Expert Feedback? Assessing The Role Of Test-Taker Reactions To A Proficiency Test For Teachers Of Japanese[1]

Noriko Iwashita and Catherine Elder
The NLLIA Language Testing Research Centre
University of Melbourne

Abstract

This paper draws on data gathered in the process of trialling of a language proficiency test for teachers of Japanese. The test comprises reading, writing, listening and oral components and is designed to test language skills which are of particular relevance to the second language classroom. The trial population consisted of approximately 300 candidates some of whom were attending specialist teacher education courses, some who had experience as classroom teachers of Japanese and others who were in the second or third year of a post Year 12 Japanese undergraduate major. Their feedback on each component of the test was gathered immediately after test administration via a questionnaire which elicited reactions. The reactions of the different types of candidate (teachers/teachers in training and undergraduates) to the various test tasks were compared and considered in relation to their actual performance on the test. Aspects of the content, construct and face validity of the test are considered in the light of the analysis and the value of test-taker feedback in the test revision process is discussed.

Introduction

Recent writings in the field of language testing have stressed the importance of drawing on multiple sources of evidence for test validity (e.g. Messick 1988). Test developers are expected to consult with experts in drawing up test specifications, and, subsequently, to consider however effectively these specifications have been operationalised by trialling the actual test items or tasks on a good-sized sample of candidates drawn from the relevant target population. Candidate performance is then assessed and subjected to statistical analysis to determine whether test items are at a suitable level of difficulty, whether they discriminate effectively between candidates of different ability and whether ability

[1] An earlier version of this paper was presented at the 9th Biennial Japanese Studies Association of Australia Conference, University of Queensland in July 1995.

Melbourne Papers in Language Testing 1997 Volume 6.1 pp. 53-67. The Language Testing Research Centre, The University of Melbourne.

estimates obtained from the analysis are stable (i.e., whether we can be confident that a second or subsequent rating of candidate ability would produce the same result).

In recent years increasing attention has been paid to candidate perceptions of the test-taking experience since this will reveal the extent to which a given test is acceptable to its users (see for example Alderson 1988; Brown 1993; Cohen 1984; Hill 1994; Kenyon & Stansfield 1991; Shohamy 1982; Zeidener 1990 and Zeidener & Bènsoussan 1988). As Hill (1994) states:

> *It is especially important in situations where test-takers have a very large investment in the outcome that the test is perceived as both fair and valid for its purposes. (p.7)*

While the feedback of test-takers is potentially valuable, thus far in the literature little mention has been made of the problems which may arise in evaluating and acting on this kind of feedback. A number of studies indicate that affective reactions to a given test may vary according to particular characteristics of the test-takers, such as gender, ethnicity and language ability. Zeidener and Bensoussan (1988) found that males responded more positively to oral tests than did females. Hill (1994) found differences between Asian and European subjects in their attitude to a tape-mediated test of English for vocational purposes. Zeidner (1990), Brown (1993), Bradshaw (1990) and Shohamy (1982) all found some relationship between the ability level of candidates and their attitude to test tasks, with weaker candidates tending to respond less favourably to them than more proficient ones. However, in the face of conflicting opinions among test-takers or of overlap between one group of test-takers and another, the task of deciding whose judgment is more valid or more relevant given the purposes of the test, may not always be straightforward. Different background variables (e.g. gender, occupation, proficiency) may interact with one another or contribute in different measure to candidates' attitudes to a given test. The status of any information gathered from test-takers must therefore be carefully considered.

Brown (1993) discovered that the kind of course taken by test candidates had a bearing on reactions to a performance-based test of Japanese for those applying for employment in service occupations within the tourism and hospitality industry. Candidates with relevant LSP training felt more positively about the test than those

who had attended generalist Japanese courses. Brown suggests that these LSP candidates are, by virtue of their specialist training, sufficiently informed about the requirements of the profession to qualify as 'expert judges' and that their views can thus be taken as evidence of the test's content validity. In professions such as foreign language teaching, however, current public rhetoric about standards required for effective professional performance may be far in advance of the skills which foreign language teachers actually possess, and a test which attempts to match these policy requirements may be met with some opposition by those who have had limited opportunities for training in what are now regarded as minimum proficiency requirements. Is it appropriate in such a situation to consider reactions from practising teachers as 'expert feedback' if the domain of expertise which the test purports to measure does not correspond to the kind of expertise owned by the target population?

Furthermore, in evaluating different kinds of evidence for test validity, the test developer may be required to make choices between the results yielded by the statistical analysis of test performance on the one hand, and of the reactions of test-takers on the other. In other words, there may be a conflict between two types of 'expertise', that of the test-taker and that of the test developer, a conflict which needs to be resolved in some principled fashion if test outcomes are to be taken seriously by all those who have a stake in them.

Context for the study

The above issues are further elaborated in this paper, which draws on data from a recently developed Japanese language proficiency test for prospective language teachers (funded by the Department of Employment Education and Training in the federal government of Australia).

The Proficiency Test for Language Teachers: Japanese is a specific-purpose test developed to measure proficiency in the five skills of reading, text-editing, writing, listening and speaking in relation to the particular requirements of classroom foreign language teachers. Test tasks are designed to reflect, in so far as is possible in the test situation, the kinds of skills required of foreign language teachers in preparing lessons and in communicating both with L2 learners and

with native speakers in performance contexts which are of relevance to the teacher role. Some of these tests are classroom specific (i.e., they relate directly to the requirements of the classroom situation) and others are more general (since it is assumed that teachers, in preparing for their classes, will need to draw on a broad knowledge/proficiency base). The test specifications for the pilot version (subsequently revised on the basis of various kinds of feedback) were developed in consultation with Japanese language teacher experts and are summarized in Appendix 1.

Research questions

Two research questions are addressed in the present study.

1. Do reactions to the various test components differ according to the background of test-takers?

2. What is the relationship (if any) between test-taker reactions to the various test components and the results yielded from an empirical analysis of test performance?

Findings will be used to inform discussion about the role of test-taker feedback in the test validation process.

Methodology

Questionnaire response

A questionnaire was administered to a subset of test-takers just after the test administration to elicit their reactions to the various test components. Candidates were asked to respond to a number of statements about such issues as the clarity of instructions, the suitability of the text, and the difficulty of test items by choosing from options on a 4 point Likert scale (strongly disagree / disagree /strongly agree / agree). Space was also provided for open-ended comments about each of the test components. A sample of the questionnaire is provided in Appendix 2 to this paper. A total of 384 questionnaires were completed. Answers were coded on a scale of 1 to 4 with 4 representing the most favourable response and 1 the least favourable. Scores for each test component were summed to produce

an overall attitude measure for each skill (ie speaking, reading, listening, writing, text-editing).[2] As well as eliciting reactions to the test, the questionnaire requested the following biodata from candidates: gender; language background (native speaker of Japanese, Chinese, Korean, Australian or other); occupational status (teacher, undergraduate student, student in training to be a Japanese language teacher or other); amount of study (in years); time elapsed since completion of study (in years); time spent in Japan (in years).

Analysis of the data

Candidates' background variables and test scores on the various test components were then cross-referenced to their attitude score. Two background variables[3] selected for the analyses were occupation (whether undergraduate student or teacher) and proficiency. The occupation variable was of course relevant since the test was designed for teachers or would-be teachers and prior experience in the professional area could be regarded as a good credential for assessing the suitability of test tasks (Brown 1993). The proficiency variable was deemed important because of findings of other studies (Zeidner 1990; Brown 1993; Bradshaw 1990 and Shohamy 1982) which suggest that affective reactions to a test may be strongly related to candidates' feelings of adequacy or inadequacy in performing test tasks. Other variables such as time spent in Japan, the number of years of tertiary study and time elapsed since completion of studies were found to be correlated with proficiency and were hence excluded from the analysis.

The data were divided into two groups (teacher and student groups) according to information provided about candidates' occupation[4].

---

[2]For each section of the test, different number of questions were asked in the questionnaire. (8 questions for the speaking task, 4 questions for the writing task, 6 questions each for listening and reading tasks and 5 questions for the text-editing task). Therefore, the full score for questionnaire response in each section of the test varies.

[3]Initially the gender variable was also selected because most teachers and language students tend to be female and it was important to consider the appropriateness of the test for this population, but it was excluded from the final analyses due to very little influence on the attitude score compared with other variables (occupation and proficiency).

[4]There were several test-takers with occupational backgrounds other than teacher, teacher in training and undergraduate student, and these test-takers were excluded from the analysis.

The teacher group was made up of both Diploma of Education students who were well into their training year, and of practising teachers. The student group was in fact made up entirely of undergraduate students. T-tests were performed to compare the attitude of each group of test-takers (teachers and students) to the various components of the test. Further analyses were then carried out in order to determine the relationship between proficiency and attitude on the listening, reading and text-editing tasks.

For research question two, the attitude scores of each group were ranked from the most favourable to the least favourable and findings were compared to those produced by a statistical analysis of test properties.

Results and discussion

Do reactions to the various test components differ according to the background of test-takers?

The T-tests revealed significant differences in attitude score between teacher and student test-takers on the listening, reading and text-editing tasks. Table 1 shows the attitude scores in all sections of two groups. The text-editing tasks were favoured more by teachers than by students whereas the reverse was the case on the listening and reading tasks (Table 1).

Table 1. Comparison of attitude scores for two groups

Section		Teacher			Students			t
	Max. score	N	Mean	S.D.	N	Mean	S.D.	
Speak.	32	29	25.7	3.76	15	24.1	2.74	1.68
Write.	16	26	11.0	1.94	115	11.3	1.81	0.82
Listen.	24	20	14.0	1.17	98	15.8	1.14	2.01*
Read.	24	36	16.0	2.06	128	17.6	1.85	4.16**
Edit.	20	18	15.9	2.36	41	14.6	1.51	-2.17*

*p<.05, **p<.0001

For the listening, reading and text editing sections a further analysis (ANCOVA) was conducted to determine the relative contribution of background variables (i.e. Occupation and Proficiency) to test takers' attitude score (Table 2). (An interaction term was also included in the analysis.)

Table 2. Background variables/attitude (ANCOVA)

	Occupation (Teacher/student)	Proficiency	Occupation & Proficiency
Listening	5.70**	9.17**	6.44**
Reading	7.73*	0.97*	0.03
Text-editing	0.83	1.05	1.22

*p<.05 **p<.001

For the text-editing task the ANCOVA analysis produced a non-significant result which means that there was no specific relationship between test-takers' background and their attitudes to this component of the test.

On the listening task, Proficiency was found to be the best predictor of attitude, which means that, as found in previous studies, high scorers tended to have the most positive attitude to the test. Occupation also emerged as a significant albeit less powerful factor contributing to the attitude score. There was also a significant interaction between Occupation and Proficiency probably caused by the fact that most of the high scorers on this component were non-teachers.

On the reading task, although Occupation was the factor that best predicted attitude, Proficiency also had a significant part to play. There was however no significant interaction between these two variables.

At this point it is worth attempting an interpretation of the findings reported above. The fact that student test takers reacted more favourably to the listening task than did teachers appears to be the result of two factors - occupation and proficiency. The listening text, as explained earlier, was a video-taped lecture on a

general topic 'Rice Imports in Japan', which bore no obvious relationship to the kind of listening a teacher might do in the context of his/her classroom practice. Our argument for including it was that teachers might be reasonably expected, as part of their professional training, to attend lectures and to view video material in order to be in a position to provide up-to-date information about the culture of the target language (which in turn would be conveyed to students). While undergraduate students proved to be quite accepting of this test task/format, teachers were quite indignant. Comments such as 'this passage assumes specialised knowledge that school-age students do not have' and 'this is way beyond the level which my learners would be expected to cope with' suggest that the teachers' sense of content relevance is confined to what they expect their learners to be able to cope with. It is interesting to note moreover that teachers involved in the trials performed at a signficantly lower level than undergraduate students. Teachers' relatively poor performance on this task may partly explain their negative reaction to this component of the test.

As far as reading is concerned, one possible explanation of the fact that teachers react more negatively than undergraduate students to the task is that, again, they may see it as being unnecessarily difficult for their students. In other words their view of content relevance relates directly to the materials used in the classroom and they may not see it as important for a teacher to be able to do more than a student is required to do. This is borne out by comments from teachers about the number of kanji in the text which is beyond what would be required at matriculation level. Whether we accept this view or not is essentially a matter of how we perceive the teacher's role: should a teacher be only expected to keep abreast of minimum current curriculum demands or should s/he have a larger bank of knowledge and ability to draw upon? Teachers' answers to this question are likely, of course, to vary depending in part on their own level of proficiency or the opportunities which they are given to maintain it. There was in fact a significant difference between teachers and students' level of performance on this task, but the difference in score between the two groups was not as large as in the listening. This may have been due to the fact that dictionaries were allowed for the reading component of the test. It was reported by a number of teachers that their level of kanji knowledge had diminished drastically since completing their undergraduate

studies and that without recourse to a bilingual dictionary they would not have been able to attempt some of the test items.

What is the relationship (if any) between test-taker reactions to the various test components and the results yielded from an empirical analysis of test performance?

In order to rank attitude scores of each group given in Table 1, mean attitude scores are produced as a percentage of the maximum possible score first.[5] (Table 3). Then the percentage of the maximum score (in columns three and five in Table 3) were ranked from the highest to the lowest. Table 4 shows the rank order of test components according to test-taker attitudes of the two groups.

Table 3. Comparison of test-taker attitudes

	Teacher		Students	
	Mean	Mean as % of max score	Mean	Mean as % of max score
Speaking	25.7	75.3	24.1	80.5
Writing	11.0	69.0	11.3	71.1
Listening	14.0	55.4	15.8	66.4
Reading	16.0	67.0	17.6	73.6
Text-editing	15.9	80.2	14.6	73.3
Mean (%)	n/a	63.4	n/a	73.0

[5]As mentioned earlier, the maximum score for each section of the test varies because of the different number of questions asked. A percentage mean score (i.e. the mean as a percentage of the maximum scores) was therefore calculated to allow for comparison of attitudes across the various test components.

Table 4. Rank order of test components according to test-taker attitudes

Test-taker attitudes	
Teacher	Students
TEXT EDITING	SPEAKING
↓	↓
SPEAKING	READING
↓	↓
WRITING	TEXT EDITING
↓	↓
READING	WRITING
↓	↓
LISTENING	LISTENING

The test component which was the least popular amongst both teachers and students was the listening test (teachers 55.4% and students 66.4%). Both groups' reaction to this component was much less positive than that elicited in response to the other test components (See columns 3 and 5 of Table 3). Both groups' relative lack of enthusiasm for this task may be linked to high levels of anxiety (and this is borne out by some of the comments made by candidates). The fact that a listening text, unlike reading, must be processed under time constraints (i.e. the candidate cannot go back to seek clarification of the parts of the text which s/he has not understood) may be stressful for candidates who may moreover be disturbed by such factors as tone of voice and style of presentation particularly when the input is presented in video mode. The higher the rate of anxiety is, the more likely it is that test-takers will respond negatively to a test task. However rate of anxiety amongst test-takers is not in itself a measure of validity of test iems. On the contrary, analysis of the trial test data revealed that the listening test had a smaller number of misfitting items and hence a greater discriminatory power than all other test components.

Conversely the item analysis revealed that the text-editing task was far less robust as a measure of proficiency (i.e., there were greater numbers of misfitting items) and yet there was a much more positive reaction to this task from both teachers and undergraduate students.

The results of the present study reveal a potential conflict between test takers' reactions and statistical evidence as to the test's internal validity (as revealed by the numbers of misfitting items). Ideally, if we are serious about providing multiple sources of evidence for test validity, we should opt only for those tasks where the opinions of test-takers and the statistical evidence converge, but given the limited resources available for experimenting with different possibilities and the range of factors which may influence the affective reactions of test-takers, this may in some cases be an unreachable goal.

Conclusion

The present study compared two groups of test-takers' reaction to a teacher proficiency test, and considered the legitimacy of teachers' status as expert judge.

The investigation took the form of an analysis of the influence of background variables on test-taker responses to a questionnaire eliciting candidates' feedback on the various tasks immediately after they had taken the test.

The findings present rather a confused picture as far as the value of test-taker feeback is concerned. While the occupational experience of teachers is potentially valuable in assessing the suitability of various types of tasks as measures of professional competence, in this case teacher feedback was revealed to be somewhat suspect because of the powerful influence of proficiency on questionnaire responses. In sum it appeared that teachers' judgements were influenced by the fact that many of them felt ill-equipped to meet the proficiency requirements of the test. In other words their questionnaire responses were not necessarily indicative of a considered and dispassionate appraisal of the suitability of the content and format of test tasks for their intended purpose.

It may nevertheless be important, if only for political reasons, to take test-taker feedback seriously, given Spolsky's (1995) observation that 'any public language test will need to satisfy several different sets of criteria, criteria drawn from different areas and representing the interests of several different groups' (p. 160). We ignore criteria of face validity or public acceptability at our peril since they may determine whether a test is ultimately adopted and/or whether its results are accepted as plausible measures of the target domain. It is nevertheless difficult to know what weighting this 'public acceptability' or 'face validity' criterion should be given compared to other information about the quality of the test, such as that generated by an item analysis, and testing scholars offer very little guidance on this issue. The divergence between test taker and other kinds of test feedback which we have documented in this paper may be merely a symptom of a larger problem, namely: that tests alone cannot be expected to drive public educational policy (e.g. in this case to raise standards of teacher competence) unless they are accompanied by other more far-reaching strategies (e.g. in this case a systematic professional development program which ensures that teachers are able to meet the requisite standards). Without such strategies, tests, if they are to be acceptable to their users, can do no more than reflect the status quo.

References

Alderson, J.C. (1988) New procedures for validating proficiency tests of ESP? Theory and Practice. Language Testing 5: 220-232.

Bachman, L.F. (1990) Fundamental considerations in language testing. Oxford: Oxford University Press.

Bradshaw, J. (1990) Test-takers' reactions to a placement test. Language Testing 7: 13-30.

Brown, A. (1993) The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. Language Testing 10: 277-303.

Cohen, A.D. (1984) On taking language tests: what the students report. Language Testing 1: 70-81.

Hill, K. (1994). The contribution of multi-informant feedback to the development and validation of an oral proficiency test in two formats. Unpublished MA thesis, University of Melbourne.

Kenyon, D.M. and Stansfield, C. (1991) A method for improving tasks on performance assessments through field testing. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago IL, April 1991.

Messick, S. (1988) Validity. In Linn, R.J. (Ed.) Educational Measurement Third Edition. N.Y.: American Council on Education / Macmillan.

McNamara, T.F. (1990) Assessing the second language proficiency of health professionals. Unpublished Ph.D thesis, University of Melbourne.

Shohamy, E (1982) Affective considerations in language testing. The Modern Language Journal 66: 13-17.

Spolsky, B. (1995) Measured words: the development of objective language testing. Oxford: Oxford University Press.

Zeidner, M. (1990) College students' reactions towards key facets of classroom testing. Assessment and Evaluation in Higher Education 15 (2): 151-169.

Zeidner, M. & Bensoussan, M. (1988) College students' attitudes towards written versus oral tests of EFL. Language Testing 5: 100-114.

Appendix 1

Outline of the Proficiency Test for Language Teachers: Japanese (Trial version)

Section 1: Text editing and Reading *time allowed 45 mins.*

Task 1.

Candidates identify and correct a range of errors (of phonology, syntax, semantics and character formation) in a quasi-authentic text produced by a school-age L2 learner.

Task 2.

Candidates read one or more texts (amounting to approximately kanji) on a topic or topics relevant to themes proposed in national curriculum materials and/or teaching texts developed for Australian learners of Japanese. Comprehension of these texts is assessed by means of a range of item types (multiple choice and open ended).

Section 2: Writing *time allowed 40 mins.*

Candidates, in the role of LOTE teacher, write two kanji pieces on a topic related to the teacher role.

Task 1 is directed to an L2 learner audience.

Task 2 is directed to a native speaker of Japanese (professional colleague).

Prompt materials are provided and the writing sample is assessed for a) content and organization b) discourse style and c) control of linguistic elements.

Section 3: Listening *time allowed 30 minutes*

Task 1.

Candidates view a videotaped lecture containing informational content of relevance to the teaching of Japanese culture/current affairs.

Task 2

Candidates view a videotaped discussion involving both native and non-native speakers of Japanese on a topic related to teaching.

Comprehension of the two texts is assessed via multiple-choice and open-ended questions.

Section 4: Speaking *time allowed 30 minutes per pair*

Candidates perform a set of tasks which are relevant to the teacher role including reading aloud, giving instructions, paraphrasing information and negotiating arrangements. Picture prompt materials are provided for each task. Tasks are performed in pairs with candidates alternating betweeen the role of teacher and learner. Candidates are assessed according to the following criteria: a) fluency b) accuracy c) appropriateness d) intelligibility e) task fulfilment.

Appendix 2

A sample of the questionnaire form used in the study

Reading

Now that you have completed the reading section please tell us about your reactions.

What did you think of the reading test? Indicate your response to the following statements by ticking the appropriate circles.

	Strongly Disagree	Disagree	Agree	Strongly Agree
It reflects my ability to read in Japanese.	O	O	O	O
The texts were suitable for my level.	O	O	O	O
The questions were difficult.	O	O	O	O
The instructions were clear.	O	O	O	O
Tasks were relevant to my needs.	O	O	O	O
There was enough time allowed.	O	O	O	O

Please make further comments on the reading tasks.

.....

.....

.....

.....