Latitude
Aviation English Services

Checkpoint

LANGUAGE ASSESSMENT FOR STUDENT SELECTION AND ADMISSIONS

Report on the performance of the listening and reading assessments

April 2024

Contents

Executive summary

- This report presents results of statistical analyses of the performance of the Checkpoint listening and reading assessments conducted in 2015 and 2024.

- Analyses conducted in 2015:

  - Estimated reliability to be between 90.2 and 100% with a standard error of measurement of between 1.7 and 3.2%;

  - Showed that it is impossible to make accurate judgements about listening ability or reading ability on the basis of a spoken performance; separate measurement of listening, reading and speaking skills is essential; and

  - Showed the listening and reading items to be discriminating well between candidates of low, medium and high ability.

- Analyses conducted in 2024:

  - Showed that the listening and reading items:

    - Cover a range of difficulty which is well-suited to the ability of the candidate population; and

    - Measure effectively with many items close to or at values ideal for measurement;

  - Resulted in revision of four of the 96 listening and reading items to further enhance their measurement properties.

1. Introduction

Checkpoint is Latitude's web-based assessment of English language proficiency for student pilot selection and admissions. Checkpoint is used by airlines, flight schools, colleges and universities to:

- Determine language proficiency for entry into primary flight training programmes; and
- Identify language training needs.

This report presents information about the consistency of measurement of Checkpoint's listening and reading skills assessments. It begins by setting out the importance of listening and reading skills for student pilots and the requirement for robust measures of the same. It then briefly explains how listening and reading is assessed in Checkpoint before describing results from statistical analyses conducted shortly after the assessment was launched in 2015. The report then goes on to describe analyses of item response data from candidates who took Checkpoint between 2017 and 2023 which provides empirical evidence for the quality of the items that comprise the Checkpoint listening and reading assessments.

This report is intended primarily to help flight training decision-makers decide if Checkpoint meets their language assessment requirements but it may be of interest to other stakeholders such as students, student sponsors, assessors, admissions officers, training managers, English language teachers and language testers.

1.1 Listening and reading for flight training

Along with spoken language proficiency, listening and reading skills are crucial to successful flight training. Instructor-led theoretical and practical training places considerable demands on students' listening skills, and the quantity of training material that students are required to read and understand during a primary flight training programme is substantial. For timely and efficient flight training, it is essential that students begin with listening and reading skills that enable rapid and automatic processing of complex spoken and written language.

1.2 Language assessment for student selection and admissions

Weak language proficiency can result in a loss of training efficiency and training repeats and failures which are costly and disruptive. For these reasons, it is important to assess students' language proficiency before they commence flight training. Along with assessment of spoken language proficiency, thorough assessment of listening and reading skills in advance of flight training is important for all stakeholders: it increases confidence in investment and minimises the risk associated with weak language proficiency.

Fair and robust decisions about the selection and admission of students into flight training programmes require valid and reliable language assessment. Scores need to:

- Reflect the range of language knowledge and skills that students need for success in flight training;
- Distribute students across a range of language levels; and
- Be accurate and consistent over time.

As decisions made on the basis of assessment scores affect future careers and have significant financial implications, it is important for providers of language assessments to demonstrate that their assessments are fit for purpose.

1.3 Checkpoint listening and reading

Checkpoint comprises three separate English language skills assessments: Listening, reading and speaking. In the listening and reading assessments, candidates:

- Listen to and answer questions on recordings of instructors, students and training centre representatives in a variety of situations (24 scored items / 40 minutes); and
- Read and answer questions on texts from a range of authentic sources including training courseware, aeronautical information publications, regulatory documentation, incident and accident reports and articles from industry media (24 scored items / 40 minutes).

Candidate responses to items are automatically scored as 'correct' or 'incorrect' according to pre-determined criteria. Checkpoint currently consists of two parallel versions of each of the listening and reading assessments – set A and set B.

2. Reliability estimates

2.1 Internal reliability

The most commonly used standard statistical measure for internal reliability is the Cronbach alpha. Cronbach alpha scores are expressed on a scale from 0 to 1 where 1 indicates perfect internal reliability. If a test with perfect internal reliability were administered twice on the same candidates, it would produce the same distribution of scores and would rank the candidates in the same order so that correlation between the two sets of scores would be perfect. However, in the field of language testing there are many sources of potential measurement error and perfect internal reliability is rarely achieved (Green, 2013, p38). Therefore, test developers work to established standards for the acceptability of Cronbach alpha results as shown in table 1.

Cronbach's alpha	Internal reliability
Above 0.90	Excellent
0.80 to 0.90	Good
0.70 to 0.80	Acceptable
0.60 to 0.70	Questionable
0.50 to 0.60	Poor
Below 0.50	Unacceptable

Table 1: Standards for the acceptability of Cronbach alpha results (George & Mallory, 2003, p231)

It should be noted that Cronbach's alpha is a lower bound estimate of reliability. This means that a result of 0.50 would mean reliability somewhere between 50% and 100%.

By the end of 2015, 702 candidates had taken the finalised versions of set A and set B. Cronbach alpha results for these two final versions of the full Checkpoint assessment were:

- Set A – **CA = 0.902** (reliability between 90.2 and 100%) N=335
- Set B – **CA = 0.915** (reliability between 91.5 and 100%) N= 367

2.2 Standard error of measurement

The internal reliability estimates above tell us how reliable scores are for a particular set of candidates. From these results a Standard Error of Measurement (SEM) can be calculated. The SEM tells us about the effect of measurement error and the extent to which we can have confidence in individual scores. Tables 2 and 3 report the SEM for green, yellow and red candidates in Checkpoint[1]. The figures show how a candidate's 'observed' score (in the assessment) may vary from the candidate's 'true' score (their actual language proficiency). For example, for candidates who score green in set A, observed scores may vary from the candidate's true score by +/- 3.2%.

Ability level	SEM
Green	+/- 3.2%
Yellow	+/- 1.8%
Red	+/- 2.8%

Table 2: SEM for Set A

[1] See the document '*Structure, platform and scores*' on the Latitude website for a description of the traffic light system.

Ability level	SEM
Green	+/- 3.0%
Yellow	+/- 1.7%
Red	+/- 2.5%

Table 3: SEM for Set B

2.3 Correlation between listening, reading and speaking

It is common in selection and admissions procedures for judgements about a candidate's readiness for flight training to be made solely on the basis of their ability to speak, often in non-test language evaluations such as an admissions interview. To understand the extent to which we can make judgements about listening and reading on the basis of spoken performance, a correlation analysis was performed.

Correlation analysis provides information about the strength and direction of the relationship between two variables. Correlation is measured on a scale of -1 to 0 and 0 to +1. Table 4 shows correlations between scores in the Checkpoint listening, reading and speaking assessments. The correlation coefficients show that there are significant (p<0.00, n=702) correlations between scores in the various skills assessments.

		Listening (%)	Reading (%)
Reading (%)	Pearson Correlation	**.668****	
	Sig. (2-tailed)	.000	
	N	702	
Speaking (overall)	Pearson Correlation	**.545****	**.553****
	Sig. (2-tailed)	.000	.000
	N	702	702

Table 4: Correlations between scores in the Checkpoint listening, reading and speaking assessments

Although the results are highly statistically significant, the relationships between the reading, listening and speaking assessments are moderate. While the different skills are clearly related (they are all part of overall English language proficiency), the correlations show that there may be a large variance between a candidate's ability in each skill. For example, a candidate may be poor at reading but good at speaking, or good at listening but poor at speaking. Speaking has a correlation of 0.55 (rounded) with both listening and reading. This means that only 55% of the variance in speaking scores can be attributed to either listening ability or reading ability.

The correlations in table 4 show that it is impossible to accurately judge listening ability or reading ability on the basis of a spoken performance. The implication is that separate measurement of listening, reading and speaking skills is essential, especially as proficiency in all three skills is a prerequisite for successful flight training.

3. Listening and reading item analysis

3.1 Item discrimination

Figure 1 shows the results of an analysis of item (question) difficulty and discrimination conducted in 2015. In figure 1, candidate scores for listening set A are presented according to the traffic-light system[2]. Figure 1 is typical example of the discrimination achieved in listening and reading sets A and B.

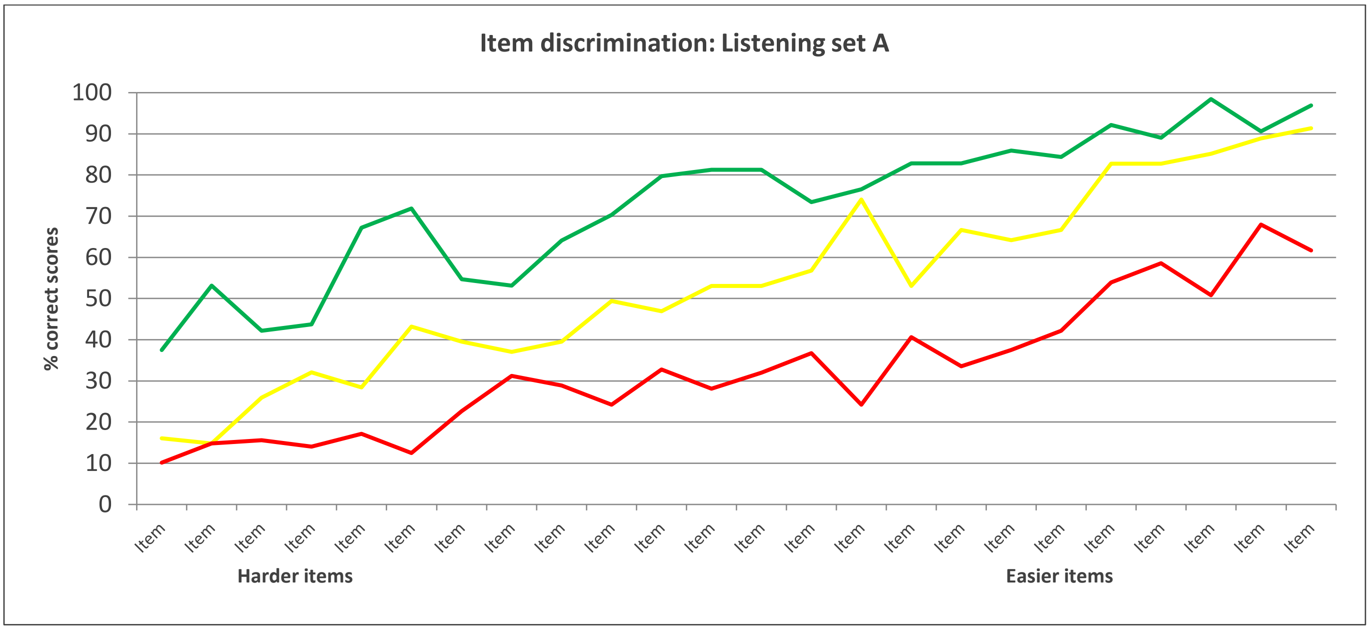

Figure 1: Item discrimination for listening set A

The item discrimination graph shows how good item (question) design, test trialling and item revision leads to virtually perfect discrimination between the chosen candidate ability ranges. On the vertical axis is the score expressed as a percentage. On the horizontal axis are the 24 items belonging to listening set A. The aspects of good test design which we can see in the graph are:

- Coverage of the complete range of item difficulty in a relatively smooth progression from more difficult items (on the left) to easier items (on the right). This is important for ability measurement: too many easy items would not discriminate between high and medium ability candidates and, vice-versa, too many difficult items would not discriminate between medium and low ability candidates.
- The more difficult items are on the left. Here, candidates of high ability (green) achieve an average score of 40% to 50% and both medium and low ability candidates are scoring, on average, below 20%. As expected, these more difficult items do not discriminate between medium and low ability candidates (the questions are too difficult for both). However, these items are required to discriminate between high and very high ability candidates. For example, let's say 2 candidates score "green". If the first candidate has an overall score of 95% and the second an overall score of 70%, it is clear that the first has higher proficiency.
- The easier items are on the right. Here, candidates of low ability (red) achieve an average score of 50% to 60% and both medium and high ability candidates are scoring, on average, above 80%. As expected, these easier items do not discriminate between medium and high ability candidates (the questions are too easy for both). However, these items are required to discriminate between low and very low ability candidates. For example, let's say 2 candidates score "red". If the first candidate has an overall score of 15% and the second an overall score of 40%, it is clear that the second has higher proficiency.
- All the mid-range difficulty items discriminate extremely well between candidates in all three ability categories.

[2] See the document '*Structure, platform and scores*' on the Latitude website for a description of the traffic light system.

3.2 Item performance

To build on the analyses conducted in 2015 and to investigate the performance of individual items in greater detail, item score data were exported from the assessment platform for listening and reading sets A and B for all candidates who took Checkpoint between January 2017 and December 2023. The dataset comprises responses for 1,075 candidates who took set A (listening and reading) and 195 candidates who took set B (listening and reading). The difference in exposure to the two versions is due to the fact that set A is presented for first-time candidates and set B is presented for repeat candidates.

The data for each assessment (listening and reading) and each version (set A and set B) were analysed using the Rasch software programme Winsteps (Linacre, 2023). Output files from the analyses along with brief explanations and discussions are presented below.

3.2.1 Listening: person and item distributions

Using Rasch statistical modelling, Winsteps generates measures of person ability and item difficulty based on response data. Table 5 presents the Wright maps for listening sets A and B. The Wright maps distribute candidates (represented by '#') by measures of ability on the left-hand side of the maps, and items by measures of difficulty on the right-hand side. The maps are arranged vertically where candidates with higher ability and more difficult items appear towards the top, and candidates with lower ability and easier items appear towards the bottom.

Listening set A

```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    4             .  +
                     |
                     |
               ####  |
                     |
                     |
                     |
                    T|
    3                +
               .###  |
                     |
                     |
                     |
          .########  |
                     |
                    S|T
    2   .##########  +
                     |
                     |
      .############  |  L1a.2  L3a.4
                     |  L2a.2
      .############  |
                     |  L1a.8  L3a.3
      .############  |S L1a.4
    1               M+  L1a.7
       .###########  |  L3a.6
                     |
        .##########  |
                     |
       .###########  |
         .#########  |  L2a.6
                     |  L1a.3  L2a.8
    0      .#######  +M L2a.5
                    S|
            .######  |  L1a.5  L3a.8
              .####  |  L1a.6  L2a.1
                     |
                ###  |  L2a.3  L2a.4
                     |  L3a.7
                 .#  |  L2a.7
   -1                +
                 .#  |S L1a.1
                    T|
                  .  |
                     |
                  .  |  L3a.1
                     |  L3a.2
                     |
   -2             .  +
                     |T L3a.5
                     |
                  .  |
                     |
                     |
                     |
                     |
   -3                +
                     |
                     |
                     |
                     |
                  .  |
```

Listening set B

```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    4                +
                     |
                     |
                     |
                  .  |
                     |
                     |
                     |
                     |
    3                +
                     |
                  .  |
                     |
                     |
                     |
                     |
                 .#  |
                    T|
    2                +T
                .##  |
                     |  L2b.5
                     |  L2b.2
                 ##  |
                     |
            #######  |
                     |
             ###### S|  L2b.6
    1                +S L2b.7  L3b.7
              .####  |
                     |
             ######  |
                     |  L2b.1  L2b.8
            #######  |  L1b.1
                     |  L2b.3  L3b.2  L3b.4
             ######  |  L1b.2
                    M|
    0    .#########  +M L3b.6
                     |
           ########  |  L1b.3  L1b.8  L3b.8
           .#######  |  L3b.1
                     |
         ##########  |  L1b.5  L2b.4
                     |
             .#####  |  L3b.3
                     |
   -1           .## S+S
                     |
                     |
              .####  |  L1b.6  L1b.7  L3b.5
                     |
                .##  |
                     |
                     |
                .##  |
   -2               T+T
                     |
                     |
                  .  |
                     |  L1b.4
                     |
                     |
                  #  |
                    |
                    |
  -4              . +
              <less>|<freq>
 EACH "#" IS 8: EACH "." IS 1 TO 7
```

```
                          |
   -3                     +
                   <less>|<freq>
 EACH "#" IS 2: EACH "." IS 1
```

Table 5: Wright maps for listening sets A and B.

The Wright maps show that:

- The items in both versions cover a range of difficulty which is well-suited to the ability of the candidate population; and
- For set A, person ability is higher than item difficulty whereas for set B person ability is only slightly higher than item difficulty. This shows that set B is slightly more difficult than set A (see section 3.2.2).

3.2.2 Listening: item statistics

The item measurement reports present information about how each item is functioning. Tables 6 and 7 present the item measurement reports for listening sets A and B. The names of the items are listed on the right-hand side. The tables are arranged vertically with the easier items at the bottom and the more difficult items at the top (this corresponds with the Wright maps above). An important statistic is the Infit MNSQ. This tells us about the degree of fit of the item to the model. The infit MNSQ parameters for productive measurement are from 0.50 to 1.50 with 1.00 being ideal.

```
-----------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL    JMLE   MODEL|   INFIT  |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM |
|------------------------------------+----------+----------+-----------+-----------+------|
|    20    398   1075    1.63     .07|1.10  3.22|1.16  3.13|  .33   .42| 67.4  70.6| L3a.4|
|     2    402   1075    1.61     .07|1.01   .20|1.25  4.90|  .40   .42| 72.4  70.5| L1a.2|
|    10    414   1075    1.55     .07|1.17  5.75|1.25  5.08|  .27   .42| 64.2  70.1| L2a.2|
|     8    486   1075    1.21     .07|1.04  1.74|1.06  1.44|  .38   .42| 66.7  68.2| L1a.8|
|    19    486   1075    1.21     .07| .91 -3.70| .89 -2.88|  .50   .42| 72.7  68.2| L3a.3|
|     4    509   1075    1.11     .07|1.04  1.55|1.07  1.68|  .38   .42| 66.3  67.7| L1a.4|
|     7    536   1075     .98     .07|1.07  2.72|1.07  1.87|  .36   .42| 64.7  67.7| L1a.7|
|    22    549   1075     .92     .07| .94 -2.34| .95 -1.37|  .46   .42| 71.7  67.6| L3a.6|
|    14    692   1075     .26     .07| .99  -.35| .98  -.50|  .41   .40| 69.7  70.4| L2a.6|
|    16    709   1075     .17     .07|1.00  -.09| .99  -.10|  .40   .39| 70.5  71.1| L2a.8|
|     3    727   1075     .08     .07|1.19  6.01|1.36  6.12|  .21   .39| 65.2  72.0| L1a.3|
|    13    752   1075    -.05     .07| .89 -3.52| .88 -2.19|  .47   .38| 76.6  73.4| L2a.5|
|     5    779   1075    -.19     .07| .95 -1.56| .98  -.35|  .41   .38| 77.0  75.0| L1a.5|
|    24    792   1075    -.26     .08|1.02   .69|1.06   .86|  .34   .37| 76.0  75.9| L3a.8|
|     9    804   1075    -.33     .08| .92 -2.27| .82 -2.77|  .45   .37| 77.3  76.7| L2a.1|
|     6    812   1075    -.38     .08|1.00   .11|1.09  1.24|  .35   .36| 77.3  77.3| L1a.6|
|    11    846   1075    -.59     .08| .96  -.95| .88 -1.57|  .39   .35| 80.2  79.8| L2a.3|
|    12    860   1075    -.68     .08| .98  -.45| .92 -1.00|  .37   .34| 81.0  81.0| L2a.4|
|    23    864   1075    -.71     .08| .97  -.62| .98  -.23|  .37   .34| 81.9  81.3| L3a.7|
|    15    893   1075    -.92     .09| .95  -.88| .90 -1.07|  .38   .33| 83.9  83.7| L2a.7|
|     1    916   1075   -1.10     .09| .95  -.93| .78 -2.14|  .38   .32| 85.8  85.7| L1a.1|
|    17    972   1075   -1.66     .11| .88 -1.61| .68 -2.45|  .39   .28| 90.8  90.6| L3a.1|
|    18    982   1075   -1.78     .11| .99  -.10|1.05   .36|  .28   .27| 91.6  91.6| L3a.2|
|    21   1002   1075   -2.07     .13| .93  -.69| .71 -1.73|  .32   .25| 93.5  93.4| L3a.5|
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN   715.9 1075.0     .00     .08| .99   .08| .99   .26|           | 76.0  76.2|      |
| P.SD   191.2     .0    1.07     .02| .08  2.40| .16  2.48|           |  8.6   7.9|      |
-----------------------------------------------------------------------------------------
```

Table 6: Item measurement report for listening set A.

```
-----------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL    JMLE   MODEL|   INFIT  |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM |
|------------------------------------+----------+----------+-----------+-----------+------|
|    13     38    195    1.75     .19|1.13  1.19|1.44  2.04|  .19   .35| 81.5  81.6| L2b.5|
|    10     39    195    1.71     .19| .83 -1.57| .73 -1.48|  .50   .36| 85.1  81.2| L2b.2|
|    14     59    195    1.06     .17|1.15  1.87|1.55  3.72|  .22   .39| 68.7  73.7| L2b.6|
|    15     60    195    1.04     .17|1.02   .24| .98  -.16|  .39   .39| 73.3  73.5| L2b.7|
|    23     63    195     .95     .17|1.25  3.13|1.64  4.52|  .13   .40| 66.7  72.8| L3b.7|
|     9     78    195     .55     .16|1.15  2.30|1.19  1.93|  .27   .41| 60.5  69.5| L2b.1|
|    16     78    195     .55     .16|1.03   .45|1.07   .73|  .38   .41| 68.7  69.5| L2b.8|
|     1     84    195     .39     .16| .96  -.69| .97  -.31|  .44   .41| 72.8  68.5| L1b.1|
|    11     85    195     .37     .16| .99  -.19|1.04   .52|  .41   .41| 69.2  68.4| L2b.3|
|    18     86    195     .34     .16| .89 -1.93| .85 -1.72|  .51   .41| 74.9  68.3| L3b.2|
|    20     87    195     .32     .16| .97  -.52| .97  -.27|  .44   .41| 69.2  68.2| L3b.4|
|     2     90    195     .24     .16|1.03   .47|1.03   .40|  .39   .41| 68.7  67.9| L1b.2|
|    22     98    195     .05     .16| .84 -2.98| .80 -2.54|  .56   .41| 75.9  67.4| L3b.6|
|     3    107    195    -.18     .16|1.09  1.47|1.11  1.19|  .33   .41| 63.6  67.5| L1b.3|
|     8    108    195    -.20     .16|1.05   .89|1.05   .61|  .36   .41| 65.1  67.6| L1b.8|
|    24    108    195    -.20     .16| .99  -.11| .98  -.16|  .42   .41| 68.2  67.6| L3b.8|
|    17    115    195    -.38     .16| .94 -1.03| .93  -.76|  .46   .41| 72.8  68.3| L3b.1|
|    12    122    195    -.56     .16| .87 -2.09| .84 -1.56|  .52   .40| 75.9  69.5| L2b.4|
|     5    123    195    -.59     .16|1.00   .04| .96  -.33|  .40   .40| 68.2  69.7| L1b.5|
|    19    132    195    -.83     .17| .87 -1.81| .82 -1.55|  .50   .39| 75.4  72.3| L3b.3|
|     7    147    195   -1.28     .18| .99  -.11| .98  -.04|  .37   .37| 79.5  77.4| L1b.7|
|    21    148    195   -1.31     .18|1.03   .32|1.12   .77|  .33   .36| 76.4  77.8| L3b.5|
|     6     150     195    -1.38      .18| .81 -2.12| .74 -1.64|  .54    .36| 81.5   78.6| L1b.6|
|     4     174     195    -2.42      .24| .92   -.46| .89  -.26|  .35    .28| 89.2   89.4| L1b.4|
|-----------------------------------+----------+----------+-----------+-----------+------|
| MEAN     99.1  195.0       .00      .17| .99   -.13|1.03    .15|           | 73.0   72.4|      |
| P.SD     34.8     .0       .98      .02| .11  1.48| .23  1.65|           |  6.8    5.7|      |
------------------------------------------------------------------------------------------
```

Table 7: Item measurement report for listening set B.

The item measurement reports show that all 48 items in the two versions of the listening assessment are measuring effectively, with many items close to or at the ideal infit MNSQ value of 1.00. The two highlighted items (L2b.6 and L3b.7) show raised outfit MNSQ values. As outfit MNSQ is sensitive to outliers exhibiting unexpected response patterns, for example, candidates 'correctly answering items which are far from their ability level such as might occur when guessing' (Green, 2013, p169), these items were flagged for qualitative review (see section 3.3).

The item measurement reports show that item difficulty measures range from:

- -2.07 to +1.63 logits in set A (with a mean of 0.08); and
- -2.42 to +1.75 logits in set B (with a mean of 0.17).

As mentioned above, this shows that set B is slightly more difficult than set A. This slight difference in version difficulty was identified in the initial analyses in 2015. At that stage, to ensure that the scores we report for the listening assessment are equivalent regardless of the version the candidates take, we implemented an upwards score adjustment for listening set B scores. The 2024 analysis confirms the requirement to preserve this adjustment to maintain version equivalence.

3.2.3 Reading: person and item distributions

Table 8 presents the Wright maps for reading sets A and B.

Reading set A

```
MEASURE     PERSON - MAP - ITEM
                    <more>|<rare>
    4                  #  +
                          |
                          |
                          |
                     .##  |
                          |
                          |
                          |
    3                     +
                          |
                         T|
                  .#####  |
                          |
                          |
                  .#####  |
                          |
    2                     +
                  .#####  |
                          |T
                          |  R3a.6
                   ##### S|  R2a.2
                          |
                .#######  |
                          |
    1          #########  +
                          |S R4a.4  R4a.5
               .########  |
               .########  |  R3a.4  R3a.8
                          |  R2a.4  R3a.3
               .########  |  R3a.2
           .############ M|  R2a.3  R4a.6  R4a.7
                          |  R3a.1
    0      .############  +M R3a.7
             .##########  |  R4a.9
                          |  R1a.3
             .##########  |
                          |  R1a.2  R4a.8
              ##########  |  R3a.5  R4a.2
               .########  |
                          |S
   -1          .######## S+  R2a.1
                          |  R1a.4
                 .######  |  R1a.1
                          |
                    .###  |
                          |
                          |T
                     .##  |
   -2                     +
                          |  R4a.1
                       # T|
                          |
                          |
```

Reading set B

```
MEASURE     PERSON - MAP - ITEM
                    <more>|<rare>
    4                  #  +
                          |
                          |
                          |
                      ##  |
                          |
                          |
                          |
                          |
    3                     +
                         T|
                      .#  |
                          |
                          |
                          |
                          |  R4b.4
                     ###  |
                          |
    2                     +T
                      ##  |
                          |
                          |
                   ##### S|
                          |
                  .#####  |  R4b.3
                          |
                     ###  |
    1                     +S R2b.2
                          |  R2b.4  R3b.6  R4b.6
              ##########  |  R4b.2
                     ###  |
                          |  R3b.3
                  .#####  |  R3b.5
                          |
               .######## M|  R1b.3
                          |  R3b.7  R4b.8
    0             ######  +M R3b.4  R4b.7
                          |
                  .#####  |  R2b.1  R3b.8
                          |  R1b.4
                .#######  |
                          |
                  .#####  |  R3b.1
                     ###  |  R2b.3
                          |  R4b.5
   -1                     +S
                    ####  |  R4b.1
                         S|
                    ####  |
                          |
                   #####  |  R3b.2
                          |  R1b.2
                          |  R1b.1
```

```
                 |
               . |
                 |
  -3             +
                 |
                 |
                 |
               . |
                 |
                 |
                 |
  -4           . +
          <less>|<freq>
 EACH "#" IS 7: EACH "." IS 1 TO 6
```

```
                  ####  |
   -2                   +T
                        |
                        |
                    .## |
                        |
                       T|
                        |
                      . |
                        |
   -3                   +
                  <less>|<freq>
 EACH "#" IS 2: EACH "." IS 1
```

Table 8: Wright maps for reading sets A and B.

The wright maps show that for both set A and set B:

- The items cover a range of difficulty which is well-suited to the ability of the candidate population; and
- Person ability is slightly higher than item difficulty.

3.2.4 Reading: item statistics

Tables 9 and 10 present the item measurement reports for reading sets A and B.

```
-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL     JMLE   MODEL|   INFIT  |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE   S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM |
|------------------------------------+----------+----------+-----------+-----------+------|
|    14    280   1075     1.59     .08|1.17  3.90|1.45  5.73|  .31   .45| 76.1  78.8| R3a.6|
|     6    291   1075     1.52     .08| .96  -.88| .91 -1.35|  .49   .46| 79.1  78.2| R2a.2|
|    20    410   1075      .87     .07| .96 -1.42| .93 -1.47|  .50   .46| 73.8  72.4| R4a.5|
|    19    413   1075      .85     .07|1.03  1.10|1.01   .17|  .44   .46| 70.7  72.3| R4a.4|
|    16    456   1075      .64     .07| .89 -4.06| .86 -3.33|  .54   .46| 74.4  70.9| R3a.8|
|    12    457   1075      .63     .07|1.05  1.58|1.10  2.13|  .42   .46| 69.5  70.9| R3a.4|
|    11    490   1075      .47     .07|1.11  3.99|1.16  3.65|  .37   .46| 66.7  70.1| R3a.3|
|     8    493   1075      .46     .07|1.00   .01|1.00   .02|  .46   .46| 69.5  70.0| R2a.4|
|    10    523   1075      .31     .07| .97  -.97| .95 -1.11|  .48   .46| 70.5  69.5| R3a.2|
|    21    526   1075      .30     .07| .94 -2.39| .90 -2.55|  .50   .45| 72.8  69.4| R4a.6|
|     7    538   1075      .24     .07|1.01   .38| .98  -.35|  .45   .45| 67.6  69.2| R2a.3|
|    22    543   1075      .22     .07| .87 -4.93| .83 -4.23|  .54   .45| 74.1  69.1| R4a.7|
|     9    561   1075      .13     .07|1.06  2.15|1.10  2.25|  .40   .45| 66.8  69.1| R3a.1|
|    15    600   1075     -.05     .07| .97 -1.32| .94 -1.25|  .47   .44| 69.1  69.1| R3a.7|
|    24    615   1075     -.12     .07| .93 -2.91| .87 -2.92|  .50   .44| 71.8  69.3| R4a.9|
|     3    652   1075     -.30     .07|1.12  4.18|1.13  2.37|  .35   .43| 64.9  70.0| R1a.3|
|     2    680   1075     -.44     .07| .96 -1.50|1.00  -.01|  .45   .42| 72.4  70.7| R1a.2|
|    23    704   1075     -.56     .07| .97 -1.13|1.00   .00|  .44   .42| 72.7  71.5| R4a.8|
|    18    714   1075     -.61     .07| .85 -5.46| .76 -4.30|  .53   .41| 77.4  72.0| R4a.2|
|    13    718   1075     -.63     .07| .95 -1.76| .86 -2.30|  .46   .41| 73.1  72.1| R3a.5|
|     5    790   1075    -1.02     .08| .93 -2.00| .84 -2.20|  .45   .39| 77.6  75.9| R2a.1|
|     4    810   1075    -1.14     .08|1.10  2.60|1.42  4.57|  .28   .38| 75.4  77.3| R1a.4|
|     1    818   1075    -1.19     .08|1.25  5.99|1.94  8.79|  .15   .38| 73.7  77.8| R1a.1|
|    17    947   1075    -2.17     .10| .90 -1.46| .75 -1.81|  .39   .31| 88.6  88.5| R4a.1|
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN   584.5 1075.0      .00     .07|1.00  -.26|1.03   .02|           | 72.9  72.7|      |
| P.SD   163.7     .0      .86     .01| .10  2.89| .26  3.18|           |  4.9   4.5|      |
-------------------------------------------------------------------------------------------
```

Table 9: Item measurement report for reading set A.

```
-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL     JMLE   MODEL|   INFIT  |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE   S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM |
|------------------------------------+----------+----------+-----------+-----------+------|
|    20     32    195     2.28     .22| .82 -1.35| .66 -1.25|  .55   .45| 88.1  86.0| R4b.4|
|    19     55    195     1.37     .18| .95  -.53| .90  -.57|  .51   .48| 78.2  77.5| R4b.3|
|     6     68    195      .96     .17| .91 -1.15| .84 -1.16|  .55   .48| 74.1  74.1| R2b.2|
|     8     71    195      .87     .17|1.10  1.25|1.13   .99|  .42   .49| 69.4  73.3| R2b.4|
|    14     71    195      .87     .17|1.39  4.59|1.74  4.61|  .21   .49| 60.1  73.3| R3b.6|
|    22     71    195      .87     .17| .97  -.32| .90  -.72|  .51   .49| 75.6  73.3| R4b.6|
|    18     75    195      .76     .17|1.05   .66|1.06   .49|  .46   .49| 69.4  72.6| R4b.2|
|    11     84    195      .50     .17| .96  -.59| .91  -.74|  .52   .49| 72.0  71.5| R3b.3|
|    13     86    195      .45     .17|1.01   .21|1.02   .22|  .48   .49| 73.1  71.3| R3b.5|
|     3     95    195      .20     .16|1.02   .31|1.03   .26|  .47   .48| 70.5  70.5| R1b.3|
|    15     97    195      .15     .16|1.01   .17|1.00   .08|  .47   .48| 69.4  70.5| R3b.7|
|    24    100    195      .07     .16|1.10  1.55|1.08   .76|  .42   .48| 65.8  70.6| R4b.8|
|    12    103    195     -.01     .16| .89 -1.73| .86 -1.30|  .55   .48| 76.7  70.6| R3b.4|
|    23    104    195     -.04     .16| .90 -1.47| .83 -1.50|  .55   .48| 73.1  70.7| R4b.7|
|     5    110    195     -.20     .17| .93  -.95| .87 -1.07|  .52   .47| 74.6  71.0| R2b.1|
|    16    112    195     -.26     .17| .95  -.69| .90  -.78|  .51   .47| 71.5  71.2| R3b.8|
|     4    116    195     -.37     .17|1.05   .72|1.24  1.80|  .43   .47| 66.3  71.6| R1b.4|
|     9    126    195     -.65     .17|1.09  1.14|1.39  2.45|  .39   .45| 70.5  73.4| R3b.1|
|     7    129    195     -.74     .17| .97  -.32| .93  -.41|  .47   .45| 74.1  74.0| R2b.3|
|    21    135    195     -.92     .18| .95  -.55| .87  -.75|  .47   .44| 75.1  75.4| R4b.5|
|    17    142    195    -1.14     .18|1.02   .27| .99   .01|  .41   .42| 77.7  77.2| R4b.1|
|    10    154    195    -1.57     .19| .91  -.82| .77  -.96|  .45   .39| 82.9  80.6| R3b.2|
|     2    156    195    -1.64     .20| .94  -.50| .77  -.90|  .43   .38| 80.8  81.4| R1b.2|
|     1    160    195    -1.81     .21|1.06   .53| .88  -.35|  .35   .37| 80.8  82.9| R1b.1|
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN    102.2  195.0      .00      .18|1.00    .02| .98   -.03|            | 73.7  74.4|       |
| P.SD     32.8     .0      .98      .01| .11   1.28| .22   1.37|            |  5.9   4.3|       |
--------------------------------------------------------------------------------------------------
```

Table 10: Item measurement report for reading set B.

The item measurement reports show that all 48 items in the two versions of the reading assessment are measuring effectively, with many items close to or at the ideal infit MNSQ value of 1.00. The two highlighted items (R1a.4 and R3b.6) show raised outfit MNSQ values and were flagged for qualitative review (see section 3.3 below).

Item difficulty measures range from:

- -2.17 to +1.59 logits in set A (with a mean of 0.09); and
- -1.81. to +2.28 logits in set B (with a mean of 0.18).

This means that set B is slightly more difficult than set A. As with the listening assessment, this slight difference in reading version difficulty was identified in the initial analyses in 2015. At that stage, to ensure that the scores we report for the listening assessment are equivalent regardless of the version the candidates take, we implemented an upwards score adjustment for reading set B scores. The 2024 analysis confirms the requirement to preserve this adjustment to maintain version equivalence.

3.3 Item review

Of the 96 items in the two versions of the listening and reading assessments, four items were flagged for qualitative review due to raised outfit MNSQ measures. As mentioned above, this might be due to weaker candidates successfully guessing the correct answer for more difficult items possibly because:

- There is a clue in the item itself; and/or
- The distractors are not functioning as intended.

The items were presented to a panel of judges. Following inspection and discussion, the wording of three of the items was adjusted and one item was re-drafted completely. These edits were incorporated into new issues of the Checkpoint assessment in January 2024 and they will be analysed for performance once a sufficient volume of data is available.

References

George, D. & Mallery, P. (2003).SPSS for windows step by step: A simple guide and reference. (4th Edition) Boston: Allyn & Bacon.

Green, R. (2013). Statistical analyses for language testers. Basingstoke: Palgrave Macmillan.

Linacre, J. M. (2023). Winsteps Rasch measurement computer program. Chicago: Winsteps.com.