The role of language background in the validation of a computer-adaptive test[1]

Annie Brown and Noriko Iwashita
The University of Melbourne

The use of IRT analysis has greatly facilitated the development of computer adaptive tests, where the adaptiveness is based on measures of item difficulty resulting from the performance of trial test takers. However, studies into the acquisition of L2 grammar by learners with different L1s indicate that the learners' L1 strongly influences their acquisition of grammar in the L2. Thus, it would be expected that grammar test items would present different levels of difficulty to test takers from different language backgrounds. Where a computer-adaptive grammar test is to be used with test takers from a range of language backgrounds it is, therefore, questionable whether set item difficulty measures can validly be used for all types of test taker.

The study investigates the performance of learners of Japanese from different language backgrounds, using data from a computer adaptive grammar test developed as a placement tool. The trial pen-and-paper test consisted of 225 multiple choice items. 1600 students in Australia, China and Japan and Korea (all of whom had studied Japanese for between 150 and 300 hours) each completed 50 items.

In this study, data is presented from native speakers of English, Chinese and Korean. Item difficulties drawn from the trialling were found to be quite different for the three groups of test taker. This has implications for the validity of use of computer-adaptive tests, in that where actual test takers are from a different background from that of the trial population, not only does the test fail to measure such test takers accurately in that unacceptable percentages of test takers are found to misfit, but also the measures of ability provided

[1]This research arises out of a project funded by the University of Melbourne and was made possible through the support of the National Languages and Literacy Institute of Australia. The paper was presented at the Language Testing Research Colloquium, Long Beach, California, March 1995. An earlier version of the paper, investigating learners of two language backgrounds, Chinese and English, has been published in *System*, Vol. 24,2 (1996).

Melbourne Papers in Language Testing 1996 Volume 5.1 pp. 19-34.
The Language Testing Research Centre, The University of Melbourne.

for each test taker and their relative rankings differ according to the set of item difficulties used, and will consequently affect decisions made about individual learners regarding placement or selection.

1. Introduction

The process of establishing test validity is one of the basic concerns of language testing. More recent concepts of validity see it as an 'integrated evaluative judgement' (Messick 1993: 13), based on and requiring the collection of evidence from a variety of perspectives. This evidence may take many forms and include statistical information, expert opinion, feedback from test takers and information on the impact of the test in the wider systemic sense. Nevertheless, perhaps the main approach to the establishment of a test's internal validity involves the gathering of evidence about the measurement properties of the test items and the responses produced by test takers. An issue which arises in relation to the measurement properties of items, of course, is the extent to which they are stable across groups of test takers and the issue of test bias. This paper considers the role of language background in differential item performance, and the validity implications of the effect that this might have on outcomes for candidates.

2. Influence of the L1

The influence of a learner's L1 on his/her L2 performance was first investigated in the 1950s within the paradigm of the Contrastive Analysis Hypothesis, which states that a learner's L1 plays a decisive role in the learning of an L2 and that the differences and similarities of the L2 to the L1 predict difficulties in learning the L2. More recent research, however, recognises that the influence of the L1 in learning an L2 is not necessarily born out of the similarities and differences between the L1 and the L2, and that the influence of the L1 on the L2 is an extremely complex issue.

Zobl (1982, 1983) and Ellis (1994) present a number of constraints of L1 transfers on L2 learning. While constraints identified by Zobl (1982, 1983) focus on linguistic features such as congruence of L1 form with L2's developmental structure, markedness and zero contrast, those identified by Ellis (1994) incorporate a broader ranges of factors and include language level, sociolinguistic factors,

markedness, prototypicality, language distance and developmental factors. Among the constraints proposed by Zobl and Ellis, language distance and zero contrast are the most relevant to the present study which examines the relationship between a test-takers' L1 (Chinese, English or Korean) and their performance on the Japanese grammar test.

There is substantial evidence that the distance between the L1 and L2 acts as a constraint on transfer. Corder (1981) proposes a language distance hypothesis, claiming that the mother tongue acts differentially as a facilitating agent, and the more similar the L1 is to the L2, the more rapidly learners acquire the L2. Odlin (1989) cites the different amounts of time which the Foreign Service Institute in the United States allocates to courses aimed at achieving a high level of proficiency in different languages, and claims that language distance is a major 'determinant of the amount of time students will need to become highly proficient in a language.' (Odlin 1989; 153).

The findings of a previous study (Brown & Iwashita 1996) upon which the present study builds, also support Corder's language distance hypothesis. The study investigated the performance of test takers from two different language backgrounds (Chinese and English), using data from a computer adaptive Japanese grammar test developed as a placement tool. Chinese is typologically closer to Japanese than English is, and, as expected, the Chinese test-takers performed significantly better than the Australian test-takers, despite having received the same number of hours of instructions when the test was administered.

Another finding of this study was that certain item difficulties were found to be quite different for the two groups of test-takers. These differences are explained by one of the constraints identified by Zobl, zero contrast, which would result in items being more difficult where the feature does not exist in the L1 than where it does exist in the L1. For example, the item difficulties (in logits, derived using an IRT analysis) of items testing verb forms were found to be higher for the Chinese group than for the Australian group. Japanese verbs have inflectional morphemes to indicate tense as in English (though the way the inflectional morpheme is added to the basic form of verbs is more complex than in English). Chinese languages, on the other hand, do not have inflectional verb

morphemes as in English and Japanese. Thus, as would be expected, the items were easier for those candidates familiar with the feature in their L1. Item difficulties for items testing particles also differed significantly for the Chinese and Australian groups.

These findings demonstrate that a learner's L1 can exert a powerful influence on the L2 performance. Learners from different L1 backgrounds will find different linguistic features easy or difficult according to the L1's similarity to or difference from the target language, in respect of the features.

This 'relativeness' of item difficulty is of obvious concern in the development of tests, in relation to test difficulty and bias. There is a growing body of research into the ways in which items function differently for different groups of candidates; this is generally termed DIF (differential item functioning). While DIF has the potential to tell us much about group differences in test performance and language acquisition, in a computer-adaptive test, where the measures of test taker ability are calculated on the basis of fixed measures of item difficulty, differential item functioning has the potential to affect the outcomes if the item difficulties for particular items do not reflect the real difficulty for that candidate or group of candidates. This study investigates the impact of differential item functioning in a computer-adaptive test on three such groups of language learners, each from a different language background and with an L1 at differering degrees of distance from the target language, Japanese.

3. Computer-adaptive tests

Computer adaptive tests are based upon the existence of a bank of items, all calibrated on a single ability-difficulty scale, against which the items are ranged and on which the test takers will be placed; Item Response Theory provides the tool by which this can done. The item difficulties are pre-programed into the CAT, and are derived from large-scale administration of the items to a trial population, typically at this stage in pen-and-paper format.

By matching as closely as possible the difficulty of the test item to the ability of the test taker, it is claimed (Weiss 1990) that computer-adaptive tests are more efficient measures of ability than are standard pencil-and paper tests, where many of the items may

be far above or below the ability of the candidate and hence provide little information about the candidate's ability. Thus a computer-adaptive test is said to require fewer items than a conventional test to estimate the candidates' ability. This 'greater precision of measurement', is claimed by Weiss not only to be more efficient, but also to translate into 'more accurate mastery classification' (1990:454) .

A practical advantage of computer adaptive tests is that test security is enhanced, as it is unlikely that two test takers would receive the same items in the same sequence. It is also extremely unlikely that a test taker will encounter the same items on successive administrations, thus allowing the test to be used over and over again on the same students. As a consequence, such tests, despite, or perhaps because of being costly to produce, can be expected to be marketed widely. It may be that a Spanish test developed in Australia and trialled on Australian learners will be made available on a commercial basis to institutions where Spanish is taught in other countries—Japan, for example. This, however, brings us to the question which forms the basis of our research here: to what extent do the characteristics of the trial population affect the item difficulty measures, and to what extent does this invalidate the use of the CAT with test takers of different backgrounds? This paper investigates the applicability of one set of item difficulty measures, one 'model', to test takers with different characteristics.

4. Methodology

The test used in this study is a computer adaptive grammar test which was developed by three experienced teachers of Japanese as a placement tool for incoming language students at the University of Melbourne who had studied Japanese prior to entering the university. The test consists of a bank of 225 multiple choice items which test knowledge of verb and adjective forms, conjunctions, particles, structural nouns and so on. (See Apppendix 1 for sample items.) Before the test was computerised, it was trialled on approximately 1700 learners of Japanese in Australia, China, Japan and Korea. Each trial test taker completed 50 selected items in pencil-and-paper format.

Three groups of students who took part in the trials were chosen for this study. One group consists of all the Australian students of an

English speaking background (N=650). The second group[2] consists of native Mandarin speakers in China (N=451), and the third group consisted of Korean native speakers in Korea (N=351). Test takers in all groups had had between 150 to 300 hours of formal instruction in Japanese when the test was administered. Nevertheless, a comparison of performance across the three groups using a one-parameter IRT model, QUEST (Adams & Khoo 1990) revealed substantial differences (Table 1), with the Korean students demonstrating a considerably higher level of ability than the other two groups.

	N	Mean score (logit value)	Std. Dev.	Std. Err.
English	650	.539	1.145	.045
Chinese	451	.898	1.147	.054
Korean	343	2.494	1.175	.063

Table 1. Comparison of performance—descriptive statistics

An ANOVA (Tables 2a and 2b) confirmed that the three groups were significantly different in ability. Further analysis using QUEST revealed that more Korean test takers were found to misfit (9%) than English-speaking (4%) or Chinese (5%). 'Fit' is a measure in IRT analysis, which indicates the extent to which the model 'fits' with the test takers' patterns of abilities, in other words, how well the test is able to measure their ability. The model is built on the basis of the pattern of responses by trial test takers. People who do not match this pattern will misfit. If there are a substantial number of misfitting test takers within a test population, this indicates that there are possibly two (or more) populations and the model should not be used with both of them.

[2]A substantial minority of the Australian students were bilingual Chinese Australians. These were excluded from the study on the basis that their performance will be expected to be different again, but they cannot be considered as a homogeneous group as they will vary in proficiency across English and Chinese according to their family backgrounds.

	DF	Sum of squares	Mean square	F-value	P-value
Lang	2	889.554	444.777	334.579	<.0001
Residual	1441	1915.611	1.329		

Table 2a. Comparison of performance—ANOVA

	Mean Diff	Crit. diff	P-value
Korean, Australian	1.956	.151	<.0001
Korean, Chinese	1.596	.162	<.0001
Australian, Chinese	-.359	.139	<.0001

Table 2b. Comparison of performance—Fisher's PLSD

Data from the performances of these students was used to test out three hypotheses.

4.1. Hypothesis 1

Item difficulties will differ substantially for test takers of the three language backgrounds, as a result of specific features of each language being more or less distant from the target language.

If this is found to be the case, it may affect the validity of using IRT-based tests with test takers of backgrounds different to that of the trial population if the further two hypotheses are confirmed:

4.2. Hypothesis 2

More test takers will misfit when item difficulties are developed on the basis of performance by trial test takers of a different language background.

Where a test taker is found to be misfitting, one cannot rely on the measure of ability given for that test taker to be a fair and accurate measure. Obviously, if many test takers are found to be misfitting, then the test as it stands (with this particular model of item difficulty) is clearly not an appropriate tool.

4.3. Hypothesis 3

Test takers will be ranked differently when item difficulties are developed on the basis of performance by different-language background trial test takers compared with when they are developed on the basis of performance by same-language background trial test takers.

When a computer-adaptive test is used, ability measures for test takers are based on scores derived from performance on selected items with pre-programed difficulty levels. If these difficulty levels differ, it is likely that the same performance will result in a different estimate of ability which will translate in practical terms to different ranking and placement

5. Results

5.1 Hypothesis 1

Measures of difficulty were produced for all 225 items for each group using the program QUEST (Adams & Khoo 1990). In order to determine the extent of agreement between the groups, the item difficulties were correlated using Pearson's r (Table 3)[3]. Correlations of this size, while significant, indicate a substantial mismatch between the relative item difficulties for the three groups.

[3]A standard procedure for the analysis of differential item functioning is the Mantel-Haenszel procedure. This was not used in this instance as it requires comparison of individual items across test takers of equivalent ability levels, yet in this instance candidates completed only a selection of the total number of items and differences in mean ability were found between the three groups. An IRT analysis, however, overcomes these problems in that it allows for missing data and all item difficulties are meaned to 0.

	Australian (N=650)	Chinese
Chinese (N=451)	.507*	
Korean (N=351)	.553*	.601*

*significant at p<.01

Table 3. Item difficulty correlations (r)

We then grouped the items according to the type of grammatical feature they tested—verb form, particle, adjective form, structural noun and conjunction. These are common but complex features of Japanese syntax and hence each had several items dedicated to it[4]. The mean difficulty of each category was calculated and they were ranked for each language background group in order to determine whether particular item types were relatively more or less difficult for learners of one language background than for the others (Table 4). We found that while particles were the easiest grammatical category for all learners and conjunctions were the most difficult, the other three categories varied in their relative difficulty across the three language background groups.

English	Chinese	Korean	
Conjunction	Conjunction	Conjunction	more difficult
Adjective form	Verb form	Structural noun	↕
Structural noun	Structural noun	Verb form	
Verb form	Adjective form	Adjective form	
Particle	Particle	Particle	easiest

Table 4. Relative difficulty of item types

[4]Although there were also other features of Japanese syntax included in the test, as these each had only one or two items, they were not included in this analysis.

5.2. Hypothesis 2

In order to establish whether Hypothesis 2 applied, we needed to investigate the effect of trialling the test on test takers of one language background to set item difficulties, and then, using these difficulties, administering the test to:

a) test takers of the same language background, and

b) test takers of a different language background

As it was not possible at this point to use the actual computer-adaptive version of the test, a simulation was set up using the pen-and-paper trial data. A random 144 subjects were removed from each group to become the simulated 'test test takers. The remaining test takers then became the simulated 'trial population' from which to define the item difficulties for each language background. Item difficulties were produced for all 225 items for each trial population. Then, after removing misfitting items as would be done under normal trialling procedures using these item difficulties as the benchmarks, the performance of the 144 simulated 'test takers' from the Australian, Chinese and Korean groups was analyzed and the percentage of misfitting test takers in each group was calculated (Table 5).

		Item difficulty values derived from: English	Chinese	Korean
Simulated test population	English	5.5%	32.0%	61.0%
	Chinese	31.0%	6.0%	50.0%
	Korean	11.0%	3.0%	12.0%

Table 5. Misfitting test takers

When we anchored the item difficulties derived from the three background groups and 'tested' the English-speaking students, 5% of test takers misfitted when the English-derived item difficulties were used. However, when the Chinese-derived item difficulties were used 32% of test takers misfitted and 61% when the Korean-

derived item difficulties were used. Similarly, when we 'tested' the Chinese students, only 6% test takers misfitted when the Chinese-derived item difficulties were used, but 31% did when the English-derived item difficulties were used and 50% when the Korean-derived item difficulties were used. These results conformed to the expectations of Hypothesis 2. However, the pattern of fit of the Korean test takers did not meet the expectations of Hypthesis 2; fewer test takers misfitted when the Chinese-derived item difficulties were used (3%) than when the Korean-derived item difficulties were used (12%), and a similar number misfitted when the English-derived item difficulties were used (11%). The only explanation we can offer for this is that the test as a whole does not appear suited to this Korean test population, as we saw earlier where the test was overall too easy, several test takers obtained perfect scores and a relatively large percentage (9%) misfitted.

5.3. Hypothesis 3

Assuming the test is used as a placement test (its original purpose) a practical concern is whether the test sorts test takers in the same way using item difficulties obtained from a trial population of the same or different background. So we compared the rankings of test takers in each group using the item difficulties produced when

a) the test was 'trialled' on the same background group

b) the test was 'trialled' on the other background groups

The rankings of the 144 test takers using item difficulties derived from other same-background learners were correlated with the rankings derived from different-background learners. Whilst these correlations were all highly significant, with Spearman's Rho values of over .9, we decided to simulate the placement procedure resulting from these rankings in order to find out how many students would be placed into different classes if differing item difficulties were used. We compared the division of the 144 students into 6 classes of 24 when it was done on the basis of same-language background item difficulties and other-language background item difficulties (Table 6). We found that for the Chinese test takers, when the English-derived item difficulties were used instead of the Chinese ones, 29% were placed differently, and when the Korean-derived item difficulties were used, 39% of test takers were placed

differently. Similarly, for the Australian test takers, when the Chinese-derived item difficulties were used instead of the English-derived ones, 24% were placed differently, and when the Korean ones were used 26% were placed differently. When it came to the Korean students, when English-derived item difficulties were used instead of Korean ones, 34% were placed differently, and when Chinese-derived item difficulties were used 35% were placed differently.

	English speakers using Chinese-derived item difficulties	English speakers using Korean-derived item difficulties
English speakers using English-derived item difficulties	24%	26%

	Chinese speakers using English-derived item difficulties	Chinese speakers using Korean derived item difficulties
Chinese speakers using Chinese-derived item difficulties	29%	39%

	Korean speakers using English-derived item difficulties	Korean speakers using Chinese-derived item difficulties
Korean speakers using Korean-derived item difficulties	34%	35%

Table 6. Placement: percentage of test takers placed differently according to item difficulties used

6. Discussion and conclusion

In this study, we found that some features of Japanese grammar are relatively more difficult for speakers of one L1 than for speakers of a different L1 (Hypothesis 1). We also found that language

background can play a role in the rate of acquisition of a language, Koreans with an equivalent amount of classroom exposure having a much higher mean proficiency than the Australian and Chinese learners. An implication of this is that a test cannot necessarily be marketed as suitable for all learners who have had a particular amount of exposure—language background needs to be taken into account here.

At this point we need to comment more on the performance of the Korean test takers. We suspect that the test as it stands is not suitable for them. It is overall too easy a test and does not appear to measure them well, an unacceptably high percentage of test takers (9%) having been found to misfit when the test was administered and analyzed using only Korean learners. It may be that the test would be more suitable with learners of a proficiency more comparable to that of the Chinese and Australian test takers, in other words test takers who have completed fewer hours of study of the language. However, this requires further research.

In relation to Hypothesis 2, many more Australian and Chinese test takers were found to misfit where the item difficulties used were based on performance by trial test takers of a different L1 background. In other words, the model upon which the CAT was developed was not appropriate for an unacceptably high number of test takers. Where test takers do not fit the model it is likely that they will be required to complete more items before the test can come up with an estimate of their ability, in other words the CAT is not likely to be as efficient as it could be. The extent to which this is the case will be the subject of a further investigation.

Furthermore, where test takers are found to misfit, the resulting test taker ability measure cannot be relied upon to be accurate. This leads on to findings in relation to Hypothesis 3, that the test takers are in fact ranked differently according to the model used, thus negating the claim made for CATs that they provide 'more accurate mastery classification' (Weiss 1990).

The validity of a test is dependent not only on its internal features, but on the use that is made of it in particular situations and for particular purposes. Messick refers to this as 'population generalizability', which he defines as 'the extent to which a measure's construct interpretation empirically generalizes to other

population groups' (1989: 56). No matter how internally sound a test may be, its suitability must be investigated for all potential test users. Differential item functioning is one growing area of research in construct validation which demonstrates how different groups may differ in their distribution of skills. Where, as is the case with computer adaptive tests, a model of item difficulty is built into the actual test, then of central concern must be the extent to which this model can be generalized and applied to groups and individuals different from the one on which the model was constructed, and who demonstrate different patterns of language acquisition. We believe we have demonstrated here that population generalizabilty is not feasible with computer-adaptive tests.

Recent discussions within the language testing field have focused on the ethics of test development and the responsibilities of test developers. The Code of Fair Testing Practices in Education (Joint Committee on Testing Practices 1988) has been proposed as the basis for a code of ethics for ILTA. It names as one of its principles for developing and selecting tests that 'test users should select tests that meet the purpose for which they are to be used and that are appropriate for the intended test-taking populations' (emphasis added). Studies such as ours point to problems with this which are inherent in the use of CATs with learners of different backgrounds. While the obvious solution would be to trial the test across a range of language backgrounds and to set the item difficulties for each population accordingly, this would complicate the test development process enormously, and would still not solve the problem of what to do where a cohort is of mixed language backgrounds. Whatever path is taken, however, it is important that test developers are aware of the implication of the findings such as those reported here.

References

Adams, R & Khoo, S. T. (1990) *Quest: The Interactive Test Analysis System*. Australian Council for Educational Research: Hawthorn, Vic.

Brown, A. & Iwashita, N. (1996) 'Language background and item difficulty: the development of a computer-adaptive test of Japanese'. *System* 24, 2, 199–206.

Corder, S.P. (1981) *Error Analysis and interlanguage.* Oxford: Oxford University Press.

Ellis, R. (1994) *The study of second language acquisition.* Oxford: Oxford University Press

Joint Committee on Testing Practices (1988) *The Code of Fair Testing Practices in Education.* Washington, DC: American Psychological Association.

Messick, S. (1989). 'Validity.' In Linn, R.L. (ed) *Educational Measurement.* (pp.13–103). Phoenix, Az.: The Oryx Press.

Odlin, T. (1989) *Language Transfer.* Cambridge: Cambridge University Press

Weiss, D.J. (1990) 'Adaptive testing'. In Walberg, H.J. & Haertel, G.D. (eds) The International Encyclopedia of Educational Evaluation, 1990 (pp. 454–458). Oxford: Pergamon Press.

Zobl, H. (1982) 'A direction for contrastive analysis: the comparative study of developmental sequences.' *TESOL Quarterly* 16, 2 169–183.

Zobl, H. (1983) 'Markedness and the projection problem.' *Language Learning* 33,3 293–313.

Appendix 1

Adjective form

あのみせのほうが、_____ かもしれません。

(a) やすく (b) やすい (c) やすくて (d) やすいです

Conjunction

テレビを 見(み)ている _____ ねむくなりました。

(a) あいだ (b) あとで (c) うちに (d) まえに

Particle

かいもの ___ 行(い)きます。

(a) まで (b) に (c) から (d) を

Structural noun

パーティーに この赤(あか)いきものを 着(き)る _____にしました。

(a) つもり (b) こと (c) の (d) ところ

Verb form

来年(らいねん) 日本(にほん)に _____ たいです。

(a) 行(い)き (b) 行(い)く (c) 行(い)け (d) 行(い)っ