Language Testing in Asia
a SpringerOpen Journal

RESEARCH Open Access

(Un)reliability of equivalent forms of the Taiwanese tour guide English tests

Peter J Gilks

Correspondence:
petergilks@isu.edu.tw
I-Shou University, Kaohsiung, Taiwan

Abstract

The aim of this study was to determine whether large variations in the pass rates of different versions of the Taiwanese tour guides' English test were due to inherent differences in the difficulty of the tests. Rasch analysis software was used to measure the difficulty of all items on the 2009 and 2013 versions. It was found that the ability levels corresponding to passing scores of 48, 49 and 50 (out of 80) on the 2009 test were below that required to pass the 2103 test and that it was therefore easier to pass the 2009 test. This difference was likely to have contributed to the 2009 test's significantly higher pass rate. It is argued that differences of this magnitude are avoidable, and it is recommended that the developers apply Item Response Theory to the preparation of tests in order to more reliably distinguish between those test takers whose English ability is deemed sufficient to carry out the tasks of a professional tour guide from those whose is not.

Keywords: Language testing; Tour guide; Test calibration; Rasch

Background

General backround

Taiwan's Ministry of Examinations produces two English tests each year as part of the foreign language component of a four-part certification process for domestic tour guides (導遊; *daoyou*) and tour escorts (領隊; *lingdui*), i.e., guides who accompany parties of Taiwanese tourists abroad. The tests have steadily grown in popularity over the last ten years, with over 15,000 candidates taking the tests in 2013 compared to just under 6,000 candidates in 2004 (ROC Ministry of Examinations 2013). Despite their popularity, only one study of the tests has so far been published. Using Bachman and Palmer's (1996) criteria for test "usefulness", Gilks and Trejos (2013) found that both tests were deficient in all categories under consideration except practicality. That is, although the researchers recognised the advantages of the tests' multiple choice format for large scale administration, they also found that the tests lacked reliability, construct validity, authenticity, interactiveness, and a pedagogically desirable impact on English language education.

The present study takes up the issue of reliability of the tests. A key finding by Gilks and Trejos was that large fluctuations in the pass rates were likely to be the result of variations in test difficulty rather than differences in the ability of the test takers. The tests were therefore deemed to be unreliable since a candidate who failed the test in a year when the pass rate was low may well have passed if he or she had taken the test in a year when pass rate was high. The researchers' conclusions were based on an

analysis of official data regarding overall pass rates and basic test taker characteristics such as gender, age and educational qualification but did not involve any attempt to measure or compare the difficulty of different versions of the test through an analysis of individual test taker scores. The main aim of this study, therefore, was to undertake such an analysis in order to find firmer evidence with which to support or reject the claim that, across their supposedly equivalent versions, the tests do not reliably distinguish the abilities of their test takers. Specifically, it was hypothesised that, where there was a large difference in the pass rates, there was a significant difference in the level of English ability required to achieve a threshold passing score of 48/80 (i.e., 60%).

While this study can be seen as a attempt to assess the reliability of the tour guide and tour escort tests in a more detailed manner, it must be pointed out that it is an assessment of just one particular kind of reliability, namely the consistency with which equivalent forms of a test distinguish those candidates who are deemed to have attained a certain level of proficiency from those who have not. The other common notion of test reliability, namely internal consistency, was not questioned by Gilks and Trejos and was therefore not the main focus of this study. Nevertheless, insofar issues relating to this concept of reliability are relevant, they are reviewed in the discussion on the theoretical background to the study. Moreover, since data gathered during the course of the research did allow the internal consistency of the tests under consideration to be measured, it has been noted in the results.

Theoretical background

In the first ever book published on language testing, Lado (1961) defined reliability as the dependability with which a test would yield a similar score if a student were to take the test again. Despite the many developments in communicative language testing over the years that have made Lado's structuralist approach to language assessment somewhat obsolete, his basic idea that reliability entails consistency of results remains widely accepted among many leading researchers and language test theorists, e.g., Alderson et al. (1995), Bachman (1990), Brown (1996) and Fulcher & Davidson (2007). In Classical Test Theory (CTT), the reliability of a language test is quantified as the degree to which individuals' deviation scores (or *z*-scores) remain relatively constant over repeated administration of the same test or alternate test forms (Crocker and Algina 2006). When variation does occur, factors such as test conditions, scoring methods and the test itself are said to be among the causes (Fulcher 2010). Since CTT holds that variation can be minimised when tests measure just one trait, one of the goals of language test writers has been to produce tests that are internally consistent, i.e., in all items are homogeneous in terms of what they measure. A number of techniques, such as the Guttman split-half estimate, the Kuder-Richardson coefficient and Cronbach's alpha, have been devised to calculate a tests' internal consistency (Bachman 1990; Brown 1996; Fulcher and Davidson 2007; Salmani-Nodoushan 2009). However, the underlying assumption of these estimates of reliability, namely, that a given test taker's responses on different parts of the test should be consistent with each other has been called into question by Alderson et al. (1995). Elsewhere, Anderson (1991b, cited in Alderson and Banerjee 2002) has argued that good language tests are not homogeneous since they should assess a range of different linguistic features. This is an important objection that points to deeper problems associated with the classical approach to the operationalization and measurement of reliability.

To some extent, the theoretical problems associated with measuring the internal consistency of a once-administered test can be relegated to secondary importance when considering tests that are administered regularly in several equivalent forms. Such tests may be deemed to be unreliable not only due to lack of homogeneity, but as a result of inconsistency in distinguishing candidates who have attained a certain level of mastery from those who have not (Bachman and Palmer 1996). Producing reliable equivalent versions of a test using CTT is difficult since reliability estimates are only useful for the sample on which a test is trialled and cannot be extended to other samples of students of differing proficiency (Alderson et al. 1995). In such situations, the best test developers can do to ensure reliability is to strictly adhere to detailed test specifications and/or resort to unsatisfactory *ad hoc* measures such as adjusting cut-off scores so as to ensure the percentage of test takers who pass remains constant (McNamara 1996).

For many large scale tests, Item Response Theory (IRT) now provides a satisfactory methodological alternative to CTT. In short, IRT recognises the relationship between test taker ability and item difficulty in determining the probability that an item will be correctly answered or not (Bachman 1990; Hutchison and Benton 2009). It allows for the construction of banks of items whose difficulty with respect to persons of various levels of proficiency can be accurately determined through trialling (Alderson et al. 1995; Fulcher and Davidson 2007). In the 1990s Cambridge ESOL began producing examinations using banks of items whose difficulty had been calculated using Item Response Theory combined with Rasch analysis techniques (Weir 2005). There was some debate, however, regarding the validity of the approach, with some scholars such as Goldstein & Blinkhorn (1982) and Buck (1994) objecting to the one-parameter Rasch model's unidimensional representation of the latent being measured, while others such as Henning (1989) championed the model's ability to self-regulate by identifying cases where performance did not appear to be the result of a single underlying trait. By 2000, however, the "Rasch wars" were essentially over (McNamara and Knoch 2012), with abstract technical debates over theoretical models being replaced by a pragmatic focus on practical applications (Bachman 2000), especially adaptive testing using computer technology (Bailey 1999).

There have been many applications of the one parameter Rasch model in test development. Holmes (1982) has shown how Rasch analysis can be used to vertically equate suites of tests, each of which has been designed to measure a different range of ability. Agrawal (1979) used the technique to recalibrate tests from which items that were deemed not to fit the validity construct had been removed. More recently, Rasch analysis software enabled Goh & Aryadoust (2010) to identify and remove from a test items that elicited an erratic pattern of responses (known as "underfitting" items) as well as those whose pattern of responses appeared not to be independent ("overfitting" items). Nevertheless, the main application of the Rasch analysis techniques has been to create banks of items of known difficulty from which tests can be constructed such that candidates of a certain standard of mastery consistently achieve a passing score while others below that level consistently do not (Alderson et al. 1995). It was the apparent lack of consistency is this area that led Gilks and Trejos (2013) to conclude that Taiwanese tour guide and tour escort tests were unreliable. Although there were variations in the pass rates of over 20% for the tour escort test and over 30% for the tour guide test, they did not find any significant differences in the education levels, gender

or ages of the test takers that might have accounted for such large variations in pass rates, suggesting that the by now commonplace application of IRT to trialling, calibration and item-banking had not been followed in the construction of these tests.

Methods

Data collection

The 2009 and 2013 versions of the English language component of the tour guide test were selected for comparison based on the fact that there was a large difference in the pass rates of the tests as a whole, which were 77% and 41%, respectively. Trials were then conducted on the tests using 177 subjects drawn from tourism English classes at a Taiwanese university.

Data were collected in two stages. In the first stage, the 2013 test was administered to 85 subjects. Then, the results were analysed using Rash analysis software in order to identify approximately ten suitable items from the 2013 test to be added on to the 2009 test during the second stage of data collection. The role of these items, known as 'anchors', was to function as fixed points of reference relative to which the difficulty of other items unique to either test could be calculated. What this means in practical terms, is that they enabled the difficulty measures of all items from the two tests being compared in this study to be placed on a common scale. Eight items were selected on the basis that they covered a range of difficulties and their patterns of responses closely conformed to the Rasch model's expectations in terms of discriminating between test takers' ability measures (i.e., infit mean square for the items were 0.9 to 1.1). Satisfaction of these two criteria meant that their difficulty measures could suitably serve as anchor values for calibrating the two tests.

To increase the rigour of the calibration procedure, it was decided to also use a number of person ability measures as additional anchor values. The role of these values, known as 'person anchors', was similar to that of item anchors in that they functioned as fixed points of reference relative to which the ability measures of all persons who took just one of the tests could be calculated. While such information was not in itself important for the purposes of this study, since this scale of ability measures also records the difficulty measures of all items on both tests, person anchors indirectly contribute to the calibration procedure. The theory underlying the construction of a common scale for recording both person ability and item difficulty is explained in more detail below. In order to gather suitable person anchors, a number of subjects who had completed the 2013 test were recruited to complete the 2009 test during the second stage of data collection. There were three criteria for selection. First, the data analysis had to show that their ability measures covered a broad range. Second, their patterns of responses to items on the test needed to conform to the expectations of the model in terms of getting the easier items right and the harder items wrong (i.e., infit mean squares between 0.75 and 1.25). Third, the volunteers needed to remember their 2013 test paper number. This was because all data was provided anonymously, and thus the only way of knowing which papers from each test were completed by the same person was to have the selected subjects identify their own unnamed (but numbered) test papers. In this way, nine persons were recruited to act as anchors.

Having identified suitable anchors, in stage two of the data collection process 92 subjects, including 9 persons who had already done the 2013 test, completed the extended 2009 test. The data collection procedure is represented schematically in Figure 1.

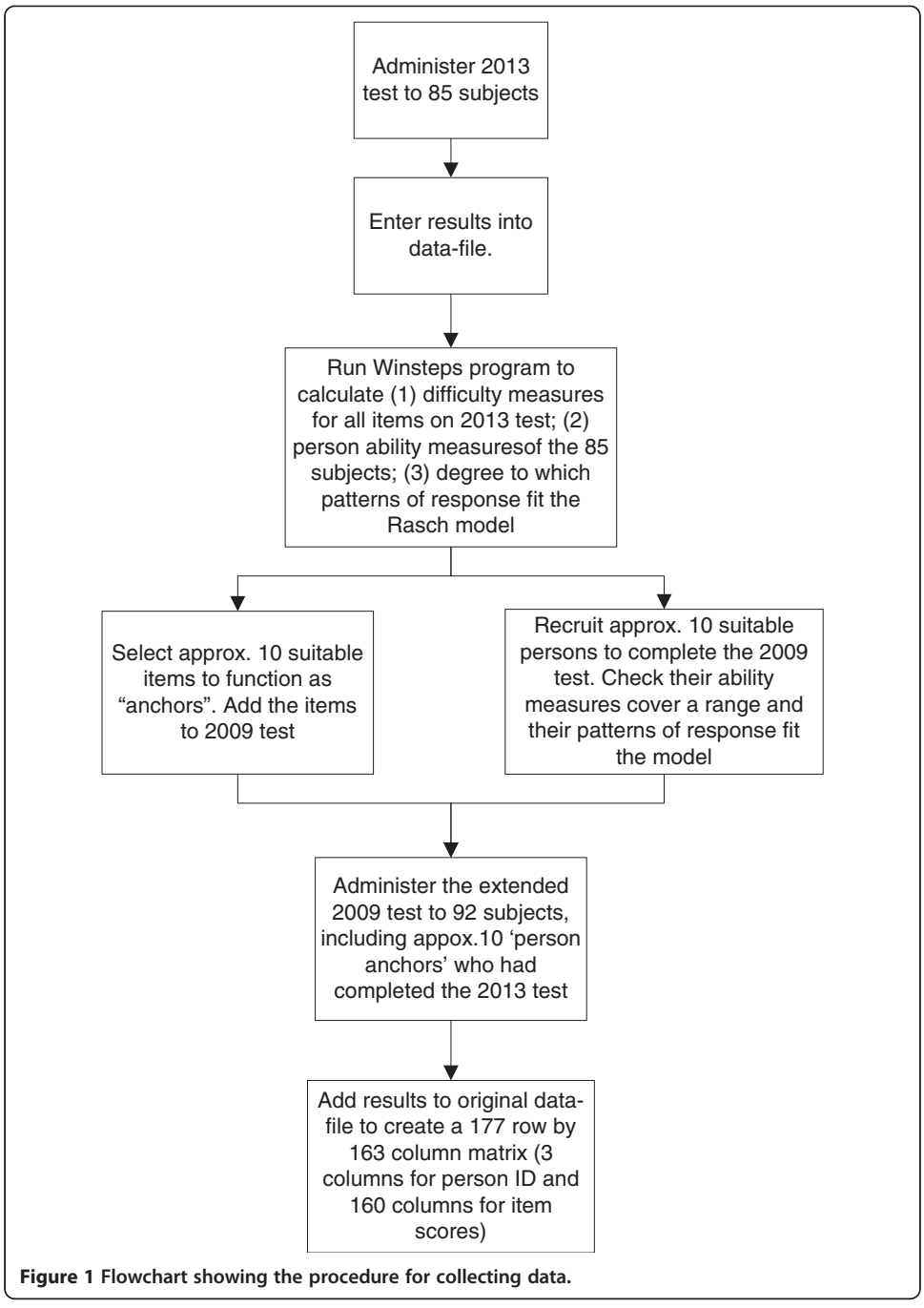

Figure 1 Flowchart showing the procedure for collecting data.

How the data were analysed

Analysis was undertaken using Winsteps, a program developed by Linacre (2012) that applies an iterative algorithm to data pertaining to each test-taker's response to each test item. These data are dichotomous: "1" means the test-taker answered correctly, while "0" means that he or she answered incorrectly. When a sufficiently large data-set is collected, the probability that a person, n, obtains a score of 1 for item, i, can be accurately calculated. This probability is expressed as P_{ni1}. Then, for each item and person, the algorithm calculates $\log_e(P_{ni1}/\ P_{ni0})$, where P_{ni0} is the probability that person, n, answers item, i, incorrectly. In theory, the measures, known as logits ("log odd units"),

could extend from -∞ to + ∞, but in practice, the range of ±5 covers all probabilities from 0.01 to 0.99 of answering a given item correctly.

The Rasch model assumes a relationship between item difficulty and a test taker's ability to answer the item correctly. This is expressed as:

$$\log_e(P_{ni1}/P_{ni0}) = B_n - D_i$$

where D_i is the difficulty of test item i and B_n is the ability of test taker n with regard to the latent trait being measured. Thus, when the probability that person n gets item i correct is 0.5, then the probability that he or she gets the item wrong will also be 0.5. In such cases,

$$P_{ni1}/P_{ni0} = 1$$

Since $\log_e(1) = 0$, the difference between D_i and B_n must also be zero. In other words, *when the probability that a person answers a given item correctly is 0.5, the item difficulty is said to equal the person's ability.* In this way, item difficulty and test taker ability can be mapped to a single scale.

Usually calibration of tests involves two stages of calculation. First, item difficulty and person ability measures are calculated for Test A. Then, a number of items and/or persons are selected as anchor values around which item difficulty and person ability measures from Test B are calibrated. However, it is also possible to use a one-step process. Although it is still necessary to have a number of persons who complete both tests and/or a number of items common to both tests, all the data are recorded in one file so that a single execution of the Winsteps program is used to simultaneously calculate the difficulty measures and ability measures for all items and persons. Then, the program's IDELETE function can be used to deselect groups of items so that score tables, which estimate the abilities corresponding to raw scores, only show results for items for one particular test. Since every difficulty measure is directly calculated in relation to data from every item from every test taker, this method provides maximum accuracy. When data from many tests and thousands of test takers are involved, data files may become too large and unwieldy for this method to be used. However, in this study, which involved just two tests of 80 items each and 177 persons, the data were not unmanageable, so the single-step process was followed.

Justification for Rasch analysis instead of *t*-test

While perhaps the most obvious method to determine whether two versions of a test are equally difficult would be to administer both versions to the same group of subjects under the same conditions and conduct a *t*-test to determine whether there was any significant difference in the pass rates, this method was not chosen because it was envisaged that data collected in the present study could be used in future studies to calibrate other versions of the test besides those of 2009 and 2013. Since it would have been impossible to gather data from the same group of subjects, Rasch analysis was chosen since it allowed for the results from other trial populations to be compared.

Results and discussion

Tables 1 and 2 are score tables that show the ability/difficulty measure and standard error corresponding to every possible score on each test. These results are represented graphically in Figure 2. They reveal a small but consistent difference in difficulty of

Table 1 Table of measures on the 2013 tour guide test (Test A)

MEASURE		SCORE S.E.	SCORE MEASURE		S.E.	SCORE MEASURE		S.E.
0	−6.00	1.83	27	-.75	.25	54	.93	.26
1	−4.77	1.01	28	-.69	.25	55	1.00	.26
2	−4.05	.72	29	-.62	.25	56	1.07	.26
3	−3.62	.60	30	-.56	.25	57	1.14	.27
4	−3.30	.52	31	-.49	.25	58	1.22	.27
5	−3.05	.47	32	-.43	.25	59	1.29	.27
6	−2.84	.44	33	-.37	.24	60	1.37	.28
7	−2.66	.41	34	-.31	.24	61	1.45	.28
8	−2.49	.39	35	-.25	.24	62	1.53	.29
9	−2.35	.37	36	-.19	.24	63	1.62	.29
10	−2.22	.35	37	-.12	.24	64	1.71	.30
11	−2.09	.34	38	-.06	.24	65	1.80	.30
12	−1.98	.33	39	-.00	.24	66	1.90	.31
13	−1.87	.32	40	.05	.24	67	2.00	.32
14	−1.77	.31	41	.11	.24	68	2.11	.33
15	−1.68	.30	42	.17	.24	69	2.22	.34
16	−1.59	.29	43	.23	.24	70	2.35	.35
17	−1.50	.29	44	.29	.24	71	2.48	.37
18	−1.41	.28	45	.35	.24	72	2.63	.39
19	−1.33	.28	46	.41	.24	73	2.79	.41
20	−1.25	.27	47	.48	.25	74	2.98	.44
21	−1.18	.27	48	.54	.25	75	3.19	.48
22	−1.10	.27	49	.60	.25	76	3.44	.53
23	−1.03	.26	50	.67	.25	77	3.76	.60
24	-.96	.26	51	.73	.25	78	4.20	.72
25	-.89	.26	52	.80	.25	79	4.92	1.01
26	-.82	.26	53	.86	.25	80	6.15	1.83

around 3 raw score points across the middle range of scores. Specifically, a raw score of 48 on the 2013 test (which had a low pass rate) corresponded to a difficulty measure or 0.54 logits, whereas a raw score on the 2009 test (which had a high pass rate) corresponded to a lower difficulty measure of 0.39 logits. Thus, candidates who passed the 2009 test with scores of 48, 49, or 50 would probably not have passed the 2013 test. In contrast, candidates who failed the 2013 test with scores of 46 or 47 would probably have passed the 2009 test.

Although the results do conform to expectations in terms of showing that it was harder to pass the 2013 test than pass the 2009 one, the difference in difficulty was not as great as might be expected given the large difference in the pass rates. There are two possible explanations.

The first is that the pass rates of 77% for the 2009 test and 41% for the 2013 test are not based solely on the tests analysed in this study but include tests of three other areas of knowledge, all of which were administered in Chinese. These three other tests cover: (1) practical matters related to working as a tour guide, such as travel safety, ticketing, tourist psychology, etc.; (2) regulations pertaining to the relationship between Taiwan

Table 2 Table of measures on the 2009 tour guide test (Test B)

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
0	−5.98	1.83	27	-.85	.25	54	.77	.25
1	−4.76	1.01	28	-.78	.25	55	.83	.26
2	−4.05	.72	29	-.72	.24	56	.90	.26
3	−3.62	.59	30	-.66	.24	57	.97	.26
4	−3.31	.52	31	-.60	.24	58	1.04	.26
5	−3.06	.47	32	-.54	.24	59	1.12	.27
6	−2.86	.43	33	-.48	.24	60	1.19	.27
7	−2.68	.40	34	-.42	.24	61	1.27	.28
8	−2.52	.38	35	-.36	.24	62	1.35	.28
9	−2.38	.36	36	-.31	.24	63	1.44	.29
10	−2.25	.35	37	-.25	.24	64	1.52	.29
11	−2.14	.33	38	-.19	.24	65	1.61	.30
12	−2.03	.32	39	-.13	.24	66	1.71	.31
13	−1.92	.31	40	-.07	.24	67	1.81	.32
14	−1.83	.30	41	-.02	.24	68	1.92	.33
15	−1.73	.30	42	.03	.24	69	2.03	.34
16	−1.65	.29	43	.09	.24	70	2.16	.35
17	−1.56	.28	44	.15	.24	71	2.29	.37
18	−1.48	.28	45	.21	.24	72	2.44	.39
19	−1.40	.27	46	.27	.24	73	2.60	.41
20	−1.33	.27	47	.33	.24	74	2.78	.44
21	−1.25	.26	48	.39	.24	75	2.99	.48
22	−1.18	.26	49	.45	.24	76	3.25	.53
23	−1.11	.26	50	.51	.24	77	3.57	.60
24	−1.04	.25	51	.57	.25	78	4.00	.73
25	-.98	.25	52	.64	.25	79	4.73	1.01
26	-.91	.25	53	.70	.25	80	5.96	1.83

and mainland China; (3) Taiwanese history and geography. Like the language test, each of these tests consists of 80 multiple choice items. If the tests were not constructed from banks of trialled items, then there may have been differences in difficulty that exerted a greater influence on overall pass-rates than that of the language component, as revealed in this study. For example, a particularly difficult test of Taiwanese history and geography on the 2013 battery could have been a greater factor contributing to the low pass rate than a slightly more difficult English test.

The smaller than expected difference may also have been the result of guessing. Unlike the actual candidates who sat the examinations in 2009 and 2013 and for whom the test was presumably a high stakes gateway to future employment, the subjects used in this study received no qualification, even if they achieved a high score. Many of them were observed to complete the test in half the allotted time, suggesting that they had not given the questions the same amount of consideration as actual test takers are likely to have given to them. If many subjects had resorted to guessing in order to finish quickly, this would have reduced the effect that any inherent differences in difficulty between the two tests would have had on the results. Some support for such an explanation comes from

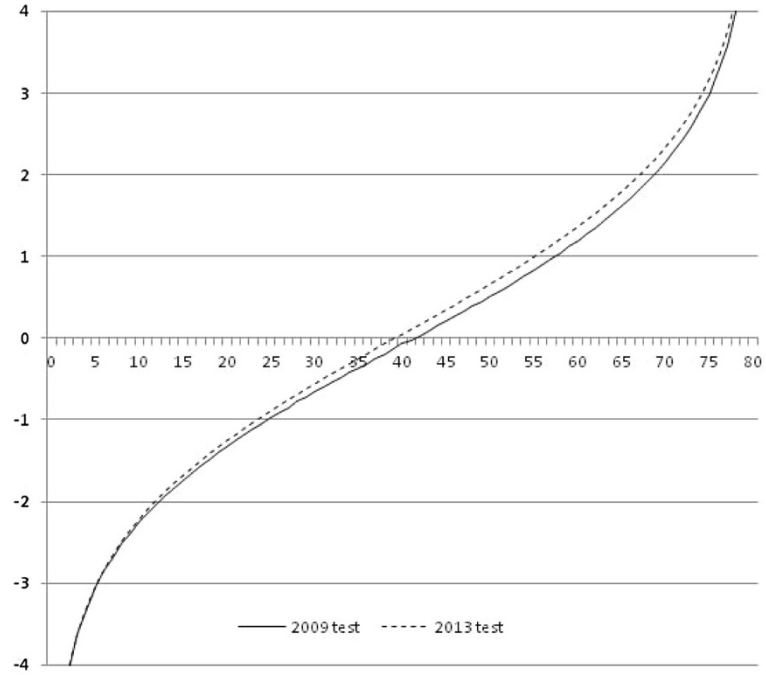

Figure 2 Difficulty measures for raw scores on the 2013 and 2009 tour guide tests compared.

the raw scores of the nine person anchors. Although their ability level was constant, there was a difference in their mean scores for both tests of 5.2 points (see Table 3). We know that these test takers had not resorted to large-scale guessing since, as explained above, one of the criteria for selecting these persons as anchors was that their patterns of responses to items on the 2013 test closely conform to the expectations of the Rasch model.

As a final note, it was observed that Winsteps returned a Cronbach's alpha score of 0.91 for the 2009 test and 0.92 for the 2013 test, showing that both tests had a high degree of internal consistency. In other words, considered individually, both tests were reliable in the classic sense. This result is not unexpected in light of Gilks and Trejos' (2013) analysis, which revealed that the overwhelming majority of items on recent versions of the tour guide tests only measure vocabulary knowledge.

Table 3 Raw scores for the nine person anchors on both tests

Year	2009	2013
	41	37
	42	47
	45	40
	57	48
	57	48
	58	53
	61	51
	64	63
	75	67
Mean	55.6	50.4

Conclusions

Given that many thousands of young Taiwanese take the tests each year hoping to obtain a certificate that would serve as an important qualification for work in the growing Taiwanese tourism industry, it is not unreasonable to expect the Ministry of Examinations to produce tests whose pass rates do not fluctuate greatly from year to year. Differences in test difficulty, which may possibly be larger than the results of this study indicate, are easily avoidable through the application of IRT to trailing of potential test items, item banking, and the compilation of new tests whose difficulty curves closely coincide with an established standard.

Although such a procedure would ensure that different versions of the test more consistently distinguish "masters" from "non-masters" at the same level of ability, it would not provide any guarantee that successful candidates really do have sufficient English proficiency to perform the tasks of a professional tour guide. In order to achieve that, reference must be made to external criteria such as the opinions of relevant stakeholders, including tour companies, tourists, and tour guides regarding what standard is considered minimally acceptable for professional tour guides. Then, subjects whose English ability is deemed to meet this standard could be used as part of a trialling process to establish a scale onto which the difficulty measures of potential test items could be placed. Subsequently, these items could be selected for inclusion in new versions of the test such that the ability required to achieve a threshold passing score remained constant at that deemed to be the minimum acceptable standard. It is hoped that this research will prompt the tests' developers to move in that direction and adopt what is now standard practice for large scale language testing.

Competing interests
The author declares that he has no competing interests.

Received: 26 September 2014 Accepted: 22 October 2014
Published online: 07 November 2014

References

Agrawal, KC. (1979). The "Short Tests of Linguistic Skills" and Their Calibration. *TESOL Quarterly, 13*(2), 185–208.
Alderson, JC, & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching, 35*(2), 79–113.
Alderson, JC, Clapham, C, & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.
Bachman, LF. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
Bachman, LF. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing, 17*(1), 1–42.
Bachman, LF, & Palmer, AS. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.
Bailey, KM. (1999). *Washback in Language Testing*. Princeton: Educational Testing Service.
Brown, JD. (1996). *Testing in Language Programs*. Upper Saddle River, NJ: Prentice Hall Regents.
Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing, 11*(3), 145–70.
Crocker, L, & Algina, J. (2006). *Introduction to Modern and Classical Test Theory*. Mason, OH: Cengage Learning.
Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.
Fulcher, G, & Davidson, F. (2007). *Language Testing and Assessment*. London and New York: Routledge.
Gilks, P, & Trejos, B. (2013). A critical look at official tour guide English examinations in Taiwan. In *Proceedings of the International Conference on English Teaching at Universities of Technology,* Cheng Shiu University, Kaohsiung.
Goh, C, & Aryadoust, SV. (2010). Investigating the Construct Validity of the MELAB Listening Test through the Rasch Analysis and Correlated Uniqueness Modeling. *Spaan Fellowship Working Papers in Second of Foreign Language Assessment, 8*, 31–68.
Goldstein, H, & Blinkhorn, S. (1982). The Rasch model still does not fit. *British Educational Research Journal, 8*(2), 167–70.
Henning, G. (1989). Does the Rasch model really work for multiple-choice items? Take another look: a response to Divgi. *Journal of Educational Measurement, 26*(1), 91–7.
Holmes, SE. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement, 19*(2), 139–47.
Hutchison, D, & Benton, T. (2009). *Parallel Universes and Parallel Measures: Estimating the Reliability of Test Results*. Coventry: Office of the Examinations and Qualifications Regulator.

Lado, R. (1961). *Language Testing: the Construction and use of Foreign Language Tests: a Teacher's Book*. London: Longman.
Linacre, JM. (2012). *Winsteps 3.74.0 Rasch Measurement Computer Program*. Chicago: Winsteps.com.
McNamara, T. (1996). *Measuring Second Language Performance*. London and New York: Longman.
McNamara, T, & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555–76.
ROC Ministry of Examinations. (2013). 各種考試統計 (Examination Statistics). http://wwwc.moex.gov.tw/main/ExamReport/wFrmExamStatistics.aspx?menu_id=158.
Salmani-Nodoushan, MA. (2009). Measurement Theory in Language Testing: Past Traditions and Current Trends. *Journal on Educational Psychology, 3*(2), 1–12.
Weir, CJ. (2005). *Language Testing and Validation*. Basingstoke: Palgrave MacMillan.

doi:10.1186/s40468-014-0007-8
Cite this article as: Gilks: (Un)reliability of equivalent forms of the Taiwanese tour guide English tests. *Language Testing in Asia* 2014 **4**:7.