PAC3
at
JALT
2001
Conference
Proceedings

MENU
Text Version
Help & FAQ

International
Conference
Centre
Kitakyushu
JAPAN
November
22-25, 2001

Which One Speaks Louder in Language Testing: Actions or Words?

GholamReza HajiPourNezhad
The University of Social Welfare Sciences, Iran

As there seem to be more unresolved than resolved issues in language testing, this study aimed to reveal the main testing-related assumptions held by the majority of testing professionals in Iran in an attempt to discover the unanimously agreed-upon testing priorities, needs, and means. This was implemented by carrying out a survey investigation into the judgments made by 1) testing experts 2) decision-making authorities of high-stakes testing, and 3) professionals in two EFL organizations to find out discrepancies as well as commonalities among the three informant groups' judgments. The patterns of judgments of the three groups contributed to a qualitative analysis of the underlying assumptions of English language testing in Iran, and were utilized as the basis for the development of a new standard proficiency test called T-Test (Tehran Test). Results show the merits of pooled judgments in language testing compared with testing decisions based on individual test constructor's decisions and preferences.

Throughout the process of test development and analysis, we are entangled in a network of judgments that we cannot avoid. Judgments in language testing are not limited to scoring; they extend into a variety of aspects such as decisions upon the difficulty level of a test or what the test actually tests. As regards the increasing attention toward judgmental devices, Hamp-Lyons & Prochnow (1994) state that "an entirely different direction in education research at the moment, however, is toward the use of judgments, attitude surveys, experiential data such as verbal protocols, and a generally humanistic orientation" (p. 1). Therefore, the purpose of the present study is not to insist on avoiding the use of judgments, but on utilizing group judgments instead of individual ones in an attempt to make judgments less faulty and deficient.

As a matter of fact, the present study is based on the premise that testing is full of gray areas, which are neither black nor white. In other words, in L2 testing, as is natural in all fields of study in the humanities, there are very few definitely answered areas. (It should be parenthetically said that, for instance, recent debates on the LTest-L mailing list on the appropriate number of options in multiple-choice items and also on different aspects of language test validity have probably more than anything else demonstrated that the number of resolved questions in testing is infinitesimal in comparison to the unresolved ones, and also these debates have shown that every one of the experts can have a word in the formation of a relatively final picture). These gray areas need judgment rather than definite answers at the moment (it is too soon to provide definite answers). Then, it is hypothesized that for these gray areas, we, language test constructors need to think together and judge together to make relatively sound decisions. In the context of the Iranian society, it seemed that the three groups involved with L2 test issues were as follows: Testing experts, Official authorities, and EFL organizations (organizers).

To put it in a different way, as the proverb "actions speak louder than words" speaks for itself, it has been customary to believe that actions are more productive than words. "Words" in the context of this study refer to judgments made by, not certain experts in isolation, but by the majority of significant experts concerned with testing. Actions, however, refer to activities of test planning, construction, validation, and administration. The content of this study is in favor of the argument that 1) words speak louder if they underlie actions; that is, test affairs will be more successful if they are based on the collective judgments of a large number of testing specialists. And 2) actions do not speak louder when they are not based on words, that is, test affairs will not really thrive unless they are based on pooled judgments. Data is reported to indicate that, as most political systems of the world are doing their best efforts to reach a democratic society, in testing, we need to resort to group judgments and pooled decision making rather than a descending hierarchical process whereby all decisions regarding test construction and validation and even very minute details of tests are made by, if not one, at best, only a few experts. The result of this state of affairs is believed to be not as satisfactory because as the saying goes "two heads are better than one". Of course, the message of the present study is " the more brains, the better" rather than "two heads instead of one".

Background

Testing professionals throughout the world are increasingly faced with new challenges in terms of test construction and validation, as well as test use interpretation and the employment of new procedures to deal with the current dilemmas in the field. However, for about twenty years, the status of testing experts in Iran has been unique. Due to certain political reasons, The TOEFL test has not been administered since the beginning years following the revolution in Iran. This has had a variety of effects, ranging from economic (the expenses of frequent trips to neighboring countries to take the TOEFL Test), through psychosocial (candidates' inconvenience with having to go on trips to take the test), to teaching and testing ones. Ever since, testing experts have made six proficiency tests to serve the purposes fulfilled by TOEFL. However, there have been certain problems with these tests. Some of the most important problems of these tests as reported in HajiPourNezhad (2000) have been as follows:

1. *Validity-related issues:*

- Definitional validity: these tests have not been totally congruent with the definitions their developers have presented for language proficiency.
- Consensual validity: these tests have not gained high general agreement from a large number of testing experts or other concerned professionals.
- Construct validity: in terms of convergent validity, that is, measuring the same trait through different test methods, in terms of factorial validity, whereby a number of the test samples are factor analyzed to see their go-togetherness in terms of common variance, and in terms of trait validity through, for instance, factor analysis.
- Congruent validity: (or predictive criterion-related validity) in the form of correlations with the TOEFL test.
- Response validity: in terms of both fit validity and general response validity.
- Wash-back validity: in terms of teaching, testing, and the satisfaction of those involved with test affairs (for instance, EFL organizers).
- Differential validity: in terms of making differentiations among subtests on certain criteria.

- Consequential or Ethical validity: Due to the fact that the candidates would count on the results of the tests as valid predictors of performance on the TOEFL Test. However, it usually took them several trips to other countries to pass the TOEFL test successfully. And also, in terms of ability interpretations with false positive classification errors and false negative classification errors.

2. *Reliability- related issues*:

- Parallel forms, i.e., parallel forms reliability. The parallel forms developed for assessment purposes did not turn out to have high correlation coefficients.
- Test-retest reliability.
- But not in terms of KR21, because as we know this kind of reliability can be easily manipulated by changing the number of items and the degree of the homogeneity of the items.

3. *Interpretation:*

- NRT frame of reference: Most of these tests were originally piloted with norm groups diametrically different from the target norms of the population to take the test.
- CRT frame of reference: Inappropriate definitions of Mastery/Non-mastery categorizations, and in terms of domain specifications.

4. *Practicality*:

- Since most of these tests tried to do the trick in any way possible, most of them were too long, which made the tests both impractical and response invalid.

5.*Test bias*:

- These tests were biased in terms of passage selection of certain backgrounds, and in terms of cultural bias.

With this general picture of the state of affairs, the purpose of the study was to investigate beliefs held by those professionals who were concerned with test construction, validation, and administration in Iran, and to find out whether this investigation could underlie the development of a standard proficiency test.

Method

Subjects

One hundred and ten people from the three mentioned fields of testing experts, decision-making authorities, and professionals in EFL organizations were randomly selected and were invited to assist. Eighty expressed their interest in the study. These eighty acted as the informants of the study.

Materials and Procedures

As the materials for the study, I initially asked the 80 respondents to list three of their main beliefs or assumptions about testing English in Iran. It was briefly clarified that these assumptions could include anything ranging from very gross to very trivial aspects of proficiency testing. Leaving commonalities, this initial survey provided about 135 assumptions on proficiency testing. Through informal contacts with other testing experts outside this group, I also gathered 45 other assumptions so that the number of assumptions amounted to 180 altogether. These assumptions were of three general types:

1. Those which were, in the light of ample research evidence, clearly erroneous such as:

- Construct validation is sufficient ground to claim the test is valid.
- We are far behind the use of an IRT modeling of our proficiency tests.
- Statistical procedures are just making things more complex.
- Integrative testing is, in practice, an abstraction from reality.
- It is neither practical nor rewarding to try to define a universe of measures in the G-theory framework.

2. Those, which were, in the light of ample research evidence, clearly correct such as:

- It is possible to integrate principles of CRT and NRT in the development of proficiency tests.
- Multiple reading passages should be included in reading subsections of test batteries to prevent Test Bias.
- Incremental validity is an indispensable part of test validation.
- The consideration of SEM is critical to making sound decisions on a true score.
- Definitional validity is an indispensable part of the validation process on proficiency tests

3. Those which were ambiguous in that there could be shades of truth in them. These called for judgments such as:

- Multiple measurement should replace single measurement in spite of its considerable expenses.
- A proficiency test tailored to the needs of the community has got to be constructed.
- Pooled (group) Judgment in test development and validation is practical and more productive although it is more burdensome.
- The inclusion of standard cloze subsections in proficiency tests will be productive.
- Language proficiency is more of pragmatic

ascription than theoretical construct.

4. Those which were used to cross-check the responses given to a previous assertion. For instance, the following assertion:

> Making differentiation among examinees and determining the mastery of specified elements are equally important

...was a restatement of a previous assertion:

> It is possible to integrate principles of CRT and NRT in the development of proficiency tests.

On the basis of the 180 assertions on proficiency testing, I developed a questionnaire to measure judgments on the mentioned assertions. The respondents were to rate the instructions on a 5-point scale with 1 meaning complete disagreement and 5, complete agreement. The completion of the questionnaires took about six months. The respondents would rate each of the assertions at the presence of the author as the interviewer. Each respondent would first rate the assertion, and would later explain her/his views regarding the given issue. The interviewer would simultaneously be completing a checklist on the knowledgeability of the respondent on the certain issue. This also had a five-point scale with "1" meaning unfamiliarity of the respondent with the given term, and "5" meaning the ability to explain and justify different approaches to the issue.

Results

As a general rule of thumb, those judgments which were accompanied by a "1" on the "knowledgeability scale" were eliminated for analysis. For instance, if an informant rated an assertion such as the first correct assertion above (combination of CRT and NRT principles) as "5" meaning "complete agreement", while s/he did not know what criterion- and norm-referenced testing frameworks were- which would give them the least knowledgeability score "1"- her/his judgment would be eliminated. This, in practice, meant that any judgment not based on knowledge would not be acceptable.

As a second rule, it was decided to take the arithmetic mean of judgment ratings on each assertion as the accepted score for that assertion. For instance, the mean on the fourth assertion under the third category above (the use of cloze in proficiency tests) was 3.8. That is, the population of the informants gave this figure as the average rating on this item.

As a third rule, it was decided to consider any mean above 3.0 as acceptable ground for inclusion into our "experts' judgments repertoire" for the development of the new proficiency test.

The mentioned rules enabled me to quantify the judgment ratings of the 80 informants on 180 assertions in the form of judgment means in a “repertoire of 180 beliefs about proficiency testing in Iran”.

As an example of the outcomes of the study, I can name the finding that, according to judgments, a valuable proficiency test, which is to be based on national needs and preferences, will have as its main feature a very high predictive validity with the TOEFL test so that candidates will not have to go to other countries back and forth to pass the TOEFL test after several tries with failing the test as the result. This is directly related to Economic effects of the prospective proficiency test.

A second preferred feature of the test was also the need to have strong wash-back on testing in Iran. One implication of this was that the test had to incorporate some of other testing purposes sought in Iran. One most important case was the effects of the new test on the English section of The National University Entrance Examination, which is another case of high-stakes testing in Iran.

During the process of the development of the new test, all the eighty experts were involved in all the steps taken for test planning, content specification, test item construction, validation, analysis, and interpretation.

The new test was named “T-test” standing for “Tehran test”. Several outstanding scholars from around the world were consulted at different stages of the job.

Further results showed much more acceptable results for the T-test compared with its predecessors (Alavi & HajiPourNezhad, 2001:11). This might be an implication of the judgment survey of the study, namely, pooled judgment is more productive than individual judgment.

References

Alavi, S.M. and HajiPourNezhad, G., (2001). Validation of the T-test. *Tehran ELT Journal*, 33(4). 24-35.

Alderson, J. C., (1993). Judgments in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing*. Alexandria, VA, USA: TESOL.

HajiPourNezhad, G., (2000). An in-depth analysis of Iranian proficiency tests. *Tehran ELT Journal*,32(2).11-22.

Hamp-Lyons, L. and Prochnow, S.M., (1994). Examining Expert Judgments of Task Difficulty on Essay Tests. *Journal of Second Language Writing*. 3(1).