The Journal of Classics Teaching (2022), **23**, 55–59
doi:10.1017/S2058631021000532

CAMBRIDGE
UNIVERSITY PRESS

Forum

Validating Communicative Tests of Reading and Language Use of Classical Greek

David Coniam, Polyxeni Poupounaki-Lappa and Tzortzina Peristeri

LanguageCert, Athens, Greece

Abstract

This paper builds on the work presented previously in this journal by Poupounaki-Lappa et al. (2021), which described the development of a communicative test of Reading and Language Use of Classical Greek calibrated to the Common European Framework of Reference (CEFR) at levels A1 and A2 (Council of Europe, 2001). In the current paper, the two tests of Classical Greek are calibrated both together and to the CEFR. In addition to describing the methodology for comparing the two separate tests of Classical Greek, the paper is also designed to be of interest to educators of other classical languages. It is hoped that they may find it useful not only by facilitating robust test design, but also by demonstrating the methods by which tests can be linked together on a common scale (as with the CEFR) or linking tests one to another (e.g., different end-of-year tests, at different points in time).

Key words: Classical Greek, reading and language use, assessment, Rasch, test linking

Introduction

This paper builds on the groundwork presented in Poupounaki-Lappa et al. (2021), which described the development of a communicative test of Reading and Language Use of Classical Greek calibrated to the Common European Framework of Reference (CEFR) at levels A1 and A2 (Council of Europe, 2001)[1].

The test outlined in Poupounaki-Lappa et al. (2021) discussed issues – in line with more 'communicative' approaches to language teaching (see Richards & Rodgers, 2014; Lloyd & Hunt, 2021) – related to the creation of a communicative testing system for Classical Greek, initially centring around Reading and Language use. Supplementary Appendix 1 presents samples of the constructs assessed in terms of reading skills, grammar and syntax, and topics in the two tests.

The major focus of the current paper involves calibrating – both to one another and to the CEFR – two tests of Classical Greek. In this light, the assessment side of the current paper's methodology may be seen to extend beyond its Classical Greek dimension. The paper will hopefully be of interest to educators of other classical languages who may also find it useful not only regarding robust test design, but also regarding linking tests together on a common scale (as with the CEFR) or in linking tests one to another (e.g., different end-of-year tests, at different points in time).

The test development process described in Poupounaki-Lappa et al. (2021) involved the construction of two tests of Reading and Language Use, with each test consisting of four parts, each using distinct task types to assess specific sub-skills. Figure 1 elaborates.

The detailed set of specifications and associated official practice material is available in the *LanguageCert Test of Classical Greek (LTCG) Qualification Handbook* for the examination (LanguageCert, 2021). The examples provided below, as well as those in Appendix 1, are drawn from this *Qualification Handbook*.

Piloting

The two tests were piloted in mid-2021, administered to groups of test takers who were judged to be at the intended level of language proficiency by the subjects' teachers. The current paper presents details of the tests, and their match with the supposed target levels of A1 and A2.

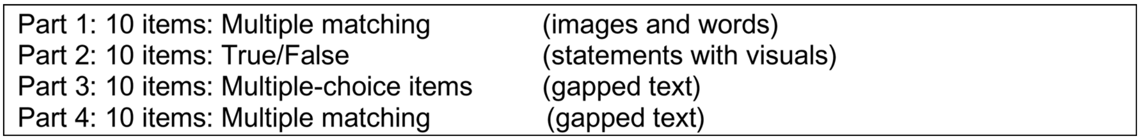
Key test qualities are validity and reliability (Bachman & Palmer, 2010). With regards to validity, central issues include how well the different parts of a test reflect what a test taker can do, and how well test scores provide an indication of test taker communicative ability (Messick, 1989; Bachman & Palmer, 2010). The Classical Greek tests assess what test takers will be expected to have control over at particular levels of ability (i.e., in relation to the CEFR). Against such a backdrop, test content needs to match target test takers' levels in terms of grammar, functions, vocabulary and topics.

As a starting point, Morrow (2012) outlines the rationale and aims of communicative language tests, stating that the aim of a communicative language test is to find out what a learner can do in the language. Moving to practicalities, if a communicative test is to be valid and reliable, it nonetheless needs to be well constructed. In addition to validity, some of the features of a 'good' test (see Hughes, 2003) are defined as tests being reliable and at an 'acceptable' level of difficulty.

Test difficulty needs to be considered from two perspectives. One, that it matches the ability level of the intended target group;

Author of correspondence: David Coniam, E-mail: coniam@eduhk.hk

Cite this article: Coniam D, Poupounaki-Lappa C and Peristeri T (2022). Validating Communicative Tests of Reading and Language Use of Classical Greek. *Journal of Classics Teaching* 23, 55–59. https://doi.org/10.1017/S2058631021000532



Part 1: 10 items: Multiple matching	(images and words)
Part 2: 10 items: True/False	(statements with visuals)
Part 3: 10 items: Multiple-choice items	(gapped text)
Part 4: 10 items: Multiple matching	(gapped text)

Figure 1. Classical Greek task types

but two, that it is sufficiently discriminating to permit the exam body (or teacher etc.) to be able to confidently make decisions about the extent to which test takers have met the language competencies required for the particular level, the pass mark and the grade they should be awarded.

Statistical Analysis

In the current study – to gauge test fitness for purpose, and to link two different tests to a common scale – two types of statistical analysis have been performed. The first of these involves classical test statistics, reporting test mean and test reliability. The second involves the use of Rasch measurement which serves the purpose of calibrating the two tests together. Both types of statistics will be briefly outlined. For assessors used only to classical test statistics, it should be noted that the development of Rasch measurement facilitates the calibration of two different tests.

Classical Test Statistics (CTS)

The test mean for a proficiency test may be expected to be within a range of 60–70%. This will depend, however, on where the pass mark is set by the exam body concerned, and the purpose for which the test is intended. A test mean of around 60–70% suggests that the test is generally appropriate to the level of a 'typical' test taker (Burton et al., 1991). Such a mean in general indicates that most test takers managed to finish the test and that test takers may be assumed to have done their best.

In terms of test reliability – where levels of reliability are associated with test length (Ebel, 1965) – expected reliability with a 40-item test is in the region of 0.67.

Rasch Measurement

The use of Rasch measurement enables different facets to be modelled together, converting raw data into measures which have a constant interval meaning (Wright, 1977). This is not unlike measuring length using a ruler, with the units of measurement in Rasch analysis (referred to as 'logits') evenly spaced along the ruler. In Rasch measurement, test takers' theoretical probability of success in answering items is gauged – scores are not derived solely from raw scores. While such 'theoretical probabilities' are derived from the sample assessed, they are able to be interpreted independently from the sample due to the statistical modelling techniques used. Measurement results based on Rasch analysis may therefore be interpreted in a general way (like a ruler) for other test taker samples assessed using the same test. Once a common metric is established for measuring different phenomena (test takers and test items in the current instance), test taker ability and item difficulty may be estimated independently of the items used, independently from the sample (Bond et al., 2020).

In Rasch analysis, test taker measures and item difficulties are placed on an ordered trait continuum. Direct comparisons between test taker abilities and item difficulties, as mentioned, may then be conducted, with results able to be interpreted with a more general meaning. One of these more general meanings involves the transferring of values from one test to another via anchor items. Anchor items are a number of items that are common to both tests; they are invaluable aids for comparing students on different tests. Once a test, or scale, has been calibrated (see e.g., Coniam et al., 2021), the established values can be used to equate different test forms.

In the current Classical Greek study, the tests developed for A1 and A2 needed to be linked to one another, and items placed on a common scale. The two tests have therefore been linked via a set of common items (the cloze passage in Part 3) (Bond, Yan & Heene, 2020).

To adequately validate a test (or tests) nonetheless requires some form of external triangulation or confirmation beyond the test. To this end, the single scale produced through Rasch measurement in the current project has been validated by a number of test takers who completed a set of *Can-do* statements (see Appendix 2) ranging from Pre-A1 to B1+ levels. These self-assessments were then regressed against test scores, providing evidence for the validity of the test constructs through the test takers' judgements of their own abilities. Detail on the *Can-do* statements and the validation procedure is reported below.

Data and Analysis

As mentioned, two tests of 40 items were constructed with a 10-item MC cloze passage common to both tests. One test at intended A1 level and another at intended A2 level were administered in spring 2021. It had been hoped that about 150 subjects from a variety of first language backgrounds would take each pilot test. Due to covid-19 pandemic restrictions, however, sample sizes were consequently smaller, with most subjects being first language speakers of modern Greek. Sample sizes were, however, large enough for statistical analyses to be able to be performed.

Classical Test Statistics

This section briefly describes key classical test statistics.

Table 1 first presents test means and reliabilities.

For 40 items, both test means were in the desirable range – in the 70 percent range. This suggests that the tests broadly fit the target population, and that most test takers finished the test and had given it their best shot. Test reliability for both tests was above 0.67 indicating that the tests may be assumed to have been well constructed. The spread of ability (indicated by the standard deviation) was narrower in the A1 cohort of test takers.

Table 2 now presents the picture of means in each of the four subtests. Each subtest comprised 10 items, with Part 3 the common section in both tests.

Table 1. Test means and reliabilities

	A1	A2
Test takers	74	89
Mean	28.9 (72%)	30.3 (76%)
Standard deviation	4.4 (11.0%)	6.2 (15.5%)
Reliability	0.72	0.88

https://doi.org/10.1017/S2058631021000532 Published online by Cambridge University Press

Table 2. Subtest means (Max = 10 on each subtest)

Part	A1	A2	Subtest type	Note
1	94%	86%	Matching words to pictures	
2	82%	81%	True/False statements with visuals	
3	59%	76%	Multiple-choice cloze passage	Common section
4	47%	60%	Multiple matching gapped text	
O'all mean	72%	76%		

As had been intended, a cline of difficulty emerged, with subtests increasing in difficulty from one subtest to the next. Part 1, *Matching words to pictures* emerged as very easy, with means close to or above 90%. Part 2, *True/False statements with visuals* emerged as comparatively easy, with means in the 80% range. Part 3, the Multiple-choice (MC) *Cloze passage* common to both tests emerged with a mean of 59% for the A1 cohort and 76% for the A2 cohort – an indication that the two groups were of differing ability. Part 4, the *Multiple matching gapped text* exercise, emerged as the most difficult. The cline of difficulty may also be seen as a reflection of intended functional demands. Part 1 involved vocabulary and items were discrete. Part 4 required that test takers operate at the more complex text level – constructing a text by matching the two lists of ten possibilities.

Raw scores are a baseline indication of how 'good' a test may be deemed. If comparisons are to be made across tests, however, or if tests are to be calibrated together, Rasch measurement needs to be employed, and it is to this statistic that the discussion now turns.

Rasch Measurement

In interpreting Rasch, the key statistic involves the 'fit' of the data in terms of how well obtained values match expected values (Bond et al., 2020). A perfect *fit* of 1.0 indicates that obtained mean square values match expected values one hundred percent[2]. Acceptable ranges of tolerance for *fit* range from 0.7 through 1.0 to 1.3, indicating a tolerance for a 30% divergence between obtained and expected values (Bond et al., 2020). *Fit* for the key statistics that are usually reported in academic journals – infit and outfit mean squares for both persons and items – were within acceptable limits. The outcomes from the Rasch analysis confirm – from a different perspective – the classical test statistics results that have been presented above. Both sets of analyses underscore and add to an appreciation of the baseline robustness of the two tests.

To provide an overview of the Rasch measurement technique, the vertical ruler (the 'facet map') produced in the Rasch output is presented below. The facet map is a visual representation of where the facets of test takers and items are located on the Rasch scale. Figure 2 below presents the map for tests A1 and A2 as seen when calibrated together. The results for the A1 test takers and items appear to the left-hand side of the Figure, while the A2 test takers and items appear to the right-hand side. Test takers are represented as asterisks or crosses, and items are represented by the item numbers

The map should be interpreted as follows. For each test, the top left-hand side of the map indicates more able test takers; in a similar manner, the top right-hand side represents more difficult items. Conversely, less able test takers appear to the bottom left-hand side of the map, and easier items to the bottom right-hand side. The green rectangles indicate the test taker midpoints, while the red ovals indicate the item midpoints.

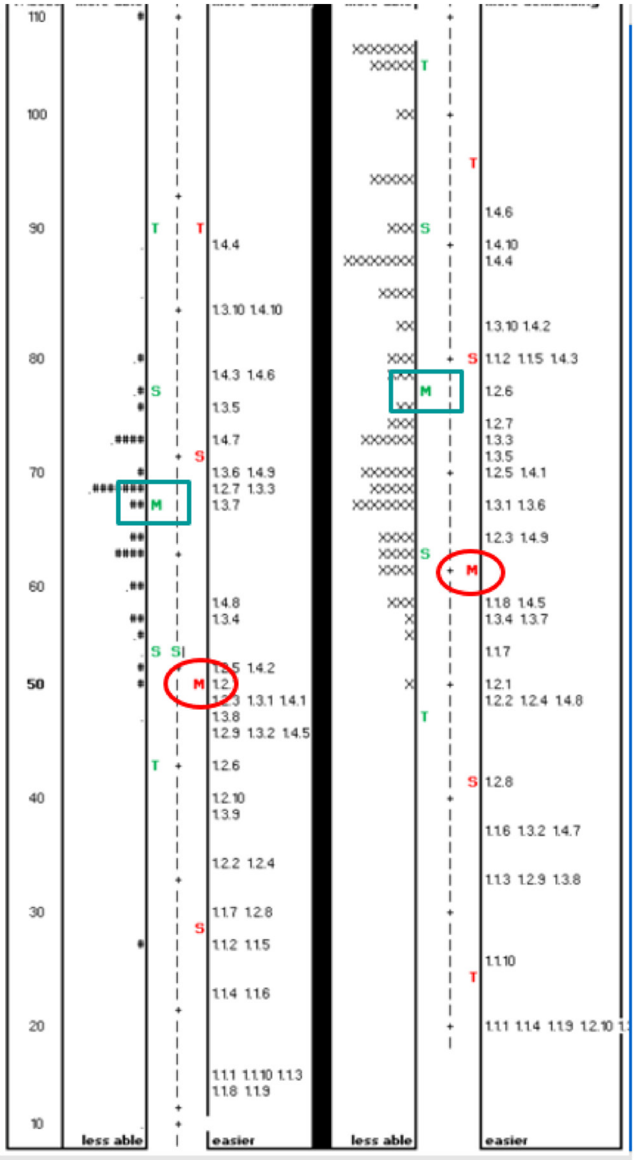


Figure 2. Facet map

To ease interpretation, Rasch measures ('logits') in the study have been rescaled to a mean of 50 with a standard deviation of 10.

As can be seen from Figure 2, the midpoint for the A1 items is 50, whereas the midpoint for the A2 items is 62. This indicates one logit of difference (10 points) between the items; the A2 items, as had been intended, have emerged as more demanding. Turning to test takers, the A1 test takers have a mean of 67, while the A2 test takers have a mean of 78. This clearly indicates that the A2 test takers are more able than the A1 cohort, again by one logit, or 10 points.

The bottom line has thus been satisfied in two respects: a) the test items differentiate between tests; and b) test taker cohorts may be seen to be of increasing ability. The item / test taker match, however, is less than optimal.

Test takers are in a comparatively narrow range. Ignoring outliers, the A1 test takers are in a three-logit range from 50 to 80; the A2 cohort show a rather wider range from 60 to 105, a 4.5 logit range. While the A1 items cover a wide difficulty range, many items – as can be seen from the map – are below 50, the bottom end of test

https://doi.org/10.1017/S2058631021000532 Published online by Cambridge University Press

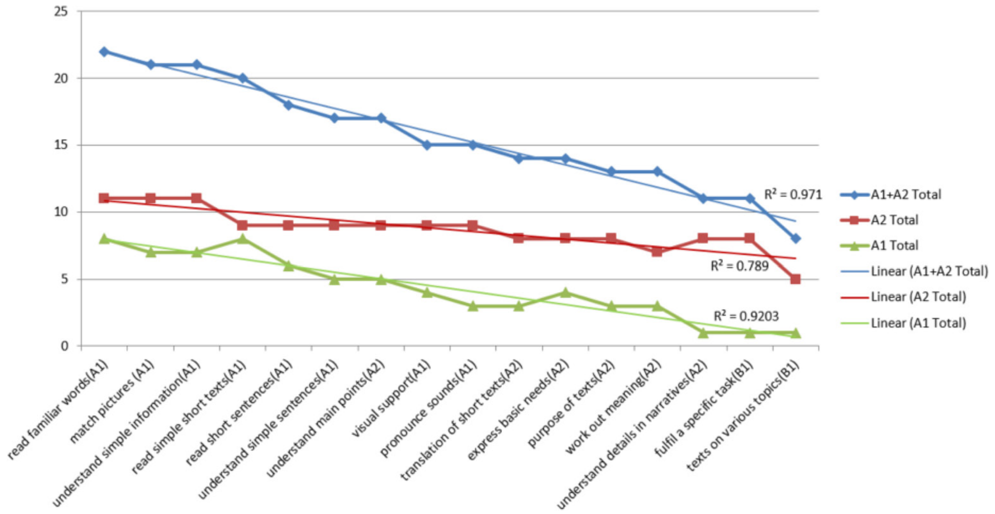


Figure 3. Self-assessment *Can-do* statements regressed against test score

Key: Linear = Regression line of best fit

taker ability. These items are too easy since they do not match with any test taker abilities, and consequently return no useful assessment 'information' on test takers. In future live tests, an attempt will be made to address this situation: the number of very easy items will be reduced, with a view to working towards a closer test taker ability / item difficulty match.

A similar, although less exacerbated, situation exists with the A2 test, where there is still a number of very easy items towards the bottom right-hand end of the map. Similar attention will be paid in order to redress the imbalance in this test.

Triangulating Test Results

As mentioned, test takers who took the two tests were subsequently approached – by email, following their consent to be contacted – and asked to complete a survey. This consisted of a series of Self-assessment *Can-do* statements, adapted from CEFR material for other languages. The use of instruments such as *Can-do* statements in self-assessment has been validated in a number of studies, see e.g., Brown et al., 2014; Summers et al., 2019. In the current study, there were 16 items, with intended difficulty levels ranging from low A1 (at the left-hand end of the figure) to high A2/low B1 (at the right-hand end of the figure) – see Supplementary Appendix 2. Respondents were asked to rate themselves for each item on a six-point scale ('6' being high) in order to demonstrate whether they felt they could master the requisite skill.

The survey was completed by only a small number of test takers (12 for A1; 15 for A2), so results may only be seen as indicative. Figure 3 presents the results of regressing the different *Can-do* statements against test scores. In Figure 3, the blue diamonds are the responses of the A1 group, the red squares those of the A2 group, and the green triangles those of the combined groups. R2 indicates the amount of variance accounted for by the regression.

Figure 3 illustrates a clear match between test takers' perceived abilities in reading and usage in Classical Greek, and their actual scores on the relevant test. The R^2 values to the bottom right of the regression line indicate a close fit between perceived abilities and test scores. Despite the small sample sizes, Figure 3 provides additional external validity evidence for the communicative traits underpinning the two tests.

Conclusion

This paper is a sequel to that introduced in Poupounaki-Lappa et al. (2021) which described the development of a communicative test of Reading and Language Use of Classical Greek calibrated to the CEFR at levels A1 and A2. Both A1 and A2 level tests comprised four parts, with the four parts designed to produce a cline of difficulty, progressing from vocabulary recognition at the lowest level through to text-based exercises in the later part of the tests. For calibration purposes, one part (the MC cloze passage) was common to both tests.

In addition to reporting on the further development, administration and analysis of the two tests and the robustness of their validity and reliability, the current paper has described an important feature of test development: the calibration, separately and then together, of the two tests of Reading and Language Use produced for levels A1 and A2.

The first part – matching words to pictures – emerged as very easy, possibly too easy, on both tests. The second part – matching True/False statements to visuals was also comparatively easy. The third part, common to both tests, was a cloze passage requiring test takers to make lexical / grammatical / syntactic contextual fits. The fourth and final part, which proved to be the most demanding, was a multiple-matching gapped text exercise, which required test takers to make sense of a whole text.

The tests were administered to comparatively small cohorts of test takers who had been estimated by their teachers to be at the approximate level for whichever level of test that they took. Subsequent to taking the test, test takers were approached and asked to complete a self-assessment of CEFR-linked *Can-do* statements reflecting abilities from pre-A1 to B1+ level.

As a baseline, based on classical test statistics, both tests were reliable; and mean scores were acceptable overall – even if some parts of the test were very easy. As a result, difficulty levels of some parts will need to be reconsidered. The cloze passage common to both tests indicated that the two cohorts were different and that the tests could be linked. Linking together was then achieved via the use of Rasch measurement, where fit statistics were good and the two tests were successfully calibrated and linked together on a common scale.

https://doi.org/10.1017/S2058631021000532 Published online by Cambridge University Press

The regressing against test scores of test takers' self-assessments on the *Can-do* statements enabled a degree of triangulation to be conducted, providing an external validation of the fit of the test to the target population. While the A2 test was seen to be pitched at a higher level to the A1 test, test results suggest that the difference between the two tests needs to be extended when future tests are produced.

In addition to presenting the development and analysis of the trialling of the two tests, the main focus in the current paper has centred around calibrating two tests of Classical Greek to each other and to the CEFR, an important issue in effective test development. Looking beyond the paper's immediate Classical Greek focus, however, it is hoped that the methodologies outlined may also be considered useful from a more general perspective, and may interest educators and teachers of other classical languages who wish to consider developing good tests. Such a 'more general' assessment perspective may involve the construction of different tests which need to be linked to other tests – possibly via a common scale of ability – or simply that of different tests being analysed together so that direct comparisons may be made between different test taker cohorts, for example.

Limitations

Two limitations alluded to in the study will now be discussed, with both linked to the covid pandemic and restrictions imposed just as the piloting of the two tests was scheduled.

The first limitation related to sample size, which had been projected to be in the region of 150 or so for each test. Ultimately, a sample of only approximately half that number was achieved. The second limitation was mother tongue, with the sample originally projected as having an international perspective, comprising subjects with a range of mother tongues. Again, as a result of the covid lockdown, this was not achieved, and the mother tongue of over 90% of test takers was modern Greek, with the test takers based in Greece.

The A1 and A2 tests have now gone live. As data becomes available with the administration of the live tests, it is anticipated that the analysis conducted in the current study will be revisited. Further, given that the Classical Greek Reading and Language Use test may be taken from anywhere worldwide via LanguageCert's Online Proctoring facility (https://www.languagecert.org/en/welcome), the sample will be further adapted to an international audience as more test takers take the test over time.

Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1017/S2058631021000532

Notes

[1] The Council of Europe's Common European Framework of Reference (CEFR) has played a decisive role in the teaching and setting standard for initially European languages. The CEFR organises language proficiency in six levels, A1 to C2. These can be regrouped into three broad levels: Basic User, Independent User and Proficient User, with levels defined through 'can-do' descriptors. See https://www.coe.int/en/web/common-european-framework-reference-languages/illustrations-of-levels.

[2] The *mean square* of a set of values is the mean of the squared differences between the mean and the values from which the mean is calculated. The reason for the squaring in the calculation is due to the fact that the sum of the actual differences between the mean and the values from which the mean is calculated would always be zero.

References

Bachman L and Palmer A (2010) *Language assessment in practice*. Oxford University Press: Oxford, UK.

Bond T, Yan Z and Heene M (2020) *Applying the Rasch model: Fundamental measurement in the human sciences*. Milton Park, UK: Routledge.

Brown NA, Dewey DP and Cox TL (2014) Assessing the validity of can-do statements in retrospective (then-now) self-assessment. *Foreign Language Annals*, 47, 261–285.

Burton SJ, Sudweeks RR and Merrill PF (1991) *How to prepare better multiple-choice test items: Guidelines for university faculty*. Brigham Young University Testing Services and the Department of Instructional Science. http://testing.byu.edu/info/handbooks/betteritems.pdf

Coniam D, Lee T, Milanovic M and Pike N (2021) *Validating the LanguageCert Test of English scale: The paper-based tests*. London, UK: LanguageCert.

Council of Europe. (2001) *Common European Framework of Reference for Languages: Learning, Teaching. Assessment*. Strasbourg Cedex, France: Council of Europe.

Ebel RL (1965) *Measuring educational achievement*. Cliffs, New Jersey: Prentice-Hall.

Hughes A (2003) *Testing for language teachers*. Cambridge University Press: Cambridge, UK.

LanguageCert.org (2021) Exam Information. Available online: https://www.languagecert.org/en/language-exams/classical-greek (accessed 21 May 2021).

Lloyd ME and Hunt S (eds) (2021) *Communicative approaches for ancient languages*. London: Bloomsbury.

Messick S (1989) Validity. In RLLinn (ed) *Educational measurement*. 3rd ed. New York: Macmillan. 13–103.

Morrow K (2012) Communicative language testing. The Cambridge guide to second language assessment, 140.

Poupounaki-Lappa P, Peristeri T and Coniam D (2021) Towards a communicative test of reading and language use for Classical Greek. *Journal of Classics Teaching*, 22 (44), 98–105.

Richards JC and Rodgers TS (2014) *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.

Summers MM, Cox TL, McMurry BL and Dewey DP (2019) Investigating the use of the ACTFL can-do statements in a self-assessment for student placement in an Intensive English Program. *System*, 80, 269–287.

Wright BD (1977) Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14 (2), 97–116.

https://doi.org/10.1017/S2058631021000532 Published online by Cambridge University Press