Development and Validation of the Hausa Speaking Test with the *ACTFL Proficiency Guidelines*

Charles W. Stansfield and Dorry Mann Kenyon
Center for Applied Linguistics, Washington DC

This article reports on the Hausa Speaking Test (HaST), a simulated oral proficiency interview (SOPI). Following careful development, trials and multiple revision of test items, a validation study was conducted. The study addressed the validity of the HaST through an examination of the ratability on the ACTFL scale of the elicited speech sample and an investigation of the nature of probes on the HaST through the speaking tasks referred to on the ACTFL Proficiency Guidelines. The results have implications for both the validity of the HaST and that of the ACTFL Proficiency Guidelines.

INTRODUCTION

The introduction to the 1986 *ACTFL Proficiency Guidelines* states that the *Guidelines* "represent a hierarchy of global characterizations of integrated performance in speaking, listening, reading and writing" (American Council on the Teaching of Foreign Languages, 1986). This article demonstrates the use of the *Guidelines'* hierarchy for speaking in developing the Hausa Speaking Test (HaST), a tape-mediated oral proficiency test developed by the Center for Applied Linguistics (CAL) in 1989. The article also reports on some preliminary research conducted to validate both the HaST as a surrogate for the Oral Proficiency Interview (commonly known as the OPI, a face-to-face assessment procedure of speaking ability in a foreign language) and to validate the *ACTFL Guidelines* as representing a consistent hierarchy of speaking proficiency.

Issues in Applied Linguistics
ISSN 1050-4273

Vol. 4 No. 1 1993 5-31

The Hausa Speaking Test (HaST) was developed by CAL as one of a series of tape-mediated speaking tests to meet the need for oral proficiency testing in the less commonly taught languages. Although Hausa is not the national language of any single country, it is an important West-African language, spoken as the mother-tongue of some 25 million speakers in northern Nigeria and southern Niger, and as a second or third language for half again that number (Newman, 1987). In 1986, after Swahili and Yoruba, it was the most widely studied African language in the United States (Brod, 1988).

Although there is much discussion in the literature about the validity of the *ACTFL Guidelines* and the OPI (Bachman & Savignon, 1986; Barnwell, 1989; Hagen, 1990; Kramsch, 1986; Lantolf & Frawley, 1985, 1988, 1992; Shohamy, 1990), the OPI and the *Guidelines* have exerted tremendous influence on the field of foreign language teaching in the United States. A bibliography published in 1989 included over 400 articles in the literature focusing on the *Guidelines* and their application to language assessment and teaching (Stansfield & Thompson, 1989). It is safe to say that the OPI has become the most influential model for assessing oral proficiency.

In less commonly taught languages such as Hausa, however, trained OPI testers are rare or nonexistent. Because a tape-mediated approach to testing oral proficiency eliminates the need for an on-site interviewer, it seemed to language testers at CAL to offer an efficient and feasible approach to oral proficiency testing in low-volume languages, providing the positive washback to be derived from oral proficiency testing and serving as an impetus for competency-based learning on the part of students of less commonly taught languages. Experience in training raters in the scoring of CAL's tape-mediated testing format has also shown that it is easier to train individuals to score such a test then to train individuals to both administer and rate an OPI.

With support provided by the U.S. Department of Education, CAL has developed tape-mediated tests in Chinese (Clark & Li, 1986), Portuguese (Stansfield & Kenyon, 1988; Stansfield, Kenyon, Paiva, Doyle, Ulsh & Cowles, 1990) , Hebrew (Shohamy, Gordon, Kenyon & Stansfield, 1989), Indonesian and Hausa (Stansfield & Kenyon, 1989). All of these tests follow a similar format, which Stansfield (1989) has called the simulated oral proficiency interview (SOPI). Through careful construction

following the hierarchy outlined in the *Guidelines*, the SOPI seeks to elicit from the examinee a speech sample ratable on the ACTFL scale. Instead of eliciting speech via a face-to-face interaction (as in the OPI), the SOPI uses recorded and printed stimuli. Yet the goal of the SOPI is the same as that of the OPI: to assess an individual's proficiency in a foreign language on the *ACTFL Guidelines* (often referred to as the ACTFL scale).[1]

The ACTFL scale is an adaptation of a scale that has been used in government agencies since 1956. The scale is commonly known as the Federal Interagency Language Roundtable (FILR) scale (Liskin-Gasparro, 1987). The FILR scale denotes eleven levels as follows: 0, 0+, 1, 1+, 2, 2+, 3, 3+, 4, 4+, and 5. The ACTFL adaptation encompasses only the FILR levels from 0 to 3. It has four main levels: Novice, Intermediate, Advanced and Superior, and several sublevels, as presented in Table 1 with the FILR scale equivalences. For the HaST, CAL has added one level above Superior (High-Superior), which is used to identify examinees approaching the level of educated native speaker (3+ to 5 on the FILR scale). Appendix A contains a copy of the scale used to score the HaST.

Table 1. The ACTFL and FILR Scales

ACTFL	FILR
Novice-Low	0
Novice-Mid	0
Novice-High	0+
Intermediate-Low	1
Intermediate-Mid	1
Intermediate-High	1+
Advanced	2
Advanced-High	2+
Superior	3
High - Superior*	3+ - 5

* Used by CAL to denote performance above Superior

Format of the HaST

The OPI follows a format tailored to the level of each examinee. Following a warm-up, the interviewer seeks to check his or her assumption about the proficiency level of the examinee by asking the examinee a series of questions at the examinee's apparent level of proficiency. To further confirm this assumption, the interviewer also presents probes, which are questions at a level slightly above the examinee's apparent level.

As a SOPI, the HaST also uses a well-defined though fixed format intended to check and probe the examinee's proficiency. The structure of the SOPI also presents the examinee with speaking tasks at different levels of speaking proficiency, as they are represented by the *ACTFL Guidelines*. Since all of the tasks on the SOPI are ones that can be effectively handled only by responding with more than isolated words and learned phrases, the SOPI is not designed for Novice-level learners. The format of the HaST can be divided into six parts:

1. Warm-up
2. Giving Directions
3. Picture Narration
4. Topical Discourse
5. Situational Discourse
6. Wind-down

Each of these parts presents examinees with speaking tasks at specific levels of the ACTFL hierarchy. The intended level of each speaking task in each part of the HaST is presented in Appendix B, which outlines the structure of the test. These parts are described in detail in the following sections.

1. *Warm-up*. After the general directions are read to the examinee from the master tape, the test begins with simple personal background questions posed on the tape in a simulated initial encounter between a native speaker of Hausa and the examinee. During a brief pause, the examinee records a short answer to each question posed on the tape. Items in this part of the test require examinees to respond to tasks ranging from formulaic speech (Novice-level responses) to giving personal information (Intermediate-level responses). This section is analogous to the warm-up section of the OPI. Its purpose is to ease the examinee

into the testing situation and allow him or her to become accustomed to the testing format.

Following the warm-up are the four core parts of the HaST. These are designed to elicit language similar to that elicited during the *level check* and *probe* phases of the OPI. Items are designed to test the examinee's ability to handle speaking tasks at the Intermediate, Advanced and Superior levels as defined by the *ACTFL Guidelines*. The directions to all the items in these four parts are read on the master tape and printed in the examinee's test booklet. All directions are given in English to ensure that the tasks required of the examinee are clear and to ensure that the examinee is given the opportunity to give his or her best performance regardless of listening proficiency (which would ideally be tested in a different format). Parts two and three also use pictures which are printed in the test booklet. Following the reading of the directions, the examinee is given between 15 and 30 seconds (depending on the difficulty of the task) to silently prepare a response. After a tone signal, the examinee has between 45 seconds and two minutes to record his or her response.

2. *Giving Directions*: The examinee is asked to give directions on the basis of a simple map. This Intermediate-level task is contextualized in that the interlocutor to whom the examinee will speak is identified and the reason for the request for directions is explicitly delineated in the prompt.

3. *Picture Narration*: The HaST contains three such items. Successful completion of the task presented in these items requires the examinee to narrate in present and past time, and to give a series of commands to help a Hausa speaker through an unfamiliar procedure. All of these are tasks at the ACTFL Advanced level.

Parts four and five of the HaST require the examinee to tailor his or her discourse strategies to selected topics and real-life situations. These last two parts assess the examinee's ability to handle the speaking tasks and content that characterize the Advanced and Superior levels of the *ACTFL Guidelines*.

4. *Topical Discourse:* The examinee is instructed to talk about selected topics involving different discourse strategies. The selection of topics is intended to probe the examinee's ability to provide information on a variety of subjects involving different vocabulary domains. Speaking tasks include explaining a process (Advanced), supporting an opinion (Superior) and talking about a

hypothetical situation (Advanced/Superior). There are five such topics, each printed in the test booklet.

"Talk about the advantages and disadvantages of using public transportation" is an example, taken from the *Hausa Speaking Test Examinee Handbook*, of a typical topical discourse item. The item's speaking task is to state advantages and disadvantages, which is intended to elicited Advanced-level performance.

5. *Situational Discourse*: The examinee reads a printed description of a real-life situation in which the background circumstances, the interlocutor or audience, and the communicative task are identified. The examinee is then instructed to carry out the specified task. The tasks range from making simple requests (Intermediate level) to giving a brief informal talk to a gathered group (Superior level). Situations differ from topics in that the situations emphasize the ability to tailor one's speech to the audience and the circumstances.

The following is an example of a situational discourse item for the Intermediate-level speaking task of making a simple request, taken from the *Hausa Speaking Test Examinee Handbook.* "You are with a Hausa friend at a market in rural Hausaland. Ask your friend to recommend a special gift for you to take home for your family in America." An example to illustrate a Superior-level speaking task (giving a brief speech) is, "At the end of a year-long stay with a family in Hausaland, you present them with a small gift and express your gratitude for all they have done for you during the past year."

The final part of the test is analogous to the wind-down of the OPI. The questions are given in Hausa, and the examinee responds directly after hearing the question, as in part one of the test.

6. *Wind-down*: This part contains three simple questions in Hausa spoken by the same individual as in the first part of the test. It is designed to put the examinee at ease and to facilitate the ending of the examination in as natural a manner as possible and is not used in the rating of the test. The wind-down permits the examinee to comment on the test and the testing experience.

Distinctive Aspects of the HaST

Through experience in developing SOPIs in the less commonly taught languages, test developers at CAL have learned

that each language presents its specific challenges. Although based on the prototypical Chinese Speaking Test (Clark & Li, 1986) and the Portuguese Speaking Test (Stansfield et al., 1990), the HaST was modified to accommodate concerns of both the local test development committee and the external review committee[2], and on the basis of data collected through extensive pilot testing. Although every attempt is made to avoid culturally loaded situations on a SOPI, as an outcome of the iterative process of review and pilot testing the test developers found that the HaST items needed to be fairly highly contextualized to Hausaland culture in order to elicit ratable speech samples. In particular, the setting for prompts needed to be "de-urbanized" as much as possible. It was found that Hausa was a language particularly tied to its cultural setting, and examinees had problems relating Hausa language use to non-Hausaland settings. This was particularly true of examinees who had learned Hausa in Africa.

In addition, the difficulty level of the test was also lowered by including more Intermediate level questions and fewer Advanced and Superior level questions than on earlier SOPIs. This was in response to the practical realization that few, if any, of the North American students of Hausa who had not spent extensive time in Hausaland would ever reach the Advanced, much less the Superior, level in Hausa. By lowering the difficulty level of the test, more examinees would feel comfortable taking the test.

Finally, in order to accommodate morphological inflections by gender required in Hausa, two versions of the master tape were made. In one version, male examinees are addressed, while the other addresses female examinees. Standard Hausa, as spoken in Kano, Nigeria, was used.

Two parallel forms of the HaST were developed (Form A and Form B). The forms are parallel in respect to the speaking tasks each item addresses (e.g., give directions or support an opinion), though the specific content of each task is different. In every case, the content of each item was designed to be accessible to adult English-speaking learners of Hausa at all proficiency levels above ACTFL Novice, so that an examinee would be able to at least say something, even if completion of the specific speaking task required proficiency in Hausa above what the examinee currently possessed.

Research on the HaST and the *ACTFL Guidelines*

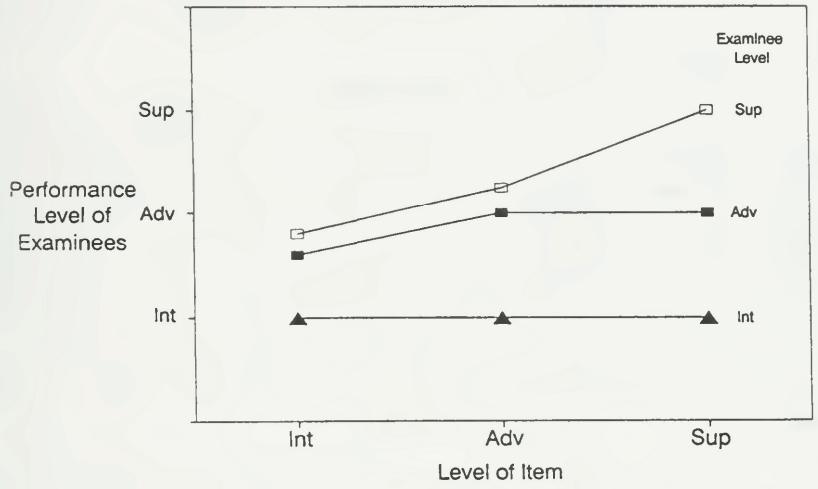
The goal of initial test development research is to validate a new test; i.e., to determine its appropriateness for the testing purposes for which it is intended. In the case of a SOPI, it is necessary to determine if the test is an appropriate surrogate for the OPI in the less commonly taught languages. To establish the comparability of the SOPIs developed by CAL in Chinese, Hebrew, Indonesian, and Portuguese, both the SOPI and an OPI were administered to a sample of language learners and scores obtained on each were compared; these were concurrent validity studies. The average correlation across languages, tests, forms, and raters between the SOPI and the OPI was .92 (Stansfield, 1989). Because there were no ACTFL-trained oral proficiency interviewers in Hausa, similar research could not be conducted for the HaST.

In lieu of a direct comparison with an OPI, the validation study of the HaST sought to answer the question of whether the test was doing what it was designed to do; i.e., to probe the various levels of proficiency as defined by the *ACTFL Guidelines* through the use of tasks specifically developed to elicit speech at the various levels of the ACTFL scale. Unlike previous studies, which only examined the final rating awarded to an examinee, this study explores the functioning of the individual items on the test.

It was hypothesized that if the HaST were functioning like an OPI in its ability to probe speaking proficiency, then examinees at the Intermediate Level would be rated as Intermediates not only on Intermediate level items, but on all items; that examinees at the Advanced Level would generally score above Intermediates at all levels, but particularly show their higher proficiency on Advanced level items; and that examinees at the Superior level would consistently show themselves to be better than both Intermediate and Advanced Level students on all items, but particularly demonstrate their Superior level ability on those items that required them to handle Superior level speaking tasks.

These hypotheses are expressed in diagram form in Figure 1. This figure shows the hypothesized mean ratings for each group of examinees (Intermediate, Advanced and Superior) on each group of items (by intended level). Three relevant specific hypotheses were delineated as follows:

Figure 1
Hypothesized Performance of
Different Level Examinees on
Different Level Items



1. Intermediate level examinees would never score above the Intermediate level on any item.
2. Advanced level examinees would perform better than Intermediate level examinees on Intermediate items, at the Advanced level on Advanced items, but not above the Advanced level on any item.
3. Superior level examinees would perform better than Intermediate and Advanced level examinees on Intermediate and Advanced level tasks, but not be able to fully demonstrate their Superior level except on Superior level items.

PROCEDURES

Thirteen subjects were administered both Form A and Form B of the HaST. Each subject was administered the appropriate version (male or female). The design controlled for order of administration, with half of the subjects receiving Form A first and Form B second, and the other half in reverse order.

Most of the subjects were administered the HaST at the Center for Applied Linguistics using two tape recorders. Some of the subjects were administered the test at the language lab at their respective universities or by their Hausa instructors. Two of the subjects administered the taped tests to themselves at home using two cassette tape recorders.

All of the subjects were adults; six were male and seven were female. Due to the scarcity of suitable subjects (i.e., Hausa students at the ACTFL Intermediate level or above[3]), the subjects could not be randomly selected. The sample included several university level students of Hausa, several subjects who had learned Hausa through experience in the Peace Corps and did not have formal academic training in the language, and several individuals who had learned Hausa in other situations and who have occasion to use Hausa in their work. Because the number of Hausa-as-a-second-language speakers is so small nationwide, it was unavoidable that a few of the subjects tested were personally known to the raters.

Due to the small number of Hausa linguists familiar with the ACTFL scale, it was necessary that the two raters used in the study be selected from the members of the local and external test development committees. Both had received some ACTFL training and one was working on ACTFL certification as an ESL oral proficiency tester at the time. However, neither was ACTFL-certified and neither had formerly rated Hausa speech samples on the ACTFL scale.

The raters scored each examinee's performance on the HaST using a form that asked them to do the following:

1. to rate each examinee's performance on each individual item, basing the judgement solely on the performance on that item;
2. to award a score for the usefulness of the speech sample elicited by each item in rating that examinee's proficiency;
3. to award a holistic proficiency rating to the examinee's entire test performance.

The 26 examinee tapes (13 examinees, 2 forms) were scored by the two raters independently in sets of five or six. Each examinee received a single holistic rating on the basis of his or her performance across the various types of items on the test. After each set of tapes was scored, however, the two raters, without

changing their original rating, compared their holistic ratings and discussed disagreements. This self-training was built into the design because the raters had not previously applied the ACTFL scale to the rating of speech samples in Hausa.

RESULTS

In the empirical analysis of the ratings, scores on these two SOPI test forms were converted to a numerical scale combining both the ACTFL and FILR scales, with weights assigned to reflect the FILR numerical scale, as follows:

ACTFL/ FILR Level	Coded as:
Novice-Low/0	0.2
Novice-Mid/0	0.5
Novice-High/0+	0.8
Intermediate-Low/1	1.0
Intermediate-Mid/1	1.5
Intermediate-High/1+	1.8
Advanced/2	2.0
Advanced-High/2+	2.8
Superior/3	3.0
High-Superior/3+to 5	3.8

This system of score coding is intended to assign an appropriate numerical value to the proficiency level descriptions. For example, proficiency at an Advanced-High/2+ level is characterized by many of the same features as at the Superior/3 level, though the examinee cannot sustain the performance. Thus, the numerical interpretation should fall closer to 3.0 than mid-way between 2.0 and 3.0, as might be expected.

Analyses of HaST Reliability

The several tables below provide descriptive statistics, interrater reliabilities, and parallel-form reliability data obtained in the study.

Table 2 shows the mean rating, standard deviation, and other descriptive statistics for each of the two raters on each of the SOPI test forms.

Table 2. Descriptive Statistics for Scoring Levels Assigned

Test Form	Rater	Minimum Score	Maximum Score	Mean	Standard Deviation
Form A	(n=13)				
	Rater 1	0.2	2.8	1.54	0.75
	Rater 2	0.5	2.8	1.53	0.65
Form B	(n-13)				
	Rater 1	0.2	3.0	1.61	0.66
	Rater 2	0.5	2.8	1.42	0.65

The mean ratings for each rater of the Form A examinee response tapes were very similar. However, on Form B, Rater 1 appears to have awarded slightly higher scores than Rater 2, as shown by her slightly higher mean ratings.

Table 3 shows the frequency of the 52 scores awarded to this sample across raters and forms (i.e., 2 raters x 2 forms x 13 examinees).

These figures illustrate the difficulty of locating suitable examinees to take the HaST. 20% of the ratings assigned were at the Novice levels, indicating that these examinees were below the suggested Intermediate Low minimum level for which the test was intended. Only 22% of the ratings were above the Intermediate level and only one Superior rating was awarded. However, there was quite a range in performances.

Table 3. Frequency Distribution of All Ratings Across 13 Subjects, 2 Raters and 2 Forms

	Rating	Frequency	Percent
0.5	Novice Mid	3	6
0.8	Novice High	5	10
1.0	Intermediate Low	7	13
1.5	Intermediate Mid	10	19
1.8	Intermediate High	14	27
2.0	Advanced	6	12
2.8	Advanced High	4	8
3.0	Superior	1	2
	Totals	52	101*

*due to rounding

The degree of agreement between the absolute ratings awarded was relatively high for these inexperienced raters. There was total agreement in 46% of the 13 paired ratings on Form A. In only one case was the disagreement greater than one step on the ACTFL scale; here, one examinee was awarded a Novice-Low by Rater 1 and a Novice-High by Rater 2. For Form A, in 92% of the cases there was either complete agreement or a difference of one step on the scale. On Form B, there was total agreement in 31% of the 13 paired ratings. Again, only one of the ratings was more than one step away from the rating awarded by the other rater.

Correlations between the ratings assigned by Rater 1 and those assigned by Rater 2 for the two SOPI test forms are shown in Table 4 below. The first is the Pearson product-moment correlation coefficient, and estimates the interrater reliability. The second, presented in parentheses, is the Spearman rank order correlation coefficient which is not affected by disagreements in score, only by disagreements in rank ordering. Since the two raters were inexperienced in rating Hausa speech samples, the rank order coefficients may give a better indication of how more experienced raters might perform. It may also be noted that the product-moment

correlation with a small sample may be heavily influenced by extreme values. The rank order correlations are less susceptible to extreme values. These correlations, both on the absolute scale and in terms of rank order, are quite high across both test forms.

Table 4. Interrater Reliabilities

Test Form	Correlation
A (n=13)	.88 (.95)
B (n=13)	.93 (.95)

Table 5 presents correlations for the same subject taking two different test forms, with the same rater scoring both forms. These can be considered parallel form reliabilities. Rank order correlations are given in parentheses.

Table 5. Parallel-Form Reliabilities (Same Rater)

	Rater 1	Rater 2
Forms A and B (n=13)	.82 (.95)	.80 (.92)

The numbers above indicate that either the rating scale may have been inconsistently applied by the raters or that some examinees did indeed perform differently on the two test forms. This can occur when an examinee attempts to do his or her best on the one form due, perhaps, to interest in the initial testing experience, but fails to make such effort when taking the second form.[4] In terms of relative ranking, the two tests placed the examinees in basically the same order for both raters. The fact that the rank-order parallel-form reliability was quite high for the two different raters supports the claim that the sample of speech elicited

by different forms consistently differentiates among performances, even if raters are inconsistent in which absolute score they assign each performance.

Table 6 shows parallel-form reliabilities for subjects taking two different test forms, with each form scored by a different rater. (Again, rank order correlation coefficients are given in parentheses.)

Table 6. Parallel Form Reliabilities (Different Forms and Raters)

Rater/Form Combination and Correlation	
Rater 1/Form A - Rater 2/Form B (n=13)	.91 (.95)
Rater 1/Form B - Rater 2/Form A (n=13)	.76 (.91)

This type of parallel-form reliability involves error that can be attributed to natural variation in examinee speech, error that can be attributed to differences in test form, and error that can be attributed to differences in raters. Thus, it may be viewed as a lower-bound estimate of the reliability of a HaST score. Although the reliabilities were not always impressively high regarding absolute ratings (i.e., the two raters at times differed both within and among themselves in severity), even under these severe conditions (different forms and different raters), the ability of the raters to place the examinees in very nearly the same rank order on the basis of the examinees' performance on the HaST is impressive.

Analyses of HaST Validity

As mentioned earlier, the HaST raters were asked to rate each item (i.e., the warm-up, the four picture items, the five topic items, and the five situation items) in terms of its usefulness in making the holistic rating for that examinee. The rating scale for item usefulness ranged from 1 (lowest) to 5 (highest), with the midpoint (3) defined as "adequate." There were 15 such ratings per examinee on each form. The mean rating given by the two raters across the 13 subjects for all the items on Form A was 3.27 and on

Form B it was 3.15. These mean ratings of usefulness indicate that in the opinion of the raters, the individual items were adequate in eliciting a ratable speech sample from the group of examinees in the validation study.

For the purposes of testing the hypotheses stated above (concerning the ability of the HaST to probe proficiency at the different ACTFL levels), it would have been best to have been able to divide the sample into groups of Intermediate, Advanced, and Superior level subjects. However, as noted above, the sample that took the Hausa test turned out to be unexpectedly low in average proficiency. Thus, for data analysis purposes the thirteen examinees were divided into three groups on the basis of similar proficiency ratings. Group 1 contained five individuals who, across both HaST forms and across both raters, had received proficiency ratings ranging between Novice-Low (0.2) and Intermediate-Mid (1.5). The mean score of group 1 members across raters and across forms was .87. This is nearest to a score of Novice High on the ACTFL scale. Group 2 contained five individuals who had received proficiency ratings at Intermediate Mid (1.5) or Intermediate High (1.8). The mean score of this group across raters and forms was 1.70, nearest to a score of Intermediate-High on the ACTFL scale. Finally, group 3 contained three individuals whose proficiency ratings ranged from Intermediate High (1.8) to Superior (3.0). The mean score of this group across raters and forms was 2.42, about midway between Advanced and Advanced-High on the ACTFL scale.

To examine the hypothesis depicted in Figure 1, it is necessary to examine the mean ratings by intended level of the item. For this analysis, all ratings were combined; i.e., scores for each individual examinee from both raters were averaged for each item, and then the average for all items at that intended level was computed. Thus, each subject had three pieces of data: his or her average on the eight Intermediate, sixteen Advanced and four Superior level items that comprised the two forms of the test. Then, the means for each of the three proficiency groups were calculated. These mean ratings are given in Table 7.

Table 7. Mean Group Performances on Items

Proficiency Group	*Intended Item Level*		
	Intermediate (8 items)	Advanced (16 items)	Superior (4 items)
1 (n=5)	0.914	0.948	0.833
2 (n=5)	1.553	1.626	1.622
3 (n=3)	1.835	2.227	2.542

These mean ratings are also presented in the diagram in Figure 2.

Figure 2
Actual Mean Performance of
Different Level Examinees on
Different Level Items

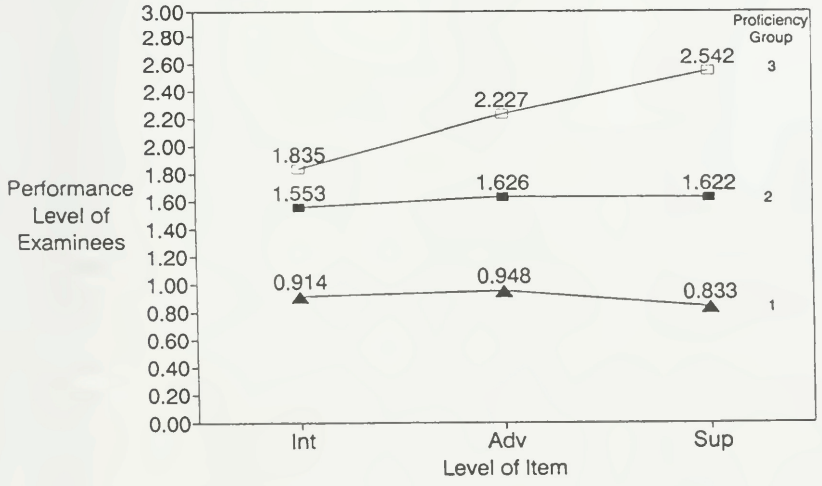


Figure 2 illustrates that the actual results appear to be similar to the hypothesized outcome. To test for the statistical significance of the results, a blocked repeated measures analysis of variance as a multivariate analysis of variance (MANOVA) was conducted using SAS. First, the test for an effect of the interaction of proficiency grouping and item level on examinee performance was significant (Wilks' lambda $F_{(4,18)}=9.86$, $p=0.0002$). Next, the test for any main effect of the different item levels on examinees was also significant (Wilks' lambda $F_{(2,9)}=24.55$, $p=.0002$). Thus, performance differed by intended item level across examinees. This indicates that the three groups were NOT equally affected by the different item levels. (If this statistic had not been significant, then the three lines in Figure 2 would be parallel.) Finally, the test for any between subject effect (i.e., difference due to proficiency grouping) was significant ($F_{(2,10)}=34.30$, $p=0.0001$). This indicates that the three groups performed differently from each other across the three item levels.

A pairwise comparison of means (Bonferroni T tests) across the three item levels reveals that the only difference in performance at any item level that was NOT significant was between groups 2 and 3 at the Intermediate item level. The mean for group 2 here was 1.553, while the mean for group 3 was 1.835. This result further supports the hypothesis that higher level examinees need items at higher levels on the proficiency scale in order for their different ability levels to be separated from each other. Had the mean proficiency of each the three groups been equal to Intermediate, Advanced, and Superior, no difference in performance on the Intermediate level items across the three groups would be expected. However, in this analysis, the mean overall performance of group 1 members was below the Intermediate level. (Recall that the mean overall rating of group 1 members was .87, which is about Novice High.) Thus, it is not surprising that group 1 scored significantly lower than groups 2 and 3 on the Intermediate level items. Likewise, had the proficiency of the three groups been equal to Intermediate, Advanced and Superior, no significant difference on the Advanced level items between groups 2 and 3 would have been expected. However, the average overall performance of group 2 members was below the Advanced level. (Recall that the mean overall rating of group 2 members was 1.70, or about Intermediate High.) Thus, it is not surprising that group 2 scored significantly lower than group 3 on the Advanced level items.

The hypotheses presented in Figure 1 predicted that for Intermediate level examinees there would be no difference in their scores across the three item levels, but that there would be an item level effect for the Advanced and Superior level examinees. To examine this, three separate single group repeated measures ANOVAs were conducted. The results indicate that there was no item level effect for proficiency group 1 (Wilks' Lambda $F_{(2,3)}$=4.63, p=.1211), nor for group 2 (Wilks' Lambda $F_{(2,3)}$=.94, p=.4825). However, there was a significant item level effect for proficiency group 3 (Wilks' Lambda $F_{(2,1)}$=554.21, p=.0300). This indicates that the lines in Figure 2 connecting the means for proficiency groups 1 and 2 should be considered statistically parallel. Considering that the mean of group 1 was in the Novice High range and the mean of group 2 was between Intermediate Mid and Intermediate High, these results do not disconfirm the original hypotheses. They do support the hypothesis that examinees at the Intermediate level remain at that level despite the ACTFL level of the item.

In summary, these findings are generally consistent with the hypotheses stated. Lower level examinees (group 1) perform at the same level across the various item levels. Given any of the HaST tasks, they would be rated lower than higher level examinees. However, higher proficiency students (group 3) would have received a lower holistic rating had they only been given Intermediate level items. Although they consistently performed better than the other groups at any item level, they needed the Superior level items to show the full extent of their ability. In short, these results indicate that the HaST items function as probes of each level as intended, and that the variety of item difficulties on the test are working to probe the examinee's overall ability to speak Hausa.[5]

In addition to providing some evidence for the validity of the HaST in a situation where concurrent validity with a face-to-face interview can not be obtained, the results of this study provide some initial support for the validity of the *ACTFL Proficiency Guidelines* as a hierarchy of performance descriptions of the speaking ability of learners of a foreign language. The items were written according to the content and speaking tasks described in the *Guidelines*. The fact that examinees were able to handle the content and speaking tasks in a way that matched the items' difficulty levels with the examinees' proficiency levels suggests that the hierarchy of tasks included in the descriptions is valid, at least for this limited sample. If the

Guidelines were without validity, then the higher level group in this study, whose mean holistic rating (2.42) was between the Advanced and Advanced High level, would not have performed any better on Superior level tasks than they did on Advanced or Intermediate level tasks. However, this was not the case.

In addition, the middle group (with a mean holistic rating at Intermediate-High) performed equally well and did not exceed the Advanced level on both Advanced and Superior items. The lowest group in this study (with a mean holistic rating of Novice-High) did not perform above the Intermediate level on Intermediate, Advanced or Superior level tasks. These results, including the fact that the low level students may have been disadvantaged by the Superior level items (Figure 2), indicate the necessity of including items on the SOPI at all levels of the ability range being tested.

DISCUSSION

Although this study was presented merely as an effort to examine the validity of the speaking tasks included on the HaST, it has been noted that the results have implications for the validity of the *ACTFL Proficiency Guidelines* as a representation of a hierarchy of skills, operationalized in the OPI. Although these results may be satisfying to those who have used the OPI and the accompanying *Guidelines* for a number of years, further studies of the *Guidelines*, making use of the methodology employed here, could be carried out. Such studies could employ certified raters and a larger sample of examinees. With a larger sample it would be possible to construct groups whose mean and range of proficiency more closely approximate the proficiency level that each group is intended to represent. With a greater spread in proficiency levels between groups, it is likely that the differences between groups in future pairwise comparisons would also be greater, if the *Guidelines* are valid.

The research methodology employed here may have broad application to the test development process. If the validity of the *Guidelines* is established through future research, then future efforts to develop SOPI tests based on the *Guidelines* can evaluate each item by comparing the performance of examinees at different proficiency levels . In such a case, if an item is intended to reflect

the Advanced level of the *Guidelines*, and Advanced and Superior level examinees do not score at the Advanced level, then the item might be revised or discarded, since it did not perform as it was designed to perform. Such a methodology could serve as a kind of item analysis that could be used for pretesting purposes. This methodology may be seen as a simple form of one parameter item response theory, with misfitting items being discarded.

The method may have further applications. If the *ACTFL Guidelines* are valid, then the method may be used to examine misfitting examinees. These would be examinees whose performance on individual items did not fit the model (for example, an Advanced level examinee who scores at the Intermediate level on a particular item). A comprehensive analysis of such individuals could provide a better understanding of any limitations to the validity of the *Guidelines*, as well as an understanding of the types of individuals for which the *Guidelines* are not valid. Thus, the methodology employed here may serve as the basis for a number of research studies on the *Guidelines*.

A further extension of this methodology beyond the sphere of the *ACTFL Guidelines* would be to present the speech samples, as individual segments, to native speakers of Hausa who are unfamiliar with the *Guidelines*. These Hausa speakers would be asked to rate each performance on a scale appropriate to the research. For example, they may be asked to make a rating from 1 to 7 for the degree to which the speaker demonstrates ability to communicate in Hausa. Would Superior level speakers, as defined by the *ACTFL Guidelines*, then outperform themselves on Superior level items (as opposed to Intermediate level items)? Would Intermediate level speakers be rated consistently across items at all three levels of proficiency? A positive outcome of such a study would support the contention that items can be at different proficiency levels, and that the hierarchy reflected in the *ACTFL Guidelines*, contrary to some criticisms in the literature, does reflect external judgments on proficiency made by native speakers of a language.

NOTES

[1] Needless to say, there are some differences in some of the aspects of a speech sample elicited in a tape-mediated mode (the SOPI) and a direct mode (the OPI). Shohamy, Shmueli & Gordon (1991) have analyzed the speech samples of 10 examinees who were administered both types of tests in Hebrew. Although certain interactive discourse features were present in the OPI and absent in the SOPI, in areas such as syntax, morphology, lexicon, and amount of speech, there were no differences in the frequencies of occurrence between the samples collected by the two different elicitation procedures. In addition, raters scored the examinees for proficiency similarly, whether listening to an OPI or a SOPI.

[2] The local test development committee was spearheaded at CAL by Charles W. Stansfield, Project Director. CAL testing staff included Dorry Mann Kenyon, Project Coordinator and Daniel Kennedy, Test Development Specialist. Local Hausa language experts were Beverly Mack (George Mason University) and Steven Lucas (Voice of America, United States Information Agency). The external reviewers of the HaST were William R. Leben (Stanford University), Roxanna Ma Newman (Indiana University) and Russell G. Schuh (University of California, Los Angeles).

[3] Although Brod (1988) listed national Hausa enrollments as totaling 60 students, the vast majority of these students were enrolled in beginning level courses. In these courses, the teachers, depending on whether they are from the department of linguistics or anthropology, either teach the language analytically or focus on both culture and language. As a result, we were advised that most students of Hausa have oral language proficiency at the ACTFL Novice level.

[4] In fact this appears to have happened. Upon analysis of individual scores, one examinee who was awarded an Advanced-High by both raters on the first form taken received an Advanced and an Intermediate High rating on the second form.

[5] Information on examinee attitudes toward the test was obtained as part of the validation study by means of a short questionnaire given to the subjects directly after completing the HaST. All subjects completed the questionnaire, providing a 100% participation rate.

The first two questions sought to determine if the subjects felt their Hausa speaking ability had been adequately and fairly probed by the HaST. Eleven of the 13 subjects (85%) responded that the descriptions, narratives, situations, and other types of questions in the test were adequate to probe their maximum level of speaking ability in Hausa. 85% also indicated that there were not any picture/descriptions, narratives, situations, or other questions they felt were in any way 'unfair'. A small majority (54%) reported feeling unduly nervous during the test. This is not surprising, since the test was above the actual proficiency level of many of the subjects and the semi-direct mode of testing was unfamiliar to the students. Twelve of the 13 subjects (92%) felt the length of the timed pauses for examinee responses was about right and 100% felt that the directions were clear. Finally, a large majority (77%) of the subjects felt that the two tests (Forms A and B) were equally difficult. This is important as the tests were designed to be alternate forms.

In summary, examinee reaction to the HaST was very positive, especially when one considers that the test tasks were inappropriately difficult for many

subjects. From the examinee's point of view the HaST probes Hausa speaking ability fairly and adequately, and it is technically sound.

REFERENCES

American Council on the Teaching of Foreign Languages. (1986). *ACTFL Proficiency Guidelines*. Hastings-on-Hudson, NY: Author.

Bachman, L. & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal, 70,* 380-390.

Barnwell, D. (1989). Proficiency and the native speaker. *ADFL Bulletin, 20*, 42-46.

Brod, R. I. (1988). Foreign language enrollments in U. S. institutions of higher education–Fall 1986. *ADFL Bulletin, 19*(2), 39-44.

Clark, J. L. D. & Li, Y. C. (1986). *Development, Validation, and Dissemination of a Proficiency-Based Test of Speaking Ability in Chinese and an Associated Assessment Model for Other Less Commonly Taught Languages.* Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 278 264)

Hagen, L. K. (1990). Logic, linguistics and proficiency testing. *ADFL Bulletin, 21*, 46-51.

Kramsch, C. (1986). From language proficiency to interactional competence. *Modern Language Journal, 70*, 366-372.

Lantolf, J. P. & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *Modern Language Journal, 64*, 337-45.

Lantolf, J. P. & Frawley, W. (1988). Proficiency: Understanding the construct. *Studies in Second Language Acquisition, 10*, 181-95.

Lantolf, J. P. & Frawley, W. (1992). Rejecting the OPI–Again: A response to Hagen. *ADFL Bulletin, 23*, 34-37.

Liskin-Gasparro, J. (1987). *Testing and Teaching for Oral Proficiency.* (Appendix B). Boston, MA: Heinle and Heinle.

Newman, P. (1987). Hausa and the Chadic languages. In B. Comrie (Ed.), *The World's Major Languages* (pp. 705-723). New York, NY: Oxford University Press.

Shohamy, E. (1990). Language testing priorities: A different perspective. *Foreign Language Annals, 23*, 385-394.

Shohamy, E., Gordon, C., Kenyon, D. & Stansfield, C. W. (1989). Development and validation of the *Hebrew Speaking Test. Bulletin of Hebrew Higher Education, 4*, 4-9.

Shohamy, E., Shmueli, D. & Gordon, C. (1991). *The Validity of Concurrent Validity of a Direct vs. a Semi Direct Test of Oral Proficiency.* Paper Presented at the 13th Language Testing Research Colloquium, Princeton, NJ.

Stansfield, C. W. (1989). *Simulated Oral Proficiency Interviews.* ERIC Digest. Washington, D. C.: ERIC Clearinghouse on Languages and Linguistics and Center for Applied Linguistics.

Stansfield, C. W. & Kenyon, D. M. (1988). *Development of the Portuguese Speaking Test.* (Final report to the U.S. Department of Education).

Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 296 586)

Stansfield, C. W. & Kenyon, D. M. (1989). *Development of Semi-Direct Tests of Oral Proficiency in Hausa, Hebrew, Indonesian and Portuguese.* (Final report to the U.S. Department of Education). Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 329 100)

Stansfield, C. W., Kenyon, D. M., Paiva, R., Doyle, F., Ulsh, I. & Cowles, M. A. (1990). The development and validation of the *Portuguese Speaking Test. Hispania, 72,* 641-651.

Stansfield, C. W. & Thompson, L. (1989). Topical bibliography of proficiency-related publications: 1987-88. In K. Buck, (Ed.), *The ACTFL Oral Proficiency Interview Tester Training Manual.* (Bibiliography). Yonkers, NY: ACTFL.

Charles W. Stansfield is the Director of the Division of Foreign Language Education and Testing at the Center for Applied Linguistics in Washington, DC. He has worked extensively in the testing of oral proficiency and in foreign and second language test development.

Dorry Mann Kenyon is an Associate Director for Testing at the Center for Applied Linguistics. He has broad experience in developing simulated oral proficiency interviews.

Appendix A: Scoring Scale for the HaST

NOVICE	The Novice level is characterized by the ability to communicate minimally with learned material. The HaST is designed for examinees who exceed this level. Any examinee not achieving the minimum ability to be rated at the Intermediate level will receive this rating.
INTERMEDIATE	The Intermediate level is characterized by the speaker's ability to: • create with the language by combining and recombining learned elements, though primarily in a reactive mode; • initiate, minimally sustain, and close in a simple way basic communicative tasks; and • ask and answer questions.
Intermediate-Low	Able to handle successfully a limited number of interactive, task-oriented and social situation. Misunderstanding frequently arise, but with repetition, the Intermediate-Low speaker can generally be understood by sympathetic interlocutors.
Intermediate-Mid	Able to handle successfully a variety of uncomplicated, basic and communicative tasks and social situation. Although misunderstandings still arise, the Intermediate-Mid speaker can generally be understood by sympathetic interlocutors.
Intermediate-High	Able to handle successfully most uncomplicated communicative tasks and social situations. The Intermediate-High speaker can generally be understood even by interlocutors not accustomed to dealing with speaker at this level, but repetition may still be required.

ADVANCED	The Advanced level is characterized by the speaker's ability to: • converse in a clearly participatory fashion - initiate, sustain, and bring to closure a wide variety of communicate tasks, including those that require an increased ability to convey meaning with diverse language strategies due to a complication or an unforeseen turn of events; • satisfy the requirement of school and work situations; and • narrate and describe with paragraph-length connected discourse.
Advanced-Plus	In addition to demonstrating those skills characteristic of the Advanced level, the Advanced Plus level speaker is able to handle a broad variety of everyday, school, and work situations. There is emerging evidence of ability to support opinions, explain in detail, and hypothesize. The Advanced-Plus speaker often shows remarkable fluency and ease of speech but under the demands of Superior-level, complex tasks, language may bread down or prove inadequate.
SUPERIOR	The Superior level is characterized by the speaker's ability to: • participate effectively and with ease in most formal and informal conversation on practical, social, professional, and abstract topics; and • support opinions and hypothesize using native-like discourse strategies.
High-Superior	This rating, which is not part of the ACTFL scale, is used in HaST scoring for examinees who clearly exceed the requirement for a rating of Superior. A rating of High-Superior corresponds to a rating of 3+ to 5 on the scale used by the Interagency Language Roundtable of the U.S. Government. The HaST is not designed to evaluate examinees above the ACTFL Superior level.

Appendix B: Structure of the HAUSA SPEAKING TEST (HaST)

Key: I = Intermediate
A = Advanced
S = Superior

Item	**Intended Level**	**Speaking Task**
Warm-up	I	Answer personal quesitons
Picture 1	I	Give directions
Picture	A	Narrrate in present time
Picture	A	Narrate in past time
Picture	A	Give instructions
Topic	I	Describe personal activities
Topic	A	State advantages and disadvantages
Topic	A	Give an explanation
Topic	S	Support an opinion
Topic	A	Hypothesize on a personal topic
Situation	I	Make simple requests
Situation	I	Make a complex request
Situation	A	Speak with tact
Situation	A	Make an apology
Situation	S	Give a brief informal speech
Wind down	I	