

**THE SENTENCE REPETITION TEST (SRT) REVISITED**  
**Evaluation of the SRT against the Second Language Oral Proficiency Evaluation (SLOPE)**  
**as an Assessment Tool for Community Bilingual Proficiency**

**Deborah H. Hatfield, Marie C. South, Stuart D. Showalter**

**SIL International**  
**2007**

SIL Electronic Working Papers 2007-008, August 2007  
Copyright © 2007 Deborah H. Hatfield, Marie C. South, Stuart D. Showalter, and SIL International  
All rights reserved

## **Contents**

Abstract

Background

Introduction

Relationship Between SLOPE and RPE Levels

    Summary of Results from Cameroon French SRT study (48 data pairs)

    Summary of Results from Jula SRT study (Showalter) (25 data pairs)

    Conclusion, Assumption, Consequence and Recommendations (SLOPE/RPE)

Relationship Between SLOPE and a Final Fifteen-Sentence SRT

    Summary of Results from Cameroon French SRT study (63 data pairs)

    Summary of Results from Jula SRT study (25 data pairs)

        Conclusions, Consequences and Recommendations (SLOPE/SRT)

        Conclusion

        Consequence

        Recommendations

    Final statement by panelists

    Recommendations for further research

References

## Abstract

Development of a Sentence Repetition Test (SRT) (Radloff 1991) has resulted in wide employment of this efficient technique for estimating the bilingual proficiency profile of an entire community. The accepted standard is the Oral Proficiency Interview as developed by the U.S. Foreign Service Institute. The Second Language Oral Proficiency Evaluation (SLOPE) was adapted from it by SIL (1987) to be used in preliterate societies.

Because of the obvious practical advantages associated with conducting SRT over SLOPE, two tests were conducted in Cameroon comparing SRT results to those of SLOPE where Cameroon French was the second language being tested (South 2007). A comparison of SLOPE and SRT was also conducted in Burkina Faso where Jula was the second language (Showalter, to appear)

In late 2000, seven people knowledgeable in testing and familiar with the foregoing tests came together to examine the results. The stated goal of the roundtable was

...to have assurance that,

SRT is a valid tool to measure proficiency of a community, distinguishing adequately those who are below ILR level 3 from those who are level 3 and above.

Of somewhat lesser importance, it is also good to measure the distribution of a representative sample of the population.

To have this assurance there needs to be a consensus from each of the participants and a meeting report must be one that each participant can support. Failing that, a clear report on the limits to the use of SRT must be prepared.

It was the consensus of the Roundtable that there is no single test, including SLOPE, that will give second language proficiency data that can be used as the only deciding factor in language development needs assessment. The results from an SRT, however, can provide important information on the L2 abilities of members of a speech community.

## Background

In the second half of the 1980s, the SIL South Asia sociolinguistic survey team with Carla Radloff taking the lead, developed a practical tool to measure a community's second language (L2) proficiency profile (Radloff 1991). The SRT involves choosing a representative sample of the community and asking them to listen to a series of 15 sentences which have been recorded in L2. After listening to the sentences, one at a time, they simply repeat them as best they can. The more sentences a subject can repeat accurately, the higher the L2 proficiency.

To calibrate the scores, the pretest L2 proficiency of these subjects must be evaluated. A procedure which Radloff and associates called Reported Proficiency Evaluation (RPE) was developed (Chapter 6, Radloff 1991) based on the early Foreign Service Institute (FSI) skill descriptions taken from the *Manual for Peace Corps Language Testers* (Educational Testing Service 1970).

The SRT proved to be highly successful at making quick assessments with relatively minimal skill required for administration and with no requirement for the subjects to be literate. SRTs have been developed in nearly two dozen languages. The correlation between SRT and RPE has

been extremely high, lending high confidence that the SRT measures the original FSI skill level descriptions accurately.

There is no absolute standard against which to measure bilingualism. But it is generally accepted that the oral proficiency evaluation as used by FSI and others is very good. It has been used to measure performance of individuals for many years in many circumstances by many agencies and in many languages with satisfactory results. This technique also uses skill level definitions that have evolved over time under the stewardship of the Interagency Language Roundtable (ILR 2005; Defense Language Institute 2000). The same numbers, 0 to 5, are used as with the original FSI, but with updated descriptions. SIL adapted this technique for use with pre-literate individuals with the help of Thea Bruhn, then head of FSI's Language Testing Unit. The resulting procedure was called Second Language Oral Proficiency Evaluation, or SLOPE (SIL 1987).

As Radloff says (1991:156–157), the SRT results need to be compared to results obtained through an oral proficiency evaluation. Two such comparisons have been conducted to date, one with two French SRTs (South 2007) and another with a Jula SRT (Showalter, to appear).<sup>1</sup>

Following is a report from the group of people who were called together to analyze these results:

## Introduction

The following is a report on roundtable discussions on second language (L2) proficiency evaluation that were held at the SIL International Linguistics Center in Dallas, Texas, from October 30 to November 2, 2000. The following people were present for the discussions:

Richard A. Berger, Professor (Retired), Temple University, Philadelphia, Pennsylvania  
 Ted Bergman, International Language Assessment Coordinator, SIL International, Dallas, Texas  
 Deborah H. Hatfield, Sociolinguistic Consultant, SIL International; Language Survey  
 Coordinator, SIL Togo-Benin  
 Stuart Showalter, Sociolinguistic Coordinator, SIL Burkina Faso, Burkina Faso  
 Marie C. South, Consultant Statistician, AstraZeneca Pharmaceuticals, Macclesfield, UK  
 Stephen Walter, Professor, Graduate Institute of Applied Linguistics, Dallas, Texas  
 G. Barrie Wetherill, Professor (Retired), University of Newcastle Upon Tyne, UK

The purpose of the roundtable was to bring together a panel of several people who have done research on L2 proficiency evaluation as used in SIL language surveys in multilingual language communities, in order to discuss several issues pertaining to these evaluations. The discussions focused on questions related to, in particular, the validity and reliability of the Sentence Repetition Test (SRT) as it was designed and has been applied in a number of SIL language surveys around the world.

---

<sup>1</sup> Indonesian using the Oral Proficiency Interview (OPI) that was developed for the American Congress of Teachers of Foreign Languages (ACTFL) in the calibration. The ACTFL OPI is more similar to SLOPE than the RPE. A correlation of 0.90 was obtained between SRT and the ACTFL OPI (Hanawalt and Susilwati, to appear).

The roundtable discussions examined only two SIL research studies, both of which used the SRT, the Reported Proficiency Evaluation (RPE), and the Second Language Oral Proficiency Evaluation (SLOPE). These are the only research studies where all three of these measures of L2 proficiency have been applied to the same population.

The first study was carried out in Yaoundé, Cameroon, in 1991. Two French SRTs were developed in the course of the study. Among the participants at the 2000 Dallas roundtable discussions, four took part in the Cameroon study: Ted Bergman, Deborah Hatfield, Marie South and Barrie Wetherill. Jürg Stalder and Carla Radloff, of SIL, were also among the researchers in Cameroon, as was Thea Bruhn, Head of Testing at the Foreign Service Institute (FSI) in the Washington, D.C. area.

The second research study was the Jula SRT development study, done in Burkina Faso for the Jula language, the primary language of wider communication for the southwest region of the country. John Berthelette of SIL was the team coordinator of this study, Stuart Showalter and Richard Berger, both participants in the roundtable discussions, provided consultant help and quality control.

Below is a very brief summary of the three measures:

SLOPE (SIL 1987) is a type of oral proficiency interview. The L2 proficiency level descriptions used are based on those of the Interagency Language Roundtable (ILR).

RPE uses level descriptions that are based on an earlier version of skill level descriptions, as they were used by the Peace Corps. Evaluation is done by third-party rating, based on previous communicative interactions with the person being rated (Radloff 1991).

SRT is a sentence repetition test, using 3 practice sentences and 15 test sentences. As used in SIL surveys, it has been calibrated against the RPE (Radloff 1991).

Ted Bergman set the following goals for the discussions:

Given that SRT will continue to be used as one of the primary tools to measure adequate second language proficiency for a community, we desire to have assurance that the SRT is a valid tool to measure the proficiency of a community, distinguishing adequately those who are below ILR level 3 from those who are level 3 and above.

Of somewhat lesser importance, it is also good to measure the distribution of proficiency levels of a representative sample of the population.

To have this assurance there needs to be a consensus from each of the participants and a meeting report must be one that each participant can support. Failing that, a clear report on the limits to the use of SRT must be prepared.

We focused on the following issues:

- the relationship between RPE and SLOPE levels as assigned or predicted by SRT in the French and Jula studies.
- the relationship of SRT scores and SLOPE levels in the French and Jula studies.

- Is the SRT a valid measure of global competence? If not, what can it tell us and what can't it tell us about a subject's competence?
- the dependency of SRT and RPE on each other through the calibration process.
- Which measure of global competence is better for calibrating the SRT scores?
- What are the limits to interpretation of SRT scores in the French and Jula studies?

## **Relationship Between SLOPE and RPE Levels**

### ***Summary of Results from Cameroon French SRT study (48 data pairs)***

There were 48 subjects whose assigned SLOPE and RPE ratings were compared in this study. Additional results from two RPE raters were rejected as there was evidence that their ratings were biased. A single team worked together to assign the SLOPE ratings for each subject.

Comparing assigned SLOPE and RPE levels: Ten subjects were rated at RPE level 2 or below. Five of these were assigned the identical numerical SLOPE level, as their RPE level. Five were assigned a SLOPE level which was "one half" level higher than their RPE level.

Thirty-eight subjects were rated at RPE level 2+ or higher. The SLOPE levels assigned to these subjects were either the same numerically or lower than their RPE levels.

In this study there was not a consistent difference between SLOPE and RPE assigned levels, but evidence of a curvilinear relationship with some scatter.

The largest spread was at SLOPE level 2+ (ten subjects), with corresponding RPE ratings ranging from 2 to 4+.

### ***Summary of Results from Jula SRT study (Showalter) (25 data pairs)***

Only those subjects who were assigned 2+ and above on RPE were selected to be evaluated using SLOPE.

Thirty subjects were then assigned SLOPE levels, but only 25 were later included in the analyses. The performances of five subjects during the test were not considered to be representative of their ability, so those results were disregarded.

The ten subjects who were assigned a SLOPE level 4 had the largest spread in RPE evaluations ranging from 3 to 5.

The scatter of RPE results for different SLOPE levels was similar to that observed in the Cameroon study. However, in the Cameroon study the SLOPE levels were consistently lower than the RPE levels for the subjects with higher L2 (French) proficiency. This was not seen in the Jula data.

## **Conclusion, Assumption, Consequence and Recommendations (SLOPE/RPE)**

Conclusion: From the results of these two studies, it was not possible to establish a one-to-one correspondence between assigned RPE and SLOPE language proficiency levels.

Assumption: Although there may be errors in both types of proficiency level assignment, it is assumed that the SLOPE level assignments are the more reliable here<sup>2</sup>. In each study a single, trained team decided on the SLOPE level assignments, so the level assigned was reached by consensus. However, the RPE levels were assigned by a variety of raters. Due to the sociolinguistic settings in which the studies were conducted, the SLOPE raters and the RPE raters were, for the most part, not first language speakers of French or Jula, respectively.

Consequence: Results obtained using SRTs which have been developed using only the RPE should be interpreted with caution: Lower SRT results may indicate lower levels of second language proficiency. SRT developed using RPE may not be used to assign specific SLOPE levels to individuals or a language group since the relationship between RPE and SLOPE seen in the two reviewed studies was not always consistent.

Recommendations: Consider whether there are any L2 situations where we may need to calibrate an SRT with SLOPE rather than RPE in future, e.g. in cases where we expect very high L2 proficiency (for other recommendations, see below).

## **Relationship Between SLOPE and a Final Fifteen-Sentence SRT**

### **Summary of Results from Cameroon French SRT study (63 data pairs)**

We observed the same curvature between SLOPE and the final fifteen-sentence SRTs (A and B) as we observed between SLOPE and RPE. This is not surprising since the SRT sentences were chosen based on the RPE.

It was possible to establish a threshold SRT score (30) below which it is safe to conclude that subjects will be SLOPE level 2+ or below.

It was possible to establish a threshold SRT score (35) above which, subjects are *most likely* to be level 3 or above. However, a small number of subjects with proficiency at level 2+ did score this highly on the SRT.

For candidates scoring 30–35 on the SRT it is not possible to confidently assign a SLOPE level.

It has been demonstrated, using extracted scores from the preliminary SRT, that it is possible to create French sentence sets where the extracted scores correlate more highly with SLOPE than the sentence sets A and B developed during the study. However, the ability of these new

---

<sup>2</sup> See Showalter's workpaper "Response to John Berthelette's objections to SLOPE results in the Burkina Faso Jula SRT research."

sentence sets to reliably predict SLOPE scores on additional subjects has not been assessed.

### **Summary of Results from Jula SRT study (25 data pairs)**

Of seven subjects who were rated SLOPE level 3 or below, six scored 25 or less on the Jula SRT.

Of sixteen subjects who were rated SLOPE level 4 or above, fifteen scored above 25 on the Jula SRT.

Because of the small sample size at SLOPE level 3+ (two subjects), it is not clear how SRT scores at this level would be distributed.

SRT testing was carried out in two villages where Jula was the first language learned. All the men scored above 25 on the SRT test. Six of twenty-six women (mostly younger women with little education) scored below 25 on the Jula SRT. It is important to note that L1 Jula speakers do not necessarily score the SRT equivalent of SLOPE 3 and above.

### **Conclusions, Consequences and Recommendations (SLOPE/SRT)**

#### **Conclusion**

In both the Cameroon and the Burkina studies, the SRT did seem to reliably distinguish between good proficiency and poor proficiency. That is, the SRT results could tell if a subject, or many in a community, spoke the L2 either poorly or well, with the cut-off point falling either at SLOPE level 3 (Cameroon) or SLOPE level 4 (Burkina). The SRT did seem to discriminate well at the lower levels, but did not reliably discriminate between levels of proficiency higher than the cut-off point

It was possible to establish a cut-off point below which we could have confidence that an individual would *most likely* have limited proficiency in French or Jula as an L2.

It was possible to establish a cut-off point above which it is *most likely* that subjects have a high proficiency in French or Jula as an L2.

In the Cameroon (French) study, there was in addition an intermediate category where the SLOPE level that could be predicted by the SRT score was indeterminate.

Since the cut-off points and the associated SLOPE levels were different for the two studies, it is not possible to generalize from these, i.e., it is not possible to make a general statement about the ability of an SRT to discriminate those who are below a specific SLOPE level from those who are above it.

Despite the above conclusion, both the Jula and the French SRTs have been found to be useful, practical tools to obtain an indication of levels of second language proficiency in a language

community. When used in conjunction with other data the SRTs have helped SIL administrators make language development decisions.

### **Consequence**

If a language community being surveyed has significant numbers of individuals, or particular subgroups, scoring higher than 30 on either French SRT A or B, or higher than 25 on the Jula SRT, the SRT results should be very carefully reviewed, in conjunction with the other socio-linguistic information relating to this language group, before it is decided that SIL will not be involved in L1 language development in that community. The person who reviews the results should be someone who has been involved in the development of the test and/or clearly understands the caution required in interpretation.

### **Recommendations**

The panel recommended that in the future, the SRT be developed and calibrated against SLOPE results if at all possible. There was some disagreement among the panelists on this point in that some thought this impractical in spite of its desirability. To carry out this recommendation would require new SLOPE training for key people in different geographical areas who would then serve as SLOPE consultants for future survey teams. (see “Future L2 proficiency assessment situations” below).

### ***Final statement by panelists***

There is no single test, including SLOPE, which will provide second language proficiency data that could be used as the only deciding factor in language development needs assessment. The results from a sentence repetition test, however, can provide important information on the L2 abilities of members of a speech community. This assumes that the SRT has been well-developed, well-administered, and well-interpreted.

This information, when accompanied by other pertinent data, can contribute to a decision that a translation project in a given language should or should not be initiated by SIL.

### ***Recommendations for further research***

#### **1. RELATIONSHIP BETWEEN SLOPE AND RPE LEVELS**

- Review the relationship between the SLOPE and RPE level descriptions in terms of the language functions described in each level
- Is it possible to show empirically a well-defined correspondence between the pertinent RPE and SLOPE levels?
- The panel was concerned that some people reading SRT/RPE results would assume that RPE levels were equivalent to SLOPE or FSI levels. Radloff (1991) is explicit in pointing out the

differences between these types of evaluations; nonetheless the level labels (2, 2+, 3, 3+, etc) are identical, which leads to easy confusion. Originally Radloff and Decker had used an alternative system for the RPE (A, A+, B, B+, etc.). The panel recommended that this system be used consistently with the RPE.

## 2. PROFICIENCY LEVEL & SCRIPTURE COMPREHENSION

Some researchers have assumed that in order for a person to comprehend written materials (including Scripture) of a complex linguistic nature they have to have an ILR (Interagency Language Roundtable) proficiency level of 3 or 4 in their L2. The level required has been disputed: is it 3 or 4? The first ILAC and the SIL International Board has supported the policy that people at ILR level 2+ or lower will not adequately understand that type of written material, and that if a significant number in a community have level 3 or higher additional sociolinguistic factors will need to be taken into account to establish need for SIL involvement. Thus by implication level 3 is judged to be adequate if additional factors are strong enough. An empirical comparison of individuals' comprehension of such L2 written materials in relation to their L2 proficiency levels has not been made.

Could we devise such a Scripture comprehension test and compare the results with the test subjects' proficiency levels? This could be attempted by giving a basic Scripture comprehension test to a sample of L2 speakers representing the 2 to 4+ range of SLOPE levels.

## 3. ADDITIONAL RESEARCH ON SRT / RPE / SLOPE

Conduct 1 or 2 additional SRT / RPE / SLOPE comparison studies in a geographical setting where:

- the L2 in focus is the L1 of a large speech community in that region
- there are highly educated L1 speakers of the language (to ensure that there will be L1 speakers to serve as RPE raters and assistants in the SLOPE procedure)
- there are many L2 speakers of the test language who will be at the whole range of proficiency levels. This will ensure greater accuracy and predictive power for the SRT in its final form.
- Focus on improving the reliability (limiting variability) of the SRT and improving the RPE and SLOPE evaluations. This would mean looking carefully at the calibration process for the SRT, working with a solid sample of L2 speakers at all levels, making the RPE evaluations more consistent, and providing the best SLOPE training possible for good SLOPE results. This would allow us to find out as best as possible where the variation in SRT/RPE/SLOPE scores is coming from: whether it is due to fundamental limitations in each type of evaluation, or from errors introduced in the evaluation process.
- Surveyors developing an SRT need good statistical tools for proper analysis. (This is being provided through the use of Minitab software, now being sold at a significantly reduced price through the Academic Bookstore in Dallas.)

- Surveyors need proper training in interpreting SRT results and the statistical tests applied to them. Surveyors who need help should contact B. Wetherill and M. South through Wycliffe Associates, UK. They are happy to help with statistical analyses of survey data.
- If a community of native speakers of the test language exists, surveyors are encouraged to give the SRT to a representative sample of this community as a baseline for comparison of results from L2 speakers.

#### 4. CROSS CHECKING OF RPE RATINGS

Investigate ways to cross-check the RPE ratings, for example:

- find two or more raters who know the same individuals to do the RPE ratings.

Compare results

- tape samples of speech from subjects (for rating by a second person later)
- perform a type of OPI (oral proficiency interview) on some / all individuals rated with RPE, and compare the results (question to address in the research design: Would the OPI need to be based on the RPE level descriptions?)

#### 5. RPE LEVEL ASSIGNMENTS – IMPROVING VALIDITY and RELIABILITY

A review of possible procedures for tightening up the RPE level assignments (including the suggestions in Radloff's book for tightening up RPE ratings during Pashto SRT development).

#### 6. SRT SENTENCE SELECTION

Further investigation into the selection of sentences for use in the final form SRT.

#### 7. SRT at HIGHER SLOPE LEVELS

Further investigation of performance on the SRT among subjects at higher SLOPE levels.

Because the SRT tests a limited set of language functions, can it ever be made to discriminate reliably between the higher SLOPE levels? (i.e., do the language functions that it tests top out at a certain SLOPE level?)

#### 8. SRT: WHICH LANGUAGE BEHAVIORS COUNT AS ERRORS?

- Was the large scatter in SRT scores at medium to high SLOPE levels a consequence of the scoring system used, that is, was the scatter related to the language behaviors counted as errors?

- Can we identify certain types of errors on SRTs which tend to be made by higher proficiency subjects?
- Others which tend to be made only by lower proficiency subjects?

#### 9. SRT: DEVELOPMENT / USING REGRESSION

SRT results in the process of development, calibration, and analysis should be plotted against the more reliable variable (SLOPE or RPE) for linear regression, not the other way around (RPE against SRT as Radloff did). The more reliable variable should be on the x-axis in a linear regression plot.

#### 10. FUTURE L2 PROFICIENCY ASSESSMENT SITUATIONS

Is there a two-stage procedure that could be applied such that if people perform well on the SRT, they could undergo some further L2 proficiency investigation?

#### 11. JULA SRT / PREDICTED SLOPE & RPE LEVELS

Examine applications of the Jula SRT with results for predicted SLOPE and RPE levels. Compare the results. Would there be any difference in the results, interpretation of results, or recommendations from SLOPE vs. RPE results?

#### 12. CAMEROON TUKI LANGUAGE RESEARCH USING FRENCH SRT

Examine the Cameroon French SRT results with the Tuki (Diller et al. 2003) and similar data with predicted SLOPE and RPE levels (based on calibrations / cut-offs from the Cameroon study). Compare the results. Would there be any difference in the results, interpretations of results, or recommendations from SLOPE vs. RPE results?

#### 13. PROFICIENCY LEVEL OF NATIVE SPEAKERS

It was noted that that L1 Jula speakers did not necessarily score the SRT equivalent of SLOPE 3 and above. This is a troubling result that deserves more study.

### References

- Bergman, T. G. 1990. *Survey reference manual*. Dallas: Summer Institute of Linguistics.
- Defense Language Institute. 2000. Interagency Language Roundtable language skill level descriptions. [http://www.dlielc.org/testing/round\\_table.pdf](http://www.dlielc.org/testing/round_table.pdf) (seen August 16, 2006).
- Diller, Jason, Kari Jordan-Diller and Cameron Hamm. 2003. "Sentence Repetition Testing (SRT) and Language Shift Survey of the Tuki Language." SIL Electronic Survey Reports. 2003-010. Dallas: SIL International. Online URL: <http://www.sil.org/silesr/abstract.asp?ref=2003-010>.

- Educational Testing Service. 1970. *Manual for Peace Corps Language Testers*. Princeton: ETS.  
Reproduced in Hendricks, Debby, G. Scholz, R. Spurling, M. Johnson and L. Vandenburg. 1980. "Oral Proficiency Testing". In J.W. Oller and K. Perkins (eds.), *Research in Language Testing in an Intensive English Language Program*. 77-90. Rowley, MA: Newbury House Publishers.
- Interagency Language Roundtable. 2005. *About the ILR*. [http://www.govtilr.org/ILR\\_History.htm](http://www.govtilr.org/ILR_History.htm) (seen August 16, 2006).
- Hanawalt, Charlie and Tanti Susiliwati. To appear. "Indonesian SRT Report."
- Radloff, Carla F. 1991. *Sentence repetition testing for studies in community bilingualism*. Dallas: The Summer Institute of Linguistics and The University of Texas at Arlington. Also in *LinguaLinks Library*.
- Showalter, Stuart. To appear. "Un profil du bilinguisme en dioula au sud-ouest du Burkina Faso." Proceedings of the Cinquième colloque inter-universitaire sur la co-existence des langues en Afrique de l'Ouest. Ouagadougou. Septembre 2004. *Electronic Survey Reports*. Dallas: SIL International.
- SIL. 1987. "The SIL Second language Oral Proficiency Evaluation (SLOPE)". *Notes on Linguistics*. 40A. Dallas: Summer Institute of Linguistics.
- SIL. 1990. "Language Assessment Criteria". In Bergman, T. G. (ed). 1990. *Survey Reference Manual*, 9.3. Dallas: Summer Institute of Linguistics.
- South, Marie C., compiler. (2007-001). "Cameroon Bilingualism Test Comparison Study Report". SIL Electronic Working Papers. Dallas: SIL. Originally presented by Deborah Hatfield, et al. *The Cameroon Study: A Comparison of Second Language Proficiency Testing Methods*, Deborah Hatfield et al. (presented at International Language Assessment Conference, Horsleys Green, UK, 1993).