A Concurrent Validity Study of Self-Assessments and the Federal Interagency Language Roundtable Oral Proficiency Interview

Charles W. Stansfield, Jing Gao, William P. Rivers

Introduction

The National Language Service Corps (NLSC) was established in 2006 as a pilot program under the auspices of the National Security Education Program (NSEP). This new organization was tasked with providing and maintaining a standing civilian corps of certified bilinguals who would be available for service to federal government agencies as they are needed, and to state and local government agencies in time of emergency. The intent of the NLSC is to fill the gap between full-time language services professionals and individuals who wish to volunteer for temporary services for short- or medium-term assignments. The NLSC recruits a readily-available pool of individuals who have expertise in languages that are important to the security and welfare of the nation. The NLSC must qualify applicants as part of its enrollment process. The NLSC uses the Federal Interagency Language Roundtable Language Proficiency Skill Level Descriptions (the ILR scale) in speaking, reading, and listening as a basis for determining eligibility for Charter membership. The NLSC requires candidates to have proficiency in English and the foreign language at ILR level 3 in the domains of speaking, reading, and listening. All NLSC applicants are asked to complete a series of language self-assessment questions in the four communicative skills and English as an initial screen of language proficiency. These self-assessments provide an indication of applicants' levels on the ILR scale, and minimize the overall formal testing requirements by eliminating the need to test individuals who self-assess at levels lower than those required by the NLSC. Formal assessment of English language skills is waived for applicants who attended an accredited high school or college in the U.S. for at least three years and graduated.

All NLSC applicants complete a basic application form, respond to a language-background questionnaire and complete a two-part self-assessment form. The first self-assessment is a series of "Can-Do statements," which are commonly referred to as "Can-Do scales" in the language testing literature. Can-Do scales require the individual to accept or reject an affirmative statement such

as "I can comprehend an oral presentation at a conference on a complex topic in my profession, and I can also comprehend the question-and-answer session immediately following the main part of the talk" (DD Form 2933, Version 4, Sep 2009, National Language Service Corps (NLSC) Detailed Skills Self-Assessment). The Can-Do scales utilized by the NLSC are grouped according to the skill and level on the ILR scale with which they are associated. These Can-Do scales, once completed, provide a useful inventory of the language skills the candidate claims to have. The candidate's responses can be analyzed in order to produce a predicted language proficiency level on the ILR scale for each skill (listening, speaking, reading, and writing). The second self-assessment is a simplified set of ILR skill level descriptions. The candidate reads the description for each skill and selects the one that best describes his or her language proficiency in that skill. An example of such a simplified ILR description follows for level 4 listening, i.e., advanced professional proficiency level, (DD Form 2934, Sep 2009, National Language Service Corps (NLSC) Pilot Global Language Self-Assessment).

> I can understand all forms and styles of speech pertinent to my social and professional needs. This includes speech involving extensive and precise vocabulary, subtleties and nuances in standard dialects of the language, and technical discussion on professional topics within the range of my knowledge. I can understand language tailored to different audiences and purposes, including persuasion, representation, counseling, and negotiating. I can readily infer meanings and implications. I can easily understand all social conversations, radio broadcasts, and phone calls. I may experience some difficulty understanding speech heard under unfavorable conditions, such as through a poor quality loudspeaker or radio or in a noisy room.

The selected skill level description can serve to confirm or disconfirm the predicted language skill level obtained from the analysis of the responses to the Can-Do statements. The self-ranking on the ILR scale is then conjoined with the predicted score on the Can-Do scales to produce a composite score and the predicted language proficiency level.

If the candidate demonstrates proficiency at ILR level 3 or higher on the predicted language proficiency rating, he or she will undergo formal testing of language skills.

The self-assessments were used as an acceptance tool during the piloting of the NLSC program. The NLSC member's language skills are further tested using a direct measure of language skills prior to any assignment. If the Can-Do self-assessments can be shown to predict scores on direct assessment of language proficiency, then the use of self-assessment could continue for some categories of languages in the permanent program and provide a significant cost reduction for the NLSC.

The validity of this kind of self-assessment instruments is well documented in the literature (ALTE, 2002; Clark, 1981; Davidson & Henning, 1985; Heilenman, 1990; LeBalanc & Painchaud, 1985; Roever & Powers, 2005; Ross, 1998; Shrauget & Osberg, 1981; Tannenbaum, Rosenfeld, Breyer, & Wilson, 1999; Wilson, 2000). In the context of the NLSC, Reed and Stansfield (2006) pointed out the potential advantages of using self-assessment techniques to meet the screening needs of the organization.

Research Design

In 2009, approximately 1,800 NLSC membership applicants across the U.S. filled out the self-assessments online. The self-assessments consist of two parts: the Can-Do statements and the global assessments. Both instruments assess four skills: listening, speaking, reading and writing. However, in the dataset, data corresponding to only three subsets of the Can-Do statements (listening, speaking, and reading) are available. The data for the writing subset of Can-Do statements (hereinafter Can-Dos) were not available for this analysis, since the Government is not currently using it. The following analyses were carried out on 323 admitted candidates to the NLSC, because only these candidates have been formally tested in the target language.

The 158 Can-Dos presented to candidates describe concrete tasks. Some of the tasks are general; and some relate to the work setting. The numbers of statements by domain are: 40 listening, 48 speaking, 32 reading, and 38 writing. The 158 Can-Dos take only a short amount of time to administer and complete. These statements are appropriate for a broad range of settings. For scoring purposes, each Can-Do statement was assigned a skill level between 1 and 5 on the ILR scale. Although there were about four Can-Do statements at each level, an examinee had to answer all statements affirmatively in order to be rated at that level.

For the global self-assessment, if the candidate's proficiency is substantially better than one level but not consistently as good as the next higher level, they are directed to select the appropriate "plus" level. There is no

description of the plus levels on the self-assessment forms. With this format, the candidate can read and understand the scale quickly, and can make a fairly accurate self-placement at one of the nine points on the scale. For purposes of statistical analysis, the plus level is interpreted as 0.6 levels higher than the base level. For example, level 3+ is converted to a numeric score of 3.6.

Characteristics of the Sample

The released dataset consists of 323 observations, including 295 OPIs for which there are corresponding self-assessments. The 323 candidates were tested for their skills in a total of eight languages: 129 for Chinese-Mandarin (39.9%), two for Hausa (0.6%), 18 for Hindi (5.6%), 18 for Indonesian (5.6%), eight for Marshallese (2.5%), 88 for Russian (27.2%), 38 for Thai (11.8%), and 22 for Vietnamese (6.8%). Of these candidates, 228 (70.6%) candidates grew up abroad; 120 (37.2%) candidates received their high school education in the U.S. and 173 (53.6%) candidates attended primary college/university in the U.S. Regarding educational background, 196 (28.5%) candidates hold bachelor's degrees, 29 (9%) hold master's degrees, and six (1.9%) hold doctoral degrees.

Results of the Self-assessments and OPIs

The self-assessments required candidates to come up with subjective estimates as to how well they could perform various tasks related to overall competence in the target language based on whether they could perform various tasks. In general, as the literature shows, respondents tend to overestimate their abilities in performing discrete tasks, in part because making accurate judgments about one's own ability relevant to something as complex as language is not easy. It is also not easy to internalize and apply an unfamiliar scale without training. Score distributions for the self-assessment variables are skewed, while score distributions for the OPI scores are closer to normal distribution. Table 1 below gives a breakdown of the self-assessments and the OPI.

The mean ratings of these NLSC candidates were 4.6 (4+) on Can-Do listening, 4.3 (a mid-range 4) on Can-Do speaking, and 4.5 (a high 4) on Can-Do reading. Their mean rating on the global self-assessments were 4.6 on Global listening, 4.2 on Global writing, 4.5 on Global reading, and 4.5 on Global speaking. These are high means, compared with a mean of 3.0 on the OPI test. At first glance, these numbers might seem unusually high. If we consider the fact these candidates are admitted members, these means do not seem as odd. The candidates generally perceived themselves to be quite competent in the target language.

Table 1: Score Distributions of Self-assessments and the OPI

ILR Level	1	1+	2	2+	3	3+	4	4+	5
					OPI				
N	1	8	27	46	105	97	10	0	1
%	0.30%	2.50%	8.40%	14.20%	32.50%	30%	3.10%		0.30%
					Can-Do: listening				
N	0	0	1	2	27	4	37	10	240
%			0.30%	0.60%	8.40%	1.20%	11.50%	3.10%	74.30%
					Can-Do: speaking				
N	0	0	7	8	45	6	53	5	197
%			2.20%	2.50%	13.90%	1.90%	16.40%	1.50%	61%
					Can-Do: reading				
N	0	1	5	7	48	4	32	3	221
%		0.30%	1.50%	2.20%	14.90%	1.20%	9.90%	0.90%	68.40%
					Global: listening				
N	0	0	0	3	19	30	25	31	211
%				0.90%	5.90%	9.30%	7.70%	9.60%	65.30%
					Global: writing				
N	4	3	8	19	30	29	22	34	157
%	1.20%	0.90%	2.50%	5.90%	9.30%	9.00%	6.80%	10.50%	48.60%
					Global: reading				
N	0	1	3	11	22	23	27	32	201
%		0.30%	0.90%	3.40%	6.80%	7.10%	8.40%	9.90%	62.20%
					Global: speaking				
N	0	0	1	9	20	24	33	30	203
%			0.30%	2.80%	6.20%	7.40%	10.20%	9.30%	62.80%

Research Design: Predictive Validity Study

Predictive validity is the extent to which a score on a scale or test predicts scores on some other measure, i.e., the criterion. The most common use of predictive validity is in selecting students for university admission. A high correlation between the admissions test and the criterion variable (grades in the year following admission) indicates that the selection procedure worked well, while a low correlation signifies that something is wrong with the selection method. For NLSC self-assessments to have predictive validity, the correlation between the self-assessment scores and formal language proficiency tests needs to be statistically significant and of at least moderate effect size. It was decided to have the Oral Proficiency Interview (OPI) score serves as the criterion measure for evaluating the validity of the self-assessments. The OPI is a standardized procedure for eliciting and rating functional speaking proficiency. It is a criterion-referenced, direct interview. The OPI measures how well a person speaks a language by comparing their performance of specific language tasks,

not with other candidates, but with the criteria for each level of the ILR scale for speaking. The OPI is a carefully structured conversation between a certified interviewer and the candidate. The validity of OPI and the skill level descriptions for speaking have been documented in research studies (Dandonoli & Henning, 1990; Kenyon & Stansfield, 1992). In this study, the OPIs were administered by Language Testing International, under contract with the NLSC and were scored using the ILR scale.

Research Questions

This validity study addresses four primary research questions:

Research Question 1: Among Can-Dos and global self-assessments, which generated higher self-ratings and which generated lower self-ratings?

Research Question 2: Are there statistically significant correlations between self-assessment scores and the direct measures of language proficiency? What is the relationship among scores on the two types of self-assessment instruments?

Research Question 3: What is the effect size and practical utility of the correlations? How do the correlations compare with those found in predictive validity studies of high stakes tests such as the GRE and the SAT?

Research Question 4: What is the predictive validity of the Global self-assessments and the Can-Dos respectively in predicting an OPI score?

Results

Research Question 1

Among Can-Dos and Global self-assessments, which generated higher self-ratings and which generated lower self-ratings?

The first research question of this study concerns whether scores from two types of self-rating instruments are alike on the ILR scale. Statistically equivalent scores on the two instruments support the utility of these measures as comparable screening tools in the NLSC certification process. The Can-Do and the Global self-rating scale assess almost the same constructs. Construct refers to the knowledge, skill, or ability that is being tested. In other words, the two types of instruments are different operationalizations of the same concepts. Both the global assessments and the Can-Dos assess the candidates' listening, speaking and reading ability.

This research question was addressed by examining the means on the Can-Do and the global self-rating scales (see Table 2). Paired sample t-tests have been conducted to compare the means of two comparable variables. For example, a paired-sample t-test was conducted to compare the self-rating of speaking scores obtained by Can-Dos and Global assessments. A paired sample t-tests computes the difference between the two variables for each case, and tests to see if the mean difference is significantly different from zero.

Table 2: Means of Can-Dos vs. Global self-assessments

		Mean	S.D.	N
Pair 1	CAN-DO LISTENING	4.66	.66	318
	GLOBAL LISTENING	4.61	.65	318
Pair 2	CAN-DO READING	4.47	.87	319
	GLOBAL READING	4.52	.78	319
Pair 3	CAN-DO SPEAKING	4.39	.86	319
	GLOBAL SPEAKING	4.56	.71	319

Table 3 gives the results of the paired sample t-tests, including the differences in means by skill. There is no significant difference in listening scores obtained by the Can-Dos (M=4.66, SD=.66) and the Global assessments (M=4.61, SD=.65); t (317) =1.78, p=.08. Similarly, there is no significant difference in reading scores obtained by the Can-Dos (M=4.47, SD=.87) and the Global assessments (M=4.52, SD=.78); t (318) =-1.67, p=0.10. However, there is a significant difference in the speaking scores obtained by the Can-Dos (M=4.39, SD=.86) and the Global assessments (M=4.56, SD=.71); t (318) =-0.16, p=0.00.

Table 3: Paired sample t-tests

		Mean	Std. Error Mean	t	df	Sig. (2-tailed)
Pair 1	CAN-DO: LISTENING-GLOBAL: LISTENING	0.05	0.03	1.78	317	0.08
Pair 2	CAN-DO: READING-GLOBAL: READING	-0.05	0.03	-1.67	318	0.10
Pair 3	CAN-DO: SPEAKING-GLOBAL: SPEAKING	-0.16	0.03	-5.102	318	0.00

These results suggest that for listening and reading, the type of self-assessment instrument used does not have an effect on the ILR-scale scores. However, for speaking, the type of self-assessment does have an effect on the scores. Specifically, our results suggest that the Global self-assessment gives a higher speaking score than the Can-Dos. The NLSC could confidently state that self-assessment scores are generally comparable across the self-assessment instruments that assess listening and reading skills.

Research Question 2

Are there statistically significant correlations between self-assessment scores and the direct measures of language proficiency? What is the relationship among scores on the two types of self-assessment instruments?

One of the most important goals of the self-assessment is to provide information on the candidates' true ability in the target language. In this study, the official OPI score is treated as the indicator of the candidates' true foreign language proficiency. The predictive validity is quantified by the correlation coefficient between the two sets of measurements obtained for the same sample—the measurement performed by the self-assessments and by the face-to-face OPI. Correlation coefficients, which fall in the range of -1.00 to +1.00, reflect the strength of the linear relationship between scores on different tests. A high positive correlation between two tests indicates that candidates who obtain a high score on one test are likely to obtain a high score on the other test.

The Pearson product moment correlation coefficients between the predictor variables and the criterion variables are shown in Table 4. Three of the constructs (listening, speaking, and reading) measured by the Can-Do statement scales, are compared to the four constructs (listening, writing, reading, and speaking) measured by the global assessment scales.

Analysis of the data indicates that the highest correlation is between the criterion (OPI scores) and the global assessment of listening ability (r=0.54). The global self-assessment of speaking ability also yields a high correlation (0.49) with the criterion measure (OPI). Next in order of magnitude is the global assessment of reading ability (r=0.47). Both reading and listening subsets of the Can-Dos yielded a correlation of 0.45 with the criterion variable.

At the time of this study, self-assessment data of non-admitted candidates was not available to the researchers. Upon receiving the data of the non-admitted applicants, we will carry out a correction of the correlation coefficients for the restriction of the range of the predictors.

Table 4: Pairwise Pearson correlation coefficients: intercorrelations of predictor variables and criterion variables

Correlations

		OPI	CAN-DO: LISTENING	CAN-DO: SPEAKING	CAN-DO: READING	GLOBAL: LISTENING	GLOBAL: WRITING	GLOBAL: READING	GLOBAL: SPEAKING
OPI	Pearson Correlation	1.00							
	N	295							
CAN-DO: LISTENING	Pearson Correlation	0.45**	1.00						
	Sig. (2-tailed)	0.00	0.00						
	N	293	321						
CAN-DO: SPEAKING	Pearson Correlation	0.41**	0.74**	1.00					
	Sig. (2-tailed)	0.00	0.00	0.00					
	N	293	321	321					
CAN-DO: READING	Pearson Correlation	0.45**	0.70**	0.80**	1.00				
	Sig. (2-tailed)	0.00	0.00	0.00	0.00				
	N	293	321	321	321				
GLOBAL: LISTENING	Pearson Correlation	0.54**	0.73**	0.71**	0.75**	1.00			
	Sig. (2-tailed)	0.00	0.00	0.00	0.00	0.00			
	N	291	318	318	318	319			
GLOBAL: WRITING	Pearson Correlation	0.43**	0.61**	0.70**	0.79**	0.72**	1.00		
	Sig. (2-tailed)	0.00	0.00	0.00	0.00	0.00	0.00		
	N	280	305	305	305	305	306		
GLOBAL: READING	Pearson Correlation	0.47**	0.70**	0.68**	0.80**	0.80**	0.85**	1.00	
	Sig. (2-tailed)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	N	292	319	319	319	319	306	320	
GLOBAL: SPEAKING	Pearson Correlation	0.49**	0.71**	0.76**	0.73**	0.89**	0.75**	0.78**	1.00
	Sig. (2-tailed)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	N	292	319	319	319	319	306	320	320

** Correlation is significant at 0.01 level (2-tailed).

Equation (1) from Lord and Novick (1968, p. 143) will be utilized. This correction uses the ratio of the observed variance of the predictor in the sample to the observed variance of the predictor in the population. The equation is

$$\rho = \sqrt{\frac{1}{1 + \frac{s_x^2}{\sigma_x^2}\left(\frac{1}{r_{yx}^2} - 1\right)}} \qquad (1)$$

where s_x^2 is the observed variance in the sample and σ_x^2 is the observed variance in the population. When statistical corrections were made to counteract the dampening effects of selection, the correlations of the predictor and the criterion measure usually increase substantially (Raju & Brand, 2003).

The correlations between OPI scores and the self-assessment scores are all statistically significant. In other words, the self-assessments correlate well with a direct measure whose validity is widely accepted. Some of the self-assessments measure the same construct as the OPI, and some measure different constructs, but the four language skills are presumably interrelated.

The relationship among the scores on the two types of self-assessment instruments is addressed by examining patterns of correlations to look for evidence of convergent and discriminant validity. Convergent validity is the degree to which scores on a variable are similar to scores on other variables that are presumed to measure similar skills. Discriminant validity is the degree to which the scores on a variable are different from scores on other variables that it theoretically measures different skills. As expected, in Table 4 the highest correlations appear in the "same construct, same instrument" column, and lowest correlations tend to appear in the "different construct, different instrument" column. For example, the correlation between two reading assessments (Can-Do and global) is higher (0.80) than the correlations between the same two reading assessments and the global speaking self-assessment (0.68 and 0.73). This supports the convergent and discriminant validity of the self-assessments.

Research Question 3

What is the effect size and practical utility of the correlations? How do the correlations compare with those found in predictive validity studies of tests such as GRE and SAT?

In the context of NLSC, it is useful to know not only whether the correlations are statistically significant, but also the size of the observed relationship. In practical situations, effect size is helpful in decision-making,

since a highly significant relationship may be uninteresting if its effect size is small. Therefore, reporting effect size is considered good practice when presenting empirical research findings. By convention, correlation coefficients of 0.10, 0.30, and 0.50 are termed small, moderate, and large respectively in terms of their effect size (Cohen, 1988). In this study, the correlation coefficients between the criterion variable (OPIs) and the self-assessments were in the range of 0.41 to 0.54 (see Table 4), which indicates the effect sizes of the correlation coefficients are from moderate to large. In interpreting these results, it should be kept in mind that the correlation coefficient results are not corrected for the restriction of range.

Clark and Swinton (1979) reported a validity coefficient (0.48), which represented the correlation between a single oral proficient interview (OPI) rating and an examinee's self-rating. Heilenman (1990) reported a correlation of 0.33 between course grades and undergraduate students' self-assessments of their French language skills (grammar, vocabulary, accuracy, and fluency). Ross (1998) conducted a meta-analysis of studies dealing with self-assessment in second and foreign languages. For reading, he located 23 correlations, with an average r=0.61. In another summary of research on second-language self-assessments, Oscarson (1997) concluded the accuracy of the assessment depends to a considerable degree on the purpose of the assessment. Compared with correlation coefficients reported in other high risk test validity studies, such as GRE and SAT (Angronow & Studley, 2007; Geiser & Studley, 2002), the NLSC self-assessment instruments have relatively high correlation coefficients with the criterion measure. This provides evidence supporting the validity of the self-assessments.

Research Question 4

What is the predictive power of the global assessments and the Can-Dos respectively?

Although the correlations between individual predictor variables and the criterion variables are interesting, they do not provide the full picture of how well a combination of predictors predicts official language test scores. To evaluate the utility of a combination of predictors, subsets of the predictor variables were entered into regression equation.

In the first step of the modeling process, all seven predictors are entered in the regression equation. Table 5 displays results of the regression of OPI scores on seven self-assessment variables. Unfortunately, only one predictor, the listening global, is significant (alpha > 0.05), which seems theoretically

questionable. One possible reason for this outcome is the high correlation between the seven predictors. For example, the correlation between global speaking self-assessment and the listening Can-Dos is as high as 0.70. In general, the correlation between the predictors is in the range of 0.60 and 0.80. When predictors correlate highly among themselves, it increases the possibility of multicollinearity. The greater the multicollinearity, the greater the standard error and the smaller the t-statistics.

Table 5: Regression of OPI scores on Can-Dos and global assessments

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	T	Sig.	Tolerance	VIF
1	(Constant)	0.69	0.23		2.97	0.00		
	CAN-DO: LISTENING	0.09	0.07	0.10	1.21	0.23	0.36	2.78
	CAN-DO: SPEAKING	-0.04	0.07	-0.06	-0.66	0.51	0.26	3.83
	CAN-DO: READING	0.04	0.07	0.07	0.61	0.54	0.22	4.57
	GLOBAL: LISTENING	0.35	0.11	0.40	3.22	0.00	0.17	5.88
	GLOBAL: WRITING	0.01	0.06	0.02	0.15	0.88	0.22	4.62
	GLOBAL: READING	0.05	0.09	0.07	0.58	0.56	0.18	5.60
	GLOBAL: SPEAKING	0.01	0.10	0.01	0.08	0.94	0.17	5.95

a. Dependent Variable: OPI

In predictive validity studies, it is desirable to find large and statistically significant multiple correlation coefficients that account for a substantial amount of variation in the predictor. For example, Linn and Hastings (1984) found a multiple correlation of 0.46 for Law School Admission Test (LSAT) scores combined with undergraduate grades in predicting first-year law school grades, which indicates LSAT and undergraduate grades account for 21.16% of the variance in first-year law school grades. In addition to gauging the strength of the predictors considered simultaneously, multiple regression can also be used to evaluate the utility of each predictor. The regression slopes provide one indicator of the relationship between a predictor and the criterion. If the regression coefficient for a particular variable is statistically significant, it can be concluded

that the variable is important for accounting for variation in the criterion. However, it is difficult to isolate the relative contributions of each predictor from the regression coefficients.

Table 6: Regression Models

		MODEL 1	MODEL 2
PREDICTED VARIABLE	OPI	X	X
PREDICTOR VARIABLES	CAN-DO: LISTENING	X	
	CAN-DO: SPEAKING	X	
	CAN-DO: READING	X	
	GLOBAL: LISTENING		X
	GLOBAL: WRITING		X
	GLOBAL: READING		X
	GLOBAL: SPEAKING		X

Since the Can-Dos and the global self-assessments are different operationalizations of the same constructs, it was decided to perform two sets of regression analyses. Model one had OPI scores regressed on the Can-Dos subsets (listening, speaking, and reading), and model two had OPI scores regressed on the global self-assessment subsets (listening, writing, reading, and speaking) (see Table 6).

Regression Model One. Three predictor variables included in the first regression model were: Can-Do listening, Can-Do speaking, and Can-Do reading. As can be seen from Table 7, two predictor variables (Can-Do listening

and Can-Do reading) were significant in their contribution to the prediction of the OPI scores. The three Can-Do statement variables altogether accounted for 23.9% (R-Square=0.239) of the variation in OPI scores.

Regression Model Two. The four predictor variables included in the second regression model were: global listening, global writing, global reading, and global speaking. These four variables altogether accounted for 30.5% (R-Square =0.305) of the variation in OPI scores. As can be seen from Table 8, the only significant predictor is global listening. The R-squared values from both model one and model two are slightly higher than R-squared (0.213) obtained in GMAT validity study, using GMAT scores to predict the first year GPA (Sireci & Talento, 2006).

Table 7: Regression of OPI scores on Can-Dos by skill

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1.17	0.21		5.49	0.000		
	CAN-DO: LISTENING	0.23	0.07	0.26	3.33	0.00	0.43	2.31
	CAN-DO: SPEAKING	-0.002	0.06	-0.003	-0.03	0.98	0.30	3.38
	CAN-DO: READING	0.18	0.06	0.28	3.10	0.00	0.33	2.99

a. Dependent Variable: OPI

Table 8: Regression of OPI scores on global assessments by skill

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	0.79	0.22		3.66	0.000		
	GLOBAL: LISTENING	0.39	0.12	0.43	3.64	0.000	0.18	5.58
	GLOBAL: WRITING	0.003	0.06	0.01	0.06	0.96	0.25	3.97
	GLOBAL: READING	0.10	0.08	0.13	1.19	0.24	0.21	4.77
	GLOBAL: SPEAKING	0.01	0.10	0.01	0.06	0.95	0.19	5.42

Conclusions and Discussion

The self-assessment scores are most useful to NLSC if they can provide information that allows NLSC to make important decisions concerning applicants' screening. The Can-Dos and Global self-assessments give a clear profile of target-language strengths and weakness. NLSC hopes to use the self-assessment scores to identify individuals with an adequate level of target language competency to perform their jobs. One major finding of this study was that there were no differences in the listening and reading global assessment ratings and the ratings produced by the Can-Dos, while global assessment of speaking produced higher speaking proficiency ratings than the Can-Dos did.

Some researchers have argued that people tend to overestimate their language skills on self-assessments (Davidson & Henning, 1985). While this study also found that to be the case, the results provide substantial empirical evidence that NLSC applicants can make reasonably effective judgments about their own language skills. As anticipated, the highly selective nature of admission to membership produced a limited range of self-assessment scores. Even with the limited range, both types of Reading/Listening/Speaking self-assessment scores (Can-Do and Global) exhibited significant positive relationships with OPI scores. The effect size of the correlation coefficients is between moderate and large. Compared with correlation coefficients reported in other high stakes test validity studies, such as those of the GRE, SAT and GMAT, the NLSC self-assessment instruments have relatively high correlation coefficients with the criterion measure. This provides evidence supporting the validity of the self-assessments.

In regression analysis, two distinct regression models were fit to the data. Model one had OPI scores regressed on the Can-Do statement skill subsets (listening, speaking, and reading), and model two had OPI scores regressed on the global self-assessment (listening, speaking, reading, and writing). The R-squared values from both model one (R-square=0.239) and model two (R-square=0.305) are slightly higher than R-squared (0.213) obtained in GMAT validity study, using GMAT scores to predict the first year grade point average (Sireci & Talento, 2006).

Overall, the implications of this study are that the Can-Dos and the global self-assessments are reasonably valid measures of language skills in NLSC target languages, and should remain as part of the NLSC screening process.

It should be noted that this validity study is based on a sample of applicants in eight languages, which may not be representative of the whole

population of candidates for the NLSC. Thus, decision-makers should use caution when applying these results to applicants with other foreign languages.

References

Angronow, S. & Studley, R. (2007). Prediction of college GPA from new SAT test scores – a first look. Paper presented at the Annual Meeting of the California Association for Institutional Research. Retrieved from: http://www.cair.org/conferences/CAIR2007/pres/Agronow.pdf

Association of Language Testers in Europe. (1992-2002). *The ALTE can do project*. Retrieved February 13, 2006, from http://alte.org/can_do/alte_cando.pdf

Clark, J. L. D. (1981). Language. In T. S. Barrows et al. (Eds.), *College students' knowledge and beliefs: A survey of global understanding* (pp. 25-35). New Rochelle, NY: Change Magazine Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals, 23*(1), 11-21.

Davidson, F., & Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. *Language Testing, 2.* 164-179.

Geiser, S. and Studley, R. (2002). UC and SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California, *Educational Assessment*, *8*(1), 1-26.

Heilenman, L. K. (1990). Field test for the ITC guidelines for adapting educational and psychological tests. *European Journal of Psychological Assessment, 15* (3), 270-276.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly, 19*, 673-687.

Linn, R. L., & Hastings, C. N. (1984). A meta-analysis of the validity of predictor of performance in law school. *Journal of Educational Measurement*, *21*, 245-259.

Raju, N.S. & Brand, P.A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement, 27*(1), 52-71.

Roever, K., & Powers, D.E. (2005, February). Effects of language of administration on a self-assessment of language skills. *ETS TOEFL*

Monograph series, *27*. Retrieved February 13, 2006, from http://www.ets.org/toefl

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15,* 1-20.

Shrauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others of psychological assessment. *Psychological Bulletin,* 90, 322-351.

Stansfield, C.W., & Kenyon, D.M. (1995). Comparing the scaling of speaking tasks by language teachers and by the ACTFL Guidelines. In A. Cumming & R. Berwick (Eds.), *The concept of validation in language testing* (pp. 124-153). Clevedon, Avon, England: Multilingual Matters.

Tannenbaum, R. J., Rosenfeld, M., Breyer, F.J., & Wilson K. (2000). *Linking TOEIC scores to self-assessments of English-language abilities: A study of score interpretation.* Unpublished manuscript.

Wilson, K.M. (1999). *Validity of global self-rating of ESL speaking proficiency based on an FSI/ILR-referenced scale* (ETS RR-99-13). Princeton, NJ: ETS.