for Stage 6 teachers emerged. KE reported that the Korean listening task was a radio interview. To assess more formal Korean, he suggested a text with more complex unfamiliar rhetorical structure that is different from the norms of real-life talk. They wanted a range of more difficult listening. AE agreed with this; She felt that the listening task was totally inappropriate as it was taken from an HSC test and involved two people having a discussion about doctors and medicine speaking in formal *al fus'ha*. The listening test for Arabic was thus totally inauthentic and inappropriate. She recommended using a media extract from *Al Jazeera* or a more authentic text in Arabic. They also commented that the listening text was too short and needed to be three to five minutes long. She also preferred the original marking sheets to the more recent ones as the original ones were more specific. CE said that the listening text, an interview, was appropriate but that there needed to be a question assessing cultural competence.

Speaking

All three experts agreed strongly with the role play in the speaking on topics tasks that the teachers would need to cope with. KE wanted an introductory conversation between the examiner and candidate which went for around five minutes. He suggested that instead of a prepared formal talk the speaking component be based on a summary presentation of the content of the listening test.

AE felt the audience needed to be made clearer so that the level of formality was clear. Whether to use MSA or dialect. The time for preparation of the talk was too short. The notion of interaction in the descriptors needed to be clearer. She also preferred the more explicit instructions in the old marking sheet.

Cultural Assessment

All experts felt that cultural competence should be assessed in all tasks and not just the third writing task. KE suggested it be assessed also in reading passages. CE added that cultural competence would also be assessed in the appropriateness of writing in Tasks 1 and 2. They also suggested that the marking criteria for each skills area needed one on cultural competence and appropriateness.

Overall, the survey and focus group provided strong evidence for the types of validity of the test. The comments have also been instrumental in making changes to improve future tests.

Summary and conclusions

The procedures, observations and analyses of data presented above provide strong support to the contention that the CLTT is both reliable and valid. The key questions were as follows:

- What's the evidence of the reliability of the test scores from the Community Languages Teachers' Test (CLTT)?

There is strong evidence for the reliability of the Community Language Teacher Test based on a study of many aspects of reliability.

- Is there sufficient evidence to support the intended interpretations and uses of the CLTT assessments?

There is strong evidence to support the intended interpretations and uses of the CLTT assessment based on findings from study of aspects of validity.

- What is the evidence of content validity? Do the tests measure what they are supposed to measure?

The content validity of the CLTT was judged to be strong given the alignment of the test with the ISLPR and similarity of items in a range of other tests.

- Is there sufficient evidence for alignment of the CLTT test with International Second Language Proficiency Rating Scale (ISLPR) Level 3 and Common European Framework of Reference (CEFR) Level BI, considered as minimum vocational proficiency?

Findings support the alignment of the CLTT test with the ISLPR level 3 and also the CEFR Level B1. In addition, the test aligns with NESA Subject Content requirements of being equivalent to two/ three years post-beginner tertiary study of the language.

To what extent this this study addressed the tender requirements? All deliverables were achieved except for two aspects: the first was the plan to interview teachers who had sat for the test. This proved to not be possible because of the ban on research in NSW government schools by the NSW Department of Education and we could thus not obtain ethic approval from SERAP in time. A second proposal involved the double marking of 2020/2021 tests. This was not done on advice from our researcher, Dr. Chris Freeman. The reason was that since there was a limited range of scores on the tests (with all candidates achieving A or B) it was decided that double marking would not provide any data that was not already available. With these two provisos, all deliverables have been achieved.

Following is a summary of the recommendations from languages experts which will be addressed in the 2024/5 revision of the tests. These will also be included in the version of the test being developed for additional languages.

1. The text types for reading should include an exposition, explanation or literary text requiring inferential comprehension and critical reading;

2. The requirements for words/ characters and time limit for writing need to be reviewed and made more realistic especially for Arabic.
3. Instructions for writing need to be more explicit in terms of audience and level of formality required.
4. The text type of the third writing task could be a recount and not the present writing an explanation of a cultural event.
5. The listening task should be authentic and more formal such as a TED talk in the community language,
6. One speaking task should be a summary presentation of the listening task.
7. One role play task in listening/ speaking should be reinstated
8. Marking rubrics should be more explicit.
9. Assessing cultural competence should be integrated across all components with specific rubrics for this. Cultural competence should be seen in its broader context embedded in communication and interaction.

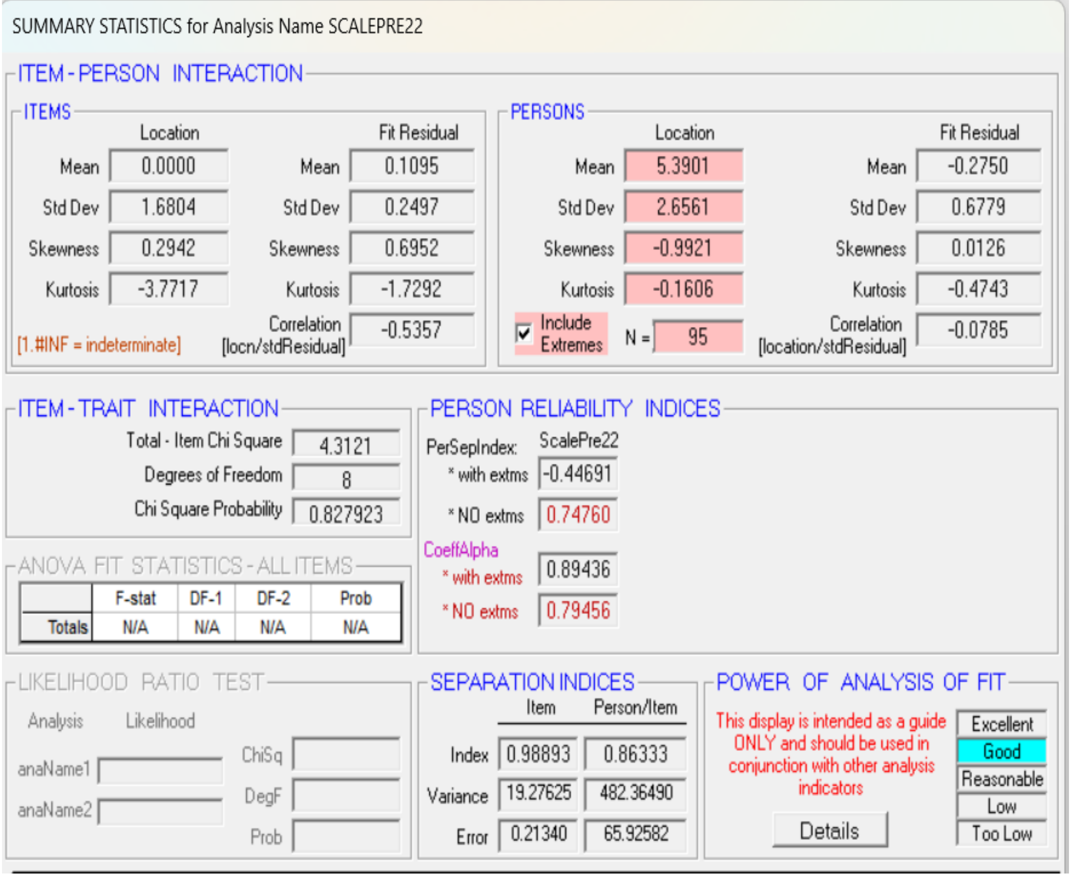
This report was prepared by the researcher, Dr. Chris Freeman and collated by Ken Cruickshank. Academic experts were Dr. Yong Moon Jung, Dr. Hongzhi (Veronica) Yang, and Dr. Nadia Selim. The Language Proficiency Test developed by SICLE is managed by Merryl Wahlin and Mark McCarthy. We acknowledge the support of the NSW Department of Education in funding this study and of Stephen Nogher, Languages Advisor K-6, Curriculum Early Years and Primary Learners.

Appendix 1
Rasch Analyses Tables

Rasch Analyses (Pre_Dec2022 Outputs)

Figure 1 Provides Summary Statistics for the analysis of the Pre-December 2022 candidates from the Rasch analysis program RUMM 2030 (Andrich et al).

Figure 1 Summary Statistics for the Pre-Dec 2022 Candidates



SUMMARY STATISTICS for Analysis Name SCALEPRE22

ITEM-PERSON INTERACTION

ITEMS

	Location		Fit Residual
Mean	0.0000	Mean	0.1095
Std Dev	1.6804	Std Dev	0.2497
Skewness	0.2942	Skewness	0.6952
Kurtosis	-3.7717	Kurtosis	-1.7292
[1.#INF = indeterminate]		Correlation [locn/stdResidual]	-0.5357

PERSONS

	Location		Fit Residual
Mean	5.3901	Mean	-0.2750
Std Dev	2.6561	Std Dev	0.6779
Skewness	-0.9921	Skewness	0.0126
Kurtosis	-0.1606	Kurtosis	-0.4743
☑ Include Extremes N =	95	Correlation [location/stdResidual]	-0.0785

ITEM-TRAIT INTERACTION

Total - Item Chi Square	4.3121
Degrees of Freedom	8
Chi Square Probability	0.827923

PERSON RELIABILITY INDICES

PerSepIndex:	ScalePre22
* with extms	-0.44691
* NO extms	0.74760
CoeffAlpha	
* with extms	0.89436
* NO extms	0.79456

ANOVA FIT STATISTICS - ALL ITEMS

	F-stat	DF-1	DF-2	Prob
Totals	N/A	N/A	N/A	N/A

LIKELIHOOD RATIO TEST

Analysis	Likelihood		
anaName1		ChiSq	
anaName2		DegF	
		Prob	

SEPARATION INDICES

	Item	Person/Item
Index	0.98893	0.86333
Variance	19.27625	482.36490
Error	0.21340	65.92582

POWER OF ANALYSIS OF FIT

This display is intended as a guide ONLY and should be used in conjunction with other analysis indicators

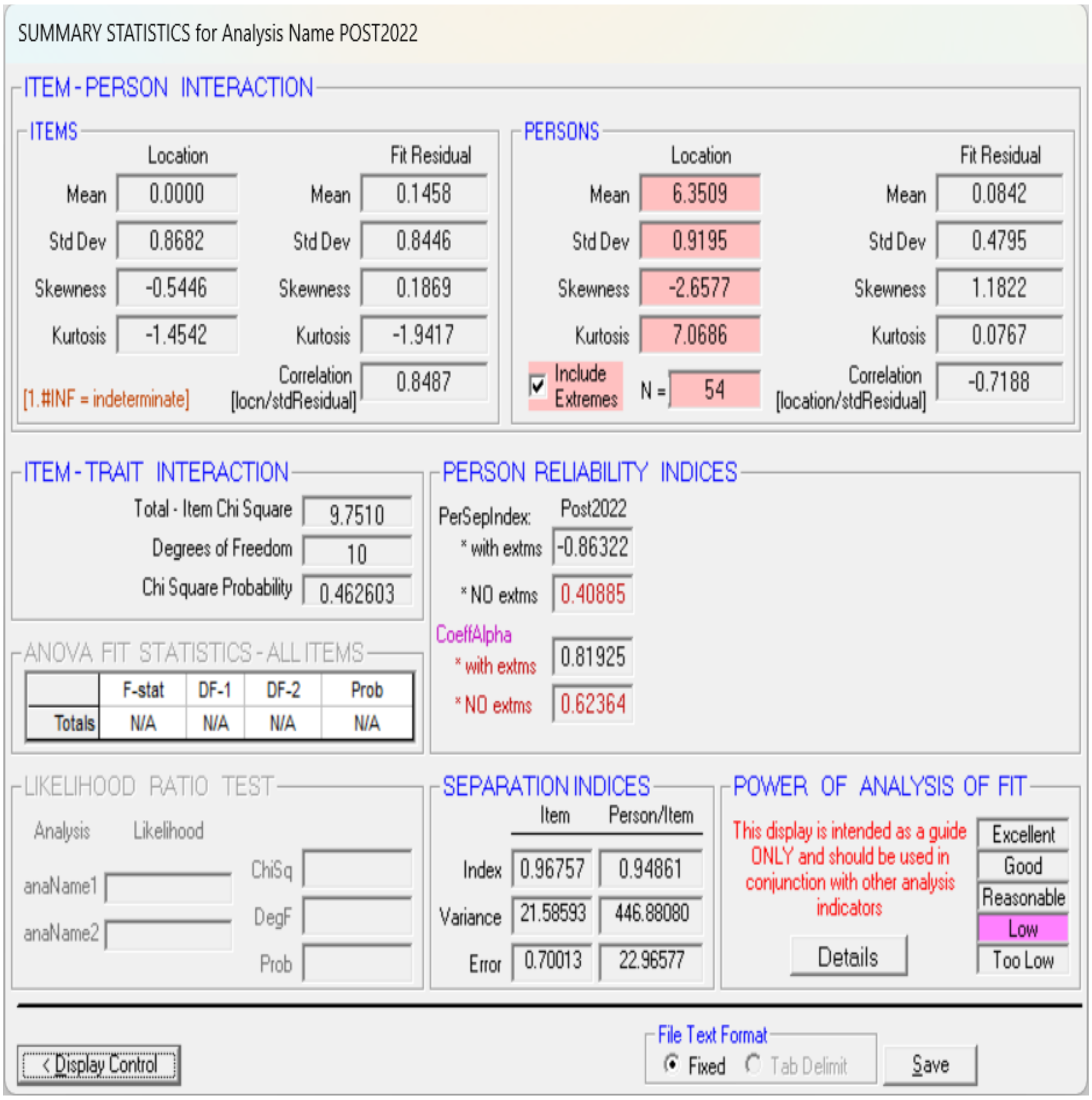
In summary, these statistics exhibited in Figure 1 indicate that the responses are consistent with the constructed modelling of a scale (language proficiency). This is indicated by the Mean of the Fit Residual which is relatively close to zero (an indication of perfect fit to the expected model); the Fit residual is a measure of how far the observed results deviate form that expected by the created model of language proficiency, and the Chi-Square Probability value of 0.828 in the ITEM-TRAIT INTERACTION which is a high indicator of the fit of the data to the model.

The Person Reliability Indices show a very low Person Separation Index (PerSepIndex) when extreme results are included. (Extreme results are perfect scores). The Person Separation Index measures the capacity of the instrument to discriminate between candidates. The low value is a function of the proportion of perfect scores in these data. By comparison the Cronbach reliability indice (α) is in the very good range with a value of 0.89. This measures the internal consistency of responses by students indicating the degree to which the responses of individual items correlate with the overall score. The mean location of the Persons (candidates) is 5.39. This indicates that students have very strong performances in their interaction with these items. (The mean of the items is zero by definition.)

Post_Dec2022 Outputs.

Analysis of the Post Dec2022 candidates was completed in the same manner to provide comparisons in outcomes and performances that may have resulted by the more robust assessment structure (five domains and two sub domains for each) and the application of the revised rubrics. Similar figures and tables are provided.

Figure 2 Summary Statistics for the Post-Dec 2022 Candidates



SUMMARY STATISTICS for Analysis Name POST2022

ITEM-PERSON INTERACTION

ITEMS

	Location		Fit Residual
Mean	0.0000	Mean	0.1458
Std Dev	0.8682	Std Dev	0.8446
Skewness	-0.5446	Skewness	0.1869
Kurtosis	-1.4542	Kurtosis	-1.9417
[1.#INF = indeterminate]		Correlation [locn/stdResidual]	0.8487

PERSONS

	Location		Fit Residual
Mean	6.3509	Mean	0.0842
Std Dev	0.9195	Std Dev	0.4795
Skewness	-2.6577	Skewness	1.1822
Kurtosis	7.0686	Kurtosis	0.0767
☑ Include Extremes	N = 54	Correlation [location/stdResidual]	-0.7188

ITEM-TRAIT INTERACTION

Total - Item Chi Square	9.7510
Degrees of Freedom	10
Chi Square Probability	0.462603

PERSON RELIABILITY INDICES

PerSepIndex:	Post2022
* with extms	-0.86322
* NO extms	0.40885
CoeffAlpha	
* with extms	0.81925
* NO extms	0.62364

ANOVA FIT STATISTICS - ALL ITEMS

	F-stat	DF-1	DF-2	Prob
Totals	N/A	N/A	N/A	N/A

LIKELIHOOD RATIO TEST

Analysis	Likelihood		
		ChiSq	
anaName1		DegF	
anaName2		Prob	

SEPARATION INDICES

	Item	Person/Item
Index	0.96757	0.94861
Variance	21.58593	446.88080
Error	0.70013	22.96577

POWER OF ANALYSIS OF FIT

This display is intended as a guide ONLY and should be used in conjunction with other analysis indicators

File Text Format ● Fixed ○ Tab Delimit

The overall pattern of summary statistics displayed in Figure 2 are very similar to those presented in Figure 1. The mean Fit residual of 0.146 is close to the perfect accord with the expected model statistic of zero. This indicates that these data fit the Model of Intent (a scale of language proficiency) quite well. Again, the Person Separation Indice is very low because these data do not enable a high degree of discrimination of ability between the candidates, again, a function of the proportion of perfect or near perfect scores. However, the reliability indice (α) is very good with a value of 0.819. The capacity of the RUMM application to analyse these data was limited by the sample size as indicated by the Power of Analysis of Fit. The candidates have a high mean ability estimate of 6.35 which indicates that even though the assessment may be more robust with the increase in items, the candidates display high levels of competence in these items.

Appendix 2
Survey questions

Reading

How appropriate do you think the reading texts and questions are for these groups of learners?
How useful/ appropriate are the instructions?
How well do you think the reading texts and questions assess these expectations of reading?
How well do the texts and questions address the range and level of reading required of primary/ secondary languages teachers?
How well do the criteria match the questions and texts?
How well do you think teachers the texts and questions align with this level description?
How well do you think applicants who pass the test would meet the requirements of Level 3 ISLPR?
The reading texts are also meant to address candidates' cultural understandings. How well do the texts do this?
What general comments do you have of the reading texts – their suitability, currency and appropriateness? Do you have any comments on how the choice of texts and questions could be improved?
In reflecting on the rubrics provided for the reading task what would you consider to be an appropriate score (out of 20) in Writing to indicate that a teacher has achieved the requirements of Level 3 ISLPR?

Writing

How appropriate do you think the writing tasks are for these groups of learners?
How useful/ appropriate are the instructions?
How appropriate do you think the time allocation for the completion of the Writing tasks are?
How well do you think the writing tasks assess this expectation of writing?
How well do the texts and questions address the range and level of writing required of primary or secondary school languages teachers?
How well do the criteria match the tasks?
How well do you think teachers the tasks align with this level description?
How well do you think applicants who pass the test would meet the requirements of Level 3 ISLPR?
The third task is also meant to address candidates' cultural understandings. How well does the task do this?
Would you prefer a picture/ photo stimulus or written question only?
What general comments do you have of the reading texts – their suitability, currency and appropriateness? Do you have any comments on how the choice of texts and questions could be improved?
In reflecting on the rubrics provided for the writing tasks what would you consider to be an appropriate score (out of 60) in Writing to indicate that a teacher has achieved the requirements of Level 3 ISLPR?

Listening

How appropriate do you think the listening text and questions are for these groups of learners?
How appropriate do you think the time allocation for the completion of the listening test is?
How useful/ appropriate are the instructions?
How well do you think the listening text assesses this expectation of listening?
How well do the text and questions address the range and level of listening required of primary/ secondary languages teachers?
How well do the criteria match the questions and text?
How well do you think the texts and questions align with this level?
How well do you think applicants who pass the test would meet the requirements of Level 3 ISLPR?
The listening text is also meant to address candidates' cultural understandings. How well do the text does this?
What general comments do you have of the listening text – its suitability, currency and appropriateness? Do you have any comments on how the choice of text and questions could be improved?
In reflecting on the rubrics provided for the Listening task what would you consider to be an appropriate score (out of 20) to indicate that a teacher has achieved the requirements of Level 3 ISLPR?

Speaking

How appropriate do you think the Task 1 speaking test and topics are for these groups of learners?
How appropriate do you think the time for the preparation and delivery of the Task 1 speaking test is?
How useful/ appropriate are the instructions for Task 1 speaking test?
How well do you think the Task 1 speaking test assesses this expectation of speaking?
How well does the Task 1 speaking test and questions address the range and level of speaking required of primary and secondary teachers?
The attached criteria are used to assess the Task 1 speaking test. How well do the criteria match the test?
How well do you think the Task 1 speaking test aligns with this level description?
How well do you think applicants who pass the Task 1 speaking test would meet the requirements of Level 3 ISLPR?
The Task 1 speaking test is also meant to address candidates' cultural understandings. How well do the test does this?
How appropriate do you think the Role Play 1 and 2 are for these groups of learners?
How useful/ appropriate are the instructions for Role Plays 1 and 2?
How well do you think the Role plays 1 and 2 assess this expectation of speaking?
How well does the Role Plays 1 and 2 address the range and level of speaking required of primary and secondary teachers?
How well do the criteria match the test?
The attached description is of level 3 ISLPR 'minimum vocational proficiency' which the test is meant to address. How well do you think the Role Plays 1 and 2 align with this level description?
How well do you think applicants who pass the Role Plays 1 and 2 would meet the requirements of Level 3 ISLPR?

Role Plays 1 and 2 are also meant to address candidates' cultural understandings. How well do these tasks do this?
In reflecting on the rubrics provided for the Speaking task what would you consider to be an appropriate score (out of 20) to indicate that a teacher has achieved the requirements of Level 3 ISLPR?

Reading aloud

How appropriate do you think the reading texts is?
How well do the criteria match the text?

General questions

The original test for K-6 teachers involve only the role plays – a separate listening test was not included and speaking/ listening was seen as a combined oral proficiency skill. To what extent is it important to have a separate listening test?
We have been discussing the differences between the role plays and the speaking test. Can you comment on the value of having both tests? Would you prefer one only? Would you suggest having both in the same test?