REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
xx-09-2012	Technical Report	Sept 2012

4. TITLE AND SUBTITLE

Analysis of the ILR *Can Do Statements* as a SOFTS Placement Tool

5a. CONTRACT NUMBER
H92222-10-D-0017

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

SWA Consulting, Inc.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

SWA Consulting Inc.
311 S. Harrington St.
Suite 200
Raleigh, NC 27603

8. PERFORMING ORGANIZATION REPORT NUMBER

2011010615

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

HQ USSOCOM, Special Operations Forces Language Office
Attn: FMD-LDEMO-TL
7701 Tampa Point Blvd.
MacDill AFB, FL 33621-5323

10. SPONSOR/MONITOR'S ACRONYM(S)

SOFLO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

A. Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

As part of an ongoing Special Operations Forces Tele-training System (SOFTS) training evaluation project, the Special Operations Forces Language Office (SOFLO) requested that SWA Consulting Inc. investigate the effectiveness of the Interagency Language Roundtable (ILR) Can Do Statements as a placement tool for SOFTS courses. Researchers asked the following questions: 1. Are the Can Do Statements measuring perceived language speaking proficiency consistently and accurately for all students? 2. Are the Can Do Statements related to similar constructs such as students' confidence in their ability to perform language tasks? The SOFTS Can Do Statements measure perceived speaking proficiency, which is an individual's perception of his or her language ability. A sample of 1710 student responses to the Can Do Statements was provided by Progressive Expert Consulting (PEC) Inc. An analysis of course feedback open ended items was also conducted. Although this study provides some support for the use of the ILR Can Do Statements as a placement tool for SOFTS courses, some limitations may restrict the usefulness of the findings. Most notably, this study used data that had already been collected before the research questions were formulated. A follow-up study could involve measuring actual proficiency with an OPI at the beginning of language training for a sub-sample of SOFTS students.

15. SUBJECT TERMS
SOFTS, ILR Can Do Statements, language proficiency

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Surface, Eric A.
U	U	U	UU (SAR)	25	**19b. TELEPHONE NUMBER** *(include area code)* 919-480-2751

Analysis of the ILR *Can Do Statements* as a SOFTS Placement Tool

SEPTEMBER 2012

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED
SPONSORED BY: SOFLO, USSOCOM
RESEARCH CONDUCTED BY: SWA CONSULTING INC.

EXECUTIVE SUMMARY

As part of an ongoing Special Operations Forces Tele-training System (SOFTS) training evaluation project, the Special Operations Forces Language Office (SOFLO) requested that SWA Consulting Inc. investigate the effectiveness of the Interagency Language Roundtable (ILR) *Can Do Statements* as a placement tool for SOFTS courses. These *Can Do Statements* measure *perceived* speaking foreign language proficiency as opposed to *actual* speaking proficiency, as measured by validated instruments such as the Oral Proficiency Interview (OPI). Perceived proficiency is an individual's perception of his or her language ability. Although some research has shown that individuals tend to overestimate their proficiency on such measures (Davidson & Henning, 1985), a study evaluating the National Language Service Corps (NLSC) pilot program found that *Can Do Statements* scores are not significantly different from actual proficiency scores (Stansfield, Gao, & Rivers, 2010)[1]. Furthermore, meta-analytic studies have found moderate correlations between perceived proficiency and actual proficiency (Ross, 1998; Surface, DuVernet, Nelson, McDaniel, & Thornhill, 2011; Surface, Nelson, DuVernet, & Thornhill, 2012). To investigate the effectiveness of this measure of perceived proficiency as a placement tool for SOFTS courses, researchers asked the following questions:

RQ1: Are the *Can Do Statements* measuring perceived language proficiency consistently and accurately for all students?

Conclusion: Overall, the current study provides initial evidence that *Can Do Statements* are a consistent and accurate measure of students' perceived language proficiency and are adequate as a placement tool for most SOFTS students. Classical Test Theory (CTT) analyses indicated that the *Can Do Statements* demonstrated acceptable psychometric properties:

- All subscales[2] had high reliability/internal consistency estimates ($\geq .82$), indicating that students' ratings were representative of their true perceived proficiency
    - Estimates < .70 imply a scale is not consistently measuring the same thing
- Most items had moderate to large item-total correlations, indicating that scale items were highly related to the construct being measured by the test
- In general, item difficulties increased as ILR Level increased (e.g., showing a pattern consistent with what would be expected for this scale)
- There was a strong correlation ($r = .73$) between students' assigned course level and their perceived speaking proficiency level
- Convergent validation evidence provides initial support for the validity of the *Can Do Statements* as a placement tool
- Item Response Theory (IRT) analyses were consistent with CTT findings

[1]The NLCS study assessed the validity of the *Can Do Statements* as a selection tool for individuals with perceived proficiency levels at or above a 3/3/3 ILR rating in listening/reading/speaking. Additional evidence is needed to validate the *Can Do Statements* as a placement tool for individuals with lower perceived proficiency levels.

[2] The SOFTS *Can Do Statements* has four subscales that assess perceived language proficiency at ILR Levels 1 (*Elementary Proficiency)* 2 (*Limited Working Proficiency)*, 3 (*General Professional Proficiency)*, and 4 (*Advanced Professional Proficiency*).

However, some of the *Can Do Statements* may need to be revised, re-assigned to a different ILR Level, or deleted in order for the *Can Do Statements* to make subtle distinctions between individuals with similar perceived proficiency levels and be a maximally effective placement tool. Additional validity evidence may also be needed to ensure that placement decisions are effective.

To this end, the following recommendations to maximize the efficiency of the *Can Do Statement*s as a placement tool were made based on study results:

> ***Recommendation 1:*** Placement decisions based on *Can Do Statements* that are not consistently assessing their assigned perceived speaking proficiency level could be incorrect or misleading. *Can Do Statements* that have perceived difficulty levels that are much greater or less than the other *Can Do Statements* within a particular ILR Level may not be consistently assessing their assigned perceived speaking proficiency level; these items should be evaluated to determine if they should be reassigned to a different ILR Level. The ILR proficiency construct definition, language testing theory, and the statistical properties of the items should be considered when making such decisions. The statistical properties, as determined by this analysis, indicate that the items listed below may be too easy or too hard for their assigned ILR Level (see pp. 11-13 for additional information) and should be considered for revision or reassignment to a different ILR Level.
>
> > ***Can Do Statements*** **that may be too easy for their assigned ILR Level:**
> >
> > **Level 2:** *Can you take and give simple messages over the telephone or leave a message on voicemail?*
> > **Level 4:** *Can you take a discussion in different directions (friendly, controversial, collaborative)?*
> >
> > ***Can Do Statements*** **that may be too hard for their assigned ILR Level:**
> >
> > **Level 2:** *Can you interview an employee, taking care of details such as salary, qualifications, hours and specific duties?*
> > **Level 3:** *Can you use the language to speculate at length about abstract topics such as how some change in history or the course of human events would have affected your life or civilization?*
> > **Level 3:** *Can you carry out any job assignment as effectively as you could in your native language?*
>
> ***Recommendation 2***: *Can Do Statements* that do not differentiate (i.e., distinguish) between individuals with different perceived speaking proficiency levels do not provide useful information for placement decisions and should be considered for revision or removal from the *Can Do Statement*s. The item that did not discriminate well is listed below (see pages pp. 14-17 for additional information).

> **Level 1:** *Are you often unable to finish a sentence because of grammatical or vocabulary limitations?*

> ***Recommendation 3:*** SOFLO language experts should evaluate the *Can Do Statements* to determine whether they effectively assess the full range of difficulty levels represented in the ILR scale. There was a moderate to high amount of variability on the extreme high (4 and 3+) and low (0+ and 1) ends of the students' perceived speaking proficiency ratings within a single course assignment. This limits the ability of the scale to make subtle discriminations between individuals with different levels of proficiency. The more subtle the distinctions are, the more accurate the placement will be. Additional items should be created to better capture the extreme ends of the ILR scale if deemed necessary (see pp. 17-19 for additional information).

RQ2: Are the *Can Do Statements* related to similar constructs such as students' confidence in their ability to perform language tasks?

Conclusion: *Perceived language speaking proficiency* and *confidence in one's ability to perform language tasks* are similar constructs. If *Can Do Statements* ratings and *Confidence* ratings are correlated with each other, this provides evidence that the *Can Do Statements* are measuring perceived speaking proficiency. Overall, there was a large correlation between students' average *Can Do Statements* ratings and their average *Confidence* ratings on the pre-training survey (r = .77, n = 147).

> ***Recommendation 4:*** Although the convergent validation evidence described above provides initial support for the validity of the *Can Do Statements* as a placement tool, additional validation evidence is needed to be confident that the *Can Do Statements* are performing as effectively as possible. SOFLO should consider conducting additional studies further exploring the convergent validity of the *Can Do Statements* and examine the discriminant validity of the *Can Do Statements* with other constructs to which perceived speaking proficiency should and should not be logically related.

Although this study provides some support for the use of the ILR *Can Do Statements* as a placement tool for SOFTS courses, some limitations may restrict the usefulness of this study's findings. Most importantly, this study was not able to use an actual measure of proficiency to investigate the effectiveness of the ILR *Can Do Statements* as a placement tool. This study also had to use data that had already been collected before the research questions were formulated. This limits what questions researchers could ask and how the questions could be answered using the information available. For example, while student *Confidence* ratings on the pre-training survey were able to provide some evidence of convergent validity, there was no data available to investigate the discriminant validity of the *Can Do Statements* as a placement tool. If SOFLO is interested in a rigorous investigation of how the *Can Do Statements* are performing as a placement tool for SOFTS courses and how they can be improved, follow-up studies should be designed to explicitly answer these questions. Some specific recommendations on the content and design of potential follow-up studies are provided below.

Recommendation 5: Although analysis of open-ended survey responses indicated that students were not reporting many issues with course placement, students were not explicitly asked questions about course placement on the surveys, which could have biased the findings. SOFLO should consider adding items to the during-training and post-training surveys that ask students whether they experienced issues that are typically experienced by students who are incorrectly placed in a course (see pp. 21-23 for additional information and a list of potential survey items).

Recommendation 6: SOFLO should consider sponsoring a follow-up study to thoroughly evaluate the *Can Do Statements* as a placement tool for SOFTS courses. As part of this study, SOFLO should consider measuring actual speaking proficiency scores at the beginning of training for a sub-sample of SOFTS students so these scores could be compared to *Can Do Statements* ratings.

TABLE OF CONTENTS

SECTION I: STUDY PURPOSE & BACKGROUND

Study Purpose

As part of an ongoing Special Operations Forces Tele-training System (SOFTS) training evaluation project, the Special Operations Forces Language Office (SOFLO) requested that SWA Consulting Inc. investigate the effectiveness of the Interagency Language Roundtable (ILR) *Can Do Statements* as a placement tool for SOFTS courses. To this end, the researchers asked the following questions:

1. Are the *Can Do Statements* measuring perceived language speaking proficiency consistently and accurately for all students?
2. Are the *Can Do Statements* related to similar constructs such as students' confidence in their ability to perform language tasks?

SOFTS Background

SOFTS is a synchronous online language-training platform that enables trainees around the world to participate in initial acquisition language training (IAT) or sustainment enhancement language training (SET) in real-time with live instructors. SOFTS courses are available in a variety of languages (e.g., Spanish, Italian, Dari, Arabic, Persian-Farsi, Chinese-Mandarin) and a range of proficiency levels. SOFTS course levels correspond to the federal ILR proficiency scale (i.e., 0, 0+, 1, 1+, 2, 2+, 3, 3+, and 4).

Potential SOFTS students who report that they have no exposure to the training language are automatically placed in the Level 0 training course. Students who report having some exposure to the training language are given a self-assessment measure to identify their perceived language proficiency level so they can be placed in a language course that is suitable for their level.

The SOFTS self-assessment measure consists of 27 *Can Do Statements* (e.g., *Can you explain or understand directions to a nearby hotel, restaurant, post office, or other establishment?*) that were adapted from the ILR *Can Do Statements* (see Form DD 2933). The SOFTS *Can Do Statements* has four subscales that assess perceived language proficiency at ILR Levels 1 (*Elementary Proficiency),* 2 (*Limited Working Proficiency),* 3 (*General Professional Proficiency),* and 4 (*Advanced Professional Proficiency*). Subscales for ILR Levels 1-3 have seven *Can Do Statements* each and the subscale for ILR Level 4 has six *Can Do Statements*. Currently, students are placed in the highest level in which they endorse five or six of the *Can Do Statements*. If students endorse three or four *Can Do Statements* in a higher level but do not endorse enough to be placed in that level, they are placed in a plus level.

Study Background

The SOFTS *Can Do Statements* measure *perceived* speaking proficiency. Perceived proficiency is an individual's perception of his or her language ability. Although some research has shown that individuals tend to overestimate their proficiency on such measures (Davidson & Henning, 1985), a study evaluating

the National Language Service Corps (NLSC) pilot program found that *Can Do Statements* scores are not significantly different from actual proficiency scores (Stansfield, Gao, & Rivers, 2010)[3]. Furthermore, meta-analytic studies have found moderate correlations between perceived proficiency and actual proficiency (Ross, 1998; Surface, DuVernet, Nelson, McDaniel, & Thornhill, 2011; Surface, Nelson, DuVernet, & Thornhill, 2012). This research suggests that in low-stakes testing environments, self-ratings can be used in place of actual proficiency ratings.

The *Can Do Statements* is a cost-saving alternative to having a trained rater conduct a one-on-one interview with potential SOFTS students to determine their language speaking proficiency and place them in an appropriate course level. However, if the *Can Do Statement*s are not effective at identifying students' perceived speaking proficiency, the long-term cost in reduced training effectiveness may be greater than the initial cost of interviewing students. When students placed in the same class have different levels of proficiency (i.e., the classroom is *multilevel)*, overall training results can be negatively affected. Inexperienced teachers may adjust the training curriculum to the average-proficiency student, which affects the other students' learning outcomes (i.e., the class is either too easy and students are bored or the class is too hard and students get frustrated, both of which lead to less effort and reduced learning outcomes [Boyd & Boyd, 1989; Wrigley & Guth, 1992] as well as attrition [Wrigley & Guth, 1992][4]).

Can Do Statement Validation Study

As part of an ongoing SOFTS training evaluation project, the Special Operations Forces Language Office (SOFLO) requested that SWA Consulting Inc. investigate the effectiveness of the *Can Do Statements* as a placement tool for SOFTS courses. Researchers conducted qualitative, psychometric and validity analyses to obtain evidence regarding the use of *Can Do Statements* ratings to place students in language training. Specifically, researchers asked the following questions:

1. Are the *Can Do Statements* measuring perceived language speaking proficiency consistently and accurately for all students?
2. Are the *Can Do Statements* related to similar constructs such as students' confidence in their ability to perform language tasks?

For the current study, Progressive Expert Consulting (PEC) Inc. provided SWA Consulting Inc. with a sample of 1710 student responses to the *Can Do Statements* on 18 JAN 11. This sample included all data on file up to that point in time.

[3] The NLCS study assessed the validity of the *Can Do Statements* as a selection tool for individuals with perceived proficiency levels at or above a 3/3/3 ILR rating in listening/reading/speaking. Additional evidence is needed to validate the *Can Do Statements* as a placement tool for individuals with lower perceived proficiency levels.

[4] In some situations, multilevel classes can improve learning outcomes. For example, the lower proficiency students could benefit from exposure to more language input from the higher proficiency students and the higher proficiency students could benefit from additional practice by helping lower proficiency students negotiate word meaning (Corley, 2005). However, for multilevel classrooms to have a positive effect on training outcomes, instructors need to receive training on how to effectively facilitate multilevel classes. This type of instruction also takes more planning, collaboration, and program support (Mathews, Van Horne, & Van Horne, 2006).

Of the initial 1710 students, 929 were categorized as having no language proficiency and were assigned to an ILR Level 0 course. These students' responses were removed from the data set because although *Can Do Statements* ratings for individuals placed in the ILR Level 0 course were included in the data file, their responses were computer generated to indicate that they had no perceived speaking proficiency in the language to be trained (i.e., none of the *Can Do Statements* were endorsed)[5].

An additional 72 students were removed from the data set because their course placement did not seem to be based on their *Can Do Statements* ratings. The response pattern for these students' ratings was consistent with the computer-generated response set (i.e., their ratings implied that they had no perceived language proficiency in the target language) but they were assigned to course levels above the ILR Level 0[6]. The remaining sample of 709 students was used for data analysis.

The majority of participants were enrolled in Spanish (n = 168), French (n = 115) or Modern Standard Arabic (n = 89) at the 0+ (n = 284), 1 (n = 144) or 2 (n = 84) course level. In addition to the *Can Do Statements* data, students' responses to two open-ended survey items and a 22-item measure assessing *students' confidence in their ability to perform language tasks* were included in data analyses. These data were collected on pre-training, mid-training and post-training surveys administered by SWA Consulting Inc. between 24 MAY 10 and 14 FEB 11. These data were collected as part of an on-going training evaluation project funded by SOFLO.

[5] Five of the 929 *Absolute Beginner* students did not have computer-generated responses.
[6] The placement decisions for these individuals did not seem to follow the current placement logic. This implies that the *Can Do Statements* scores were not used to place these students in language training or there was an error in the data set.

SECTION II: RESULTS & RECOMMENDATIONS

This section describes the results of our investigation and also provides recommendations based on those results. First, results from the investigation addressing the first research question are reviewed. The next section reviews the results examining the second research question. Third, results from a preliminary investigation of student open-ended comments provided on course feedback surveys will be reviewed. Finally, study limitations and next steps will be discussed.

RQ1: Are the *Can Do Statements* measuring perceived language speaking proficiency consistently and accurately for all students?

Classical Test Theory (CTT) Analyses. To determine if there were psychometric problems with the *Can Do Statements* at the scale level, CTT reliability analyses were performed. In general, a reliability estimate indicates the degree to which observed scores are representative of *true* or *actual* scores (i.e., there is not a significant amount of error in the estimates). For this study, high reliability estimates would indicate that students' ratings were representative of their true perceived speaking proficiency.

Reliability estimates for each of the four subscales/ILR Levels included in the *Can Do Statements* were calculated. Researchers performed an internal consistency reliability analysis (Cronbach's alpha) to determine whether the items within each level showed consistent responses. Generally, alphas larger than .70 are considered adequate (Hills, 2005). If the alphas are small, it implies that the items in a scale/subscale are not consistently measuring the same thing.

The alphas for the four ILR Levels/subscales are listed below.

- Level One = .88
- Level Two = .90
- Level Three = .87
- Level Four = .82

All of the alphas were above .70, which implies that responses were consistent within the same level of perceived speaking proficiency. It should be noted that the Level 4 subscale only had six items, while the other subscales had seven items. In CTT, decreased scale length alone can result in lower reliability. This should be taken into consideration when interpreting the Level 4 subscale results.

In addition to internal consistency estimates, *item difficulties* and *item-total correlations* provide useful information about item quality. The item difficulty in CTT is equal to the percentage of people who got the item "right" or endorsed the item. Items should get more difficult as the ILR Levels increase and items within each ILR Level/subscale should have similar difficulties. The item-total correlation is an indication of the relationship between responses on a single item and overall scores on the entire test/measure. Large, positive item-total correlations indicate that the item is highly related to the construct being measured by the test. The item difficulties and item-total correlations ranked from least to most difficult within each subscale are provided in Table 1 (p. 12) and Table 2 (p. 13).

Item Difficulty. On average, the item difficulties increased as the subscales/ILR Levels increased; however, a few items seemed to be too easy or too difficult for their specified ILR Level. In other words, students did not respond to these items the same as they responded to other items in that Level/subscale. If the *Can Do Statements* are not consistently assessing their assigned perceived speaking proficiency level, placement decisions based on these items could be incorrect.

> ***Recommendation 1:*** *Can Do Statements* that have difficulty levels that are much greater or less than the other *Can Do Statements* in a particular subscale should be evaluated to determine if they should be reassigned to a different level based on their perceived difficulty level. Language acquisition theory and the statistical properties of the item should both be considered when making such decisions. Items that may be too easy or too hard for their assigned level are listed below.
>
> ***Can Do Statements* that may be too easy for their assigned ILR level:**
>
> **Level 2:** *Can you take and give simple messages over the telephone or leave a message on voicemail?*
>
> **Level 4:** *Can you take a discussion in different directions (friendly, controversial, collaborative)?*
>
> ***Can Do Statements* that may be too hard for their assigned ILR level:**
>
> **Level 2:** *Can you interview an employee, taking care of details such as salary, qualifications, hours and specific duties?*
>
> **Level 3:** *Can you use the language to speculate at length about abstract topics such as how some change in history or the course of human events would have affected your life or civilization?*
>
> **Level 3:** *Can you carry out any job assignment as effectively as you could in your native language?*

Item-Total Correlations. Most of the items had moderate to large item-total correlations, indicating that individuals who endorsed a specific item tended to endorse more of the *Can Do Statements* overall. However, one of the Level 1 items had a small item-total correlation, indicating that there was not a strong relationship between responses to this item and responses to the scale as a whole. In other words, this *Can Do Statement* was not differentiating between individuals who endorsed a lot of *Can Do Statements* and individuals who endorsed few *Can Do Statements*. *Can Do Statements* that do not differentiate between individuals with different perceived speaking proficiency levels do not provide useful information for placement decisions and should not be included in the *Can Do Statements*.

Table 1. CTT Item Difficulties and Item-Total Correlations for Subscales/ILR Levels 1 and 2

Level 1			**Level 2**		
Item	**Diff**	**Item-Total**	**Item**	**Diff**	**Item-Total**
Can you order a meal?	0.69	0.76	Can you take and give simple messages over the telephone or leave a message on voicemail?*	0.53	0.67
Can you buy a needed item, such as bus or train ticket, groceries, or clothing?	0.67	0.80	Can you give detailed information about your job, your family, your house, and your community?	0.40	0.77
Can you make social introductions and use greeting and leave-taking expressions?	0.66	0.69	Can you talk about an everyday event that happened in the recent past or that will happen soon?	0.38	0.77
Can you ask and answer simple questions about date and place of birth, nationality, marital status, and occupation?	0.62	0.72	Can you tell a story?	0.38	0.74
Can you explain or understand directions to a nearby hotel, restaurant, post office, or other establishment?	0.61	0.71	Can you describe in detail a person or place that is very familiar to you?	0.37	0.78
Can you arrange for a hotel room or taxi ride?	0.50	0.64	Can you report on news that you have seen recently on television or read?	0.30	0.76
Are you often unable to finish a sentence because of grammatical or vocabulary limitations?**	0.49	0.39	Can you interview an employee, taking care of details such as salary, qualifications, hours, and specific duties?*	0.14	0.53
Average	**0.61**	**—**	**Average**	**0.36**	**—**

n = 709

Diff = Item Difficulty: the percentage of students who endorsed the item. The larger the item difficulty, the easier the item.

Item Total = Item-Total Correlation: A measure of how related a given item is to the measure as a whole. Items with low item-total correlations do not discriminate well between individuals with different proficiency levels.

A single asterisk (*) indicates items that may be too easy or too difficult for their assigned ILR Level.

Two asterisks (**) indicate an item does not discriminate well between individuals with different levels of language proficiency.

Table 2. CTT Item Difficulties and Item-Total Correlations for Subscales/ILR Levels 3 and 4

Level 3			**Level 4**		
Item	**Diff**	**Item-Total**	**Item**	**Diff**	**Item-Total**
Can you follow and contribute to a conversation among native speakers?	0.25	0.66	Can you take a discussion in different directions (friendly, controversial, collaborative)?*	0.18	0.63
Can you adjust your language to suit your audience, whether you're talking to diplomats, an O7, an E2, close friends, employees, or others?	0.22	0.63	Can you persuade someone effectively to take a course of action in a sensitive situation, such as to improve their health, reverse a decision, or establish a policy?	0.11	0.63
Can you discuss a hypothetical situation?	0.20	0.67	Can you naturally integrate appropriate cultural and historical references into your speech?	0.10	0.68
Can you defend personal opinions about social and cultural topics?	0.19	0.77	Can you prepare and give a lecture at a professional meeting about your area of specialization and debate complex aspects of it with others?	0.06	0.57
Can you cope with unexpected, difficult situations such as broken-down plumbing, an undeserved traffic ticket, or a serious social blunder?	0.15	0.68	In professional discussions, is your vocabulary extensive and precise enough to enable you to convey your exact meaning?	0.05	0.62
Can you use the language to speculate at length about abstract topics such as how some change in history or the course of human events would have affected your life or civilization?*	0.09	0.63	Do you practically never make a grammatical mistake?	0.03	0.47
Can you carry out any job assignment as effectively as you could in your native language?*	0.06	0.48			
Average	**0.17**	**—**	**Average**	**0.09**	**—**

n = 709

Diff = Item Difficulty: the percentage of students who endorsed the item. The larger the item difficulty, the easier the item.

Item Total = Item-Total Correlation: A measure of how related a given item is to the measure as a whole. Items with low item-total correlations do not discriminate well between individuals with different proficiency levels.

A single asterisk (*) indicates items that may be too easy or too difficult for their assigned ILR Level.

Two asterisks (**) indicate an item does not discriminate well between individuals with different levels of language proficiency.

Recommendation 2: *Can Do Statements* that do not differentiate (i.e., distinguish) well between individuals within a specific level should be revised or removed from the *Can Do Statements*. The item that did not discriminate well is listed below.

> **Level 1:** Are you often unable to finish a sentence because of grammatical or vocabulary limitations?

Item Response Theory (IRT) Analyses. To test the item properties further, researchers conducted IRT analyses on the *Can Do Statements*. IRT is a more complex approach to psychological measurement that produces more detailed information about tests, test items and test-taker characteristics. One goal of IRT is to enable practitioners to create tests that consistently and accurately measure a construct across a range of ability or trait levels.

In IRT, two statistics or estimates provide information about the psychometric properties of test items, the *item difficulty* and the *item discrimination*.

The item difficulty estimate in IRT provides similar information as the item difficulty estimate in CTT; however, it is measured on a different scale. If an item has a difficulty of zero, then someone with an average level of the construct being measured will have a 50-50 chance of endorsing that item. Items with negative difficulty levels are easier to endorse and items with positive difficulty levels are harder to endorse. Items should increase in difficulty as the levels increase and all items within a level or subscale should have similar difficulty levels.

The item discrimination estimate in IRT is a measure of how well an item can differentiate between individuals with high levels of a trait and individuals with low levels of a trait. This is similar to the item-total correlation in CTT. Another way of thinking about the item discrimination is that items with high discrimination values are more sensitive to changes in the construct being measured. If an item has a high discrimination value, we can be confident that individuals who endorse the item have a higher perceived speaking proficiency than individuals who do not endorse the item. Items (i.e., *Can Do Statements*) that do not discriminate between high perceived speaking proficiency and low perceived speaking proficiency individuals are not useful and should be rewritten or removed from the *Can Do Statement*s.

The IRT item difficulties and item discriminations for each subscale ranked from least to most difficult are provided in Table 3 (p. 15) and Table 4 (p. 16).

Items that are too easy or too difficult for their specified level (i.e., the item difficulty is closer to the average for the subscale before or after) are marked by an asterisk. Two asterisks mark items that do not differentiate well between individuals with different levels of perceived speaking proficiency.

The results of the IRT analyses are consistent with the CTT findings.

Table 3. IRT Item Difficulties and Item Discriminations for Subscales/ILR Levels 1 and 2

Level 1			**Level 2**		
Item	**Diff**	**Disc**	**Item**	**Diff**	**Disc**
Can you order a meal?	-0.48	2.15	Can you take and give simple messages over the telephone or leave a message on voicemail?*	0.00	2.46
Can you make social introductions and use greeting and leave-taking expressions?	-0.46	1.37	Can you give detailed information about your job, your family, your house, and your community?	0.33	2.86
Can you buy a needed item, such as bus or train ticket, groceries, or clothing?	-0.37	2.98	Can you talk about an everyday event that happened in the recent past or that will happen soon?	0.38	3.13
Can you ask and answer simple questions about date and place of birth, nationality, marital status, and occupation?	-0.30	1.67	Can you tell a story?	0.39	2.66
Can you explain or understand directions to a nearby hotel, restaurant, post office, or other establishment?	-0.21	2.57	Can you describe in detail a person or place that is very familiar to you?	0.40	3.17
Can you arrange for a hotel room or taxi ride?	0.05	1.26	Can you report on news that you have seen recently on television or read?	0.56	3.12
Are you often unable to finish a sentence because of grammatical or vocabulary limitations?**	0.07	0.30	Can you interview an employee, taking care of details such as salary, qualifications, hours, and specific duties?*	1.17	1.99
Average	**-0.25**	**—**	**Average**	**0.46**	**—**

n = 709

Diff = Item Difficulty: Measured in standard deviations around the mean. Items of average difficulty are equal to zero, positive values (+) are more difficult than average, and negative values (-) are easier than average.

Disc = Discrimination: High discrimination values indicate that the item discriminates well between individuals with different speaking proficiency levels. Negative values or values close to zero indicate that the item does not differentiate well between individuals with different proficiency levels.

A single asterisk (*) indicates items that may be too easy or too difficult for their assigned ILR Level.

Two asterisks (**) indicate an item does not discriminate well between individuals with different levels of language proficiency.

Table 4. IRT Item Difficulties and Item Discriminations for Subscales/ILR Levels 3 and 4

Level 3			**Level 4**		
Item	**Diff**	**Disc**	**Item**	**Diff**	**Disc**
Can you follow and contribute to a conversation among native speakers?	0.74	2.05	Can you take a discussion in different directions (friendly, controversial, collaborative)?*	0.96	2.20
Can you discuss a hypothetical situation?	0.85	2.68	Can you persuade someone effectively to take a course of action in a sensitive situation, such as to improve their health, reverse a decision, or establish a policy?	1.21	3.13
Can you defend personal opinions about social and cultural topics?	0.89	2.93	Can you naturally integrate appropriate cultural and historical references into your speech?	1.32	2.13
Can you adjust your language to suit your audience, whether you're talking to diplomats, an O7, an E2, close friends, employees, or others?	0.90	1.72	Can you prepare and give a lecture at a professional meeting about your area of specialization and debate complex aspects of it with others?	1.71	1.95
Can you cope with unexpected, difficult situations such as broken-down plumbing, an undeserved traffic ticket, or a serious social blunder?	1.06	2.41	In professional discussions, is your vocabulary extensive and precise enough to enable you to convey your exact meaning?	1.74	2.33
Can you use the language to speculate at length about abstract topics such as how some change in history or the course of human events would have affected your life or civilization?*	1.35	2.68	Do you practically never make a grammatical mistake?	2.26	1.44
Can you carry out any job assignment as effectively as you could in your native language?*	1.78	1.78			
Average	**1.08**	**—**	**Average**	**1.53**	**—**

n = 709

Diff = Item Difficulty: Measured in standard deviations around the mean. Items of average difficulty are equal to zero, positive values (+) are more difficult than average, and negative values (-) are easier than average.

Disc = Discrimination: High discrimination values indicate that the item discriminates well between individuals with different proficiency levels. Negative values or values close to zero indicate that the item does not differentiate well between individuals with different proficiency levels.

A single asterisk (*) indicates items that may be too easy or too difficult for their assigned ILR Level.

Two asterisks (**) indicate an item does not discriminate well between individuals with different levels of language proficiency.

In CTT, there is only one reliability estimate for a specific test; however, in IRT, the reliability of a test may change depending on the trait levels of the individuals taking it. For example, if a test only has items of average difficulty, it may differentiate between individuals with very low levels of a trait and individuals with moderate to high levels of a trait, but it will not make more subtle distinctions between two individuals who are both high in a trait.

For example, if the *Can Do Statements* only had items that assessed perceived speaking proficiency at ILR Level 2, it could differentiate between students with perceived speaking proficiency below Level 2 and at or above Level 2, but it could not differentiate between two individuals who both have perceived proficiencies above or below Level 2 (e.g., between a student with Level 3 perceived speaking proficiency and a student with Level 3+ perceived speaking proficiency). In order to make subtle distinctions between trait or ability levels, the psychometric properties of a test must be rigorous and the items must assess a range of difficulty levels. Currently, the *Can Do Statement*s are used to make very subtle distinctions between perceived speaking proficiency levels; however, the items may not be capable of making these distinctions. To evaluate how well the *Can Do Statements* subscales were differentiating between students, and to see if placement decisions based on the *Can Do Statements* were effective, additional analyses were conducted.

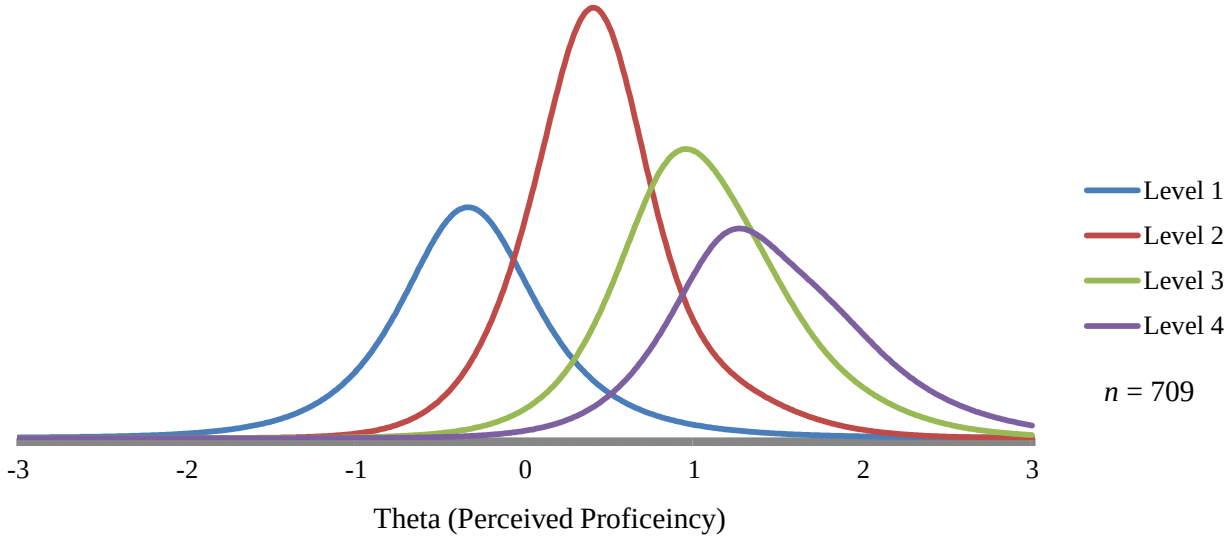
Test Characteristic Curves (TCCs) provide information about entire scales/subscales, not specific items. These curves illustrate the perceived speaking proficiency levels where the scale is able to differentiate between individuals (i.e., at what levels the scale is most reliable). The peak of the curve indicates the amount of information provided, or how well the scale differentiates between individuals. Taller curves represent scales that can make more subtle distinctions between individuals. The width of the curve indicates the speaking proficiency levels assessed by the scale. If a curve is narrow, it can only discriminate between a small range of speaking proficiency levels. If a curve is wide, it can differentiate between a larger range of speaking proficiency levels.

The TCCs for the four *Can Do Statements* subscales are provided in Figure 1 (p. 18). The four *Can Do Statements sub*scales, taken together, are able to discriminate between individuals with perceived speaking proficiency one and a half standard deviations below the mean (perceived *Memorized Proficiency*) and three standard deviations above the mean (perceived *Advanced Professional Proficiency*). The Level 2 subscale provides the most information about perceived speaking proficiency, which means that the *Can Do Statements* are most reliable when placing students who have a perceived speaking proficiency level between *Elementary Proficiency* and *Limited Working Proficiency*. Figure 1 also highlights the large amount of overlap in the Level 3 and Level 4 subscales. This overlap suggests that endorsing Level 4 *Can Do Statements* requires approximately the same perceived speaking proficiency level as endorsing Level 3 *Can Do Statements*. In other words, the Level 4 subscale is not providing a lot of unique information about perceived speaking proficiency.

To investigate whether placement decisions based on the current placement rules was consistent with students' perceived speaking proficiency based on the IRT model, we compared students' perceived speaking proficiency to their assigned course level using a scatter plot (Figure 2, p. 19). Each dot in the scatter plot represents a single student. The spread of dots represents variability in perceived speaking

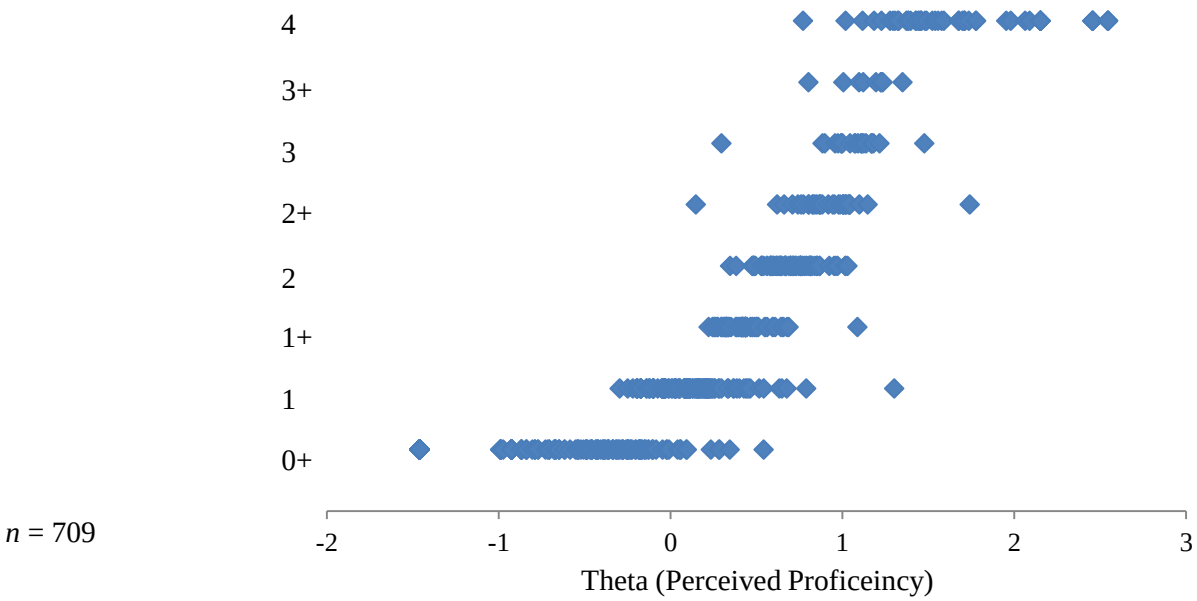
proficiency ratings within a single course level. If the dots are spread out, this implies that students who were assigned to that course level have different proficiency levels.

Figure 1. Test Characteristic Curves for *Can Do Statements* Subscales/ILR Levels 1 through 4



Theta (Perceived Proficiency) is measured in standard deviation units around the mean. Average perceived proficiency is equal to zero, positive (+) theta values indicate perceived proficiency levels that are higher than the mean, and negative (-) theta values indicate perceived proficiency levels that are below the mean.

Figure 2. Comparison of Students' Perceived Speaking Proficiency Ratings and Assigned Course Level



Theta (Perceived Proficiency) is measured in standard deviation units around the mean. Average perceived proficiency is equal to zero, positive (+) theta values indicate perceived proficiency levels that are higher than the mean, and negative (-) theta values indicate perceived proficiency levels that are below the mean.

Overall, there was a strong correlation between students' assigned course level and their perceived speaking proficiency level ($r = .73$). This means that, on average, as perceived speaking proficiency increased, assigned course level increased; however, there was still a large amount of variability in perceived speaking proficiency within each assigned course level.

Students placed in the Level 0+, 1, and 4 courses had the most variability in perceived speaking proficiency levels. Ideally, there should not be a lot of variability in the perceived speaking proficiency levels of students assigned to the same course level. To minimize the amount of variability in students' perceived speaking proficiency levels within a single course, SOFLO must be able to discriminate between individuals with different levels of speaking proficiency, so they can be placed appropriately. The more subtle the distinction in perceived proficiency levels, the more accurate the course placement decisions. In order to make subtle distinctions, the *Can Do Statements* need to be able to measure subtle differences in language ability.

> ***Recommendation 3:*** SOFLO language experts should evaluate the *Can Do Statements* to determine whether they effectively assess the range of difficulty levels represented in the ILR scale. Additional items should be created if necessary.

RQ 2: Are the *Can Do Statements* related to similar constructs such as students' confidence in their ability to perform language tasks?

The results of the statistical analyses suggested that the *Can Do Statements* subscales were consistently measuring the same construct; however, more evidence was needed to be confident that what the *Can Do Statements* were measuring was perceived language speaking proficiency. Researchers demonstrate that tests are measuring what they are supposed to be measuring by assessing construct validity. Two types or aspects of construct validity are typically assessed in an investigation of construct validity: (1) *convergent validity* and (2) *discriminant validity*. This study, however, was only able to assess convergent validity due to the lack of data available to assess divergent validity.

Convergent validation techniques evaluate whether the construct being measured by a test is related to other constructs to which it should logically be related. For the current study, the *Can Do Statements* subscales were compared to each other and to students' ratings of their *confidence in their ability to perform a range of language tasks* (*Confidence)*. The *Confidence* data were collected on the SOFTS pre-training survey distributed during students' first class meeting. The *Confidence* ratings assess students' perceived ability to perform 22 mission-specific language tasks (e.g., *In the language being trained, I am confident in my ability to communicate information about time*). The C*onfidence* measure is broken down into three factors: (1) *Basic Language Tasks (Basic),* (2) *Daily Activity Language Tasks (Daily),* and (3) *Military-Specific Language Tasks (Military-Specific)*. The *Confidence* items were created to evaluate IAT and were developed from critical task lists for SOF Army operators and leaders (SWA Consulting Inc., 2005).

Perceived language speaking proficiency and *confidence in one's ability to perform language tasks* are similar constructs. If *Can Do Statements* ratings and *Confidence* ratings are correlated with each other, this provides evidence that the *Can Do Statements* are measuring perceived speaking proficiency. Furthermore, we would expect the lower levels of the *Can Do Statements* to be more highly related to the *Confidence* factors that assess easier tasks (e.g., *Basic* and *Daily* tasks) and the higher levels of the *Can Do Statements* to be more highly related to *Confidence* factors that assess more difficult tasks (i.e., *Military-Specific tasks*). We would also expect the *Can Do Statements* subscales to be more highly correlated with each other when the levels are proximal versus distal (e.g., the Level 1 *Can Do Statements* subscale should be more highly correlated with the Level 2 *Can Do Statements* subscale than the Level 3 or 4 *Can Do Statements* subscale).

Overall, there was a large correlation between students' average *Can Do Statements* ratings and their average *Confidence* ratings on the pre-training survey (r = .77, n = 147). The correlations for the four *Can Do Statements* subscales and the three *Confidence* factors are provided in Table 5 (p. 21; sample sizes are provided in the parentheses).

Can Do Statements Comparison by Level. As expected, the *Can Do Statements* subscales were more highly correlated with other subscales that immediately precede or follow them than they were to subscales that were more distal. It should also be noted that the correlation between the Level 3 and Level

4 *Can Do Statements* subscales was quite large, providing additional evidence that students responded to these *Can Do Statements* in very similar ways.

Can Do Statements and Confidence Factor Comparison. The Level 1 and Level 2 *Can Do Statements* subscales were more highly correlated with the *Basic Confidence* factor than the Level 3 and Level 4 *Can Do Statements* subscales were. Furthermore, the Level 1 *Can Do Statements* subscale had the lowest correlation with the *Military-Specific Confidence* factor compared to the Level 2 through Level 4 *Can Do Statements* subscales.

Table 5. Correlations for the Four *Can Do Statements* Subscales and Three *Confidence* Factors

	Can Do Level 1	Can Do Level 2	Can Do Level 3	Can Do Level 4	Basic Language Tasks	Daily Language Tasks	Military-Specific Language Tasks
Can Do Level 1	1	–	–	–	–	–	–
Can Do Level 2	.744 (709)	1	–	–	–	–	–
Can Do Level 3	.521 (709)	.753 (709)	1	–	–	–	–
Can Do Level 4	.389 (709)	.597 (709)	.808 (709)	1	–	–	–
Basic Language Tasks	.731 (147)	.655 (147)	.517 (147)	.406 (147)	1	–	–
Daily Language Tasks	.672 (147)	.709 (147)	.567 (147)	.512 (147)	.945 (316)	1	–
Military-Specific Language Tasks	.559 (147)	.705 (147)	.603 (147)	.595 (147)	.842 (316)	.934 (316)	1

All correlations were statistically significant.

> ***Recommendation 4:*** Although the convergent validation evidence described above provides initial support for the validity of the *Can Do Statements* as a placement tool, additional validation evidence is needed to be confident that the *Can Do Statements* are performing as effectively as possible.

Analysis of Course Feedback Open-Ended Items

Researchers were interested in whether students reported problems with course placement on the mid-training and post-training surveys. The rationale for this preliminary investigation was that, if errors in placement were occurring, students would reference issues with placement when providing course feedback.

To assess whether students were reporting issues with placement, researchers analyzed students' open-ended survey comments using the following four codes:

- Course is/was too hard for the individual
- Course is/was too easy for the individual
- Students enrolled in one course have different proficiency levels
- Not related to the course/proficiency level

The following comment is an example of a response that was double-coded as *Course is/was too hard for the individual* and *Students enrolled in one course have different proficiency levels*:

> *"The class was more advanced than I was initially prepared for so there was a learning curve, one of the students had 3+ years of college level Arabic. (2) had 2+ years college Arabic so I have had some challenges but hope to match their skill levels by the end of the course. A good challenge."*
>
> *Arabic Student*

For the training period investigated in this study (i.e., 24 MAY 10 through 14 FEB 11), 76 students responded to the mid-training survey and of these, 53 provided a response to the prompt, *Please provide any additional comments about how your language training course can be improved or made more effective.* For the post-training time point, 131 students responded to survey and of these, 95 provided a response to the prompt, *Please provide any additional comments or recommendations that PEC and/or the training designers can use to improve SOFTS course.* Only six (11%) of the mid-training survey comments and 10 (9.5%) of the post-training survey comments referenced problems with the course level or students' proficiency level[7]. The remaining comments were coded as, *Not related to the course/proficiency level*. Although analysis of open-ended survey responses indicated that students were not reporting many issues with course placement, students were not explicitly asked questions about course placement issues, which could have biased the findings.

> ***Recommendation 5:*** SOFLO should consider adding items to the during-training and post-training surveys that ask students whether they experienced issues that are typically experienced by students who are incorrectly placed in a course. Potential survey items are listed below.
>
> > **Proposed Item 1:** Do you think you were placed in a course level that was appropriate for your language proficiency level? (Yes/No)
> >
> > > **Proposed Item 1A (If response to New Item 1 is *No)*:** Was your course level too easy or too difficult for your language proficiency level? (Too easy/Too difficult)

[7] The students who reported issues with placement on the mid-training survey were not the same individuals who reported issues with placement on the post-training survey.

Proposed Item 2: Did students in your course have different language proficiency levels? (Yes/No)

Proposed Item 2A (If response to Proposed Item 2 is *Yes*): Please elaborate on how students' language proficiency levels differed in your course. (Open-ended response)

Proposed Item 2B (If response to Proposed Item 2 is *Yes*): Did your instructor respond to the differences in students' proficiency levels appropriately (e.g., did your instructor assign tasks or activities that students with different levels of proficiency could all benefit from)? (Yes/No)

Proposed Item 2C (If response to Proposed Item 2 is *Yes*): Please elaborate on how your instructor effectively or ineffectively responded to the differences in students' language proficiency. (Open-ended response)

Current Study Limitations and Next Steps

Although this study provides some support for the use of the ILR *Can Do Statements* as a placement tool for SOFTS courses, some limitations may restrict the usefulness of the findings. Most notably, this study used data that had already been collected before the research questions were formulated. This limits what questions researchers could ask and how the questions could be answered using the information available. If SOFLO is interested in a rigorous investigation of how the *Can Do Statements* are performing as a placement tool for SOFTS courses and how they can be improved, a follow-up study should be designed to explicitly answer these questions.

A follow-up study could involve measuring actual proficiency with an OPI at the beginning of language training for a sub-sample of SOFTS students. These scores could be compared with *Can Do Statements* ratings to determine if the constructs were significantly different. This comparison could provide additional evidence that the *Can Do Statements* ratings are similar enough to actual speaking proficiency scores to be used for placement decisions.

Recommendation 6: SOFLO should consider sponsoring a follow-up study to thoroughly evaluate the *Can Do Statements* as a placement tool for SOFTS courses. As part of this study, SOFLO should consider measuring actual speaking proficiency scores at the beginning of training for a sub-sample of SOFTS students so these scores could be compared to *Can Do Statements* ratings.

REFERENCES

Boyd, J. R., & Boyd, M. A. (1989). *Input-output teacher's manual*. Normal, IL: Abaca Books.

Corley, M. (2005). Differential instruction: Adjusting to the needs of all learners. *Focus on Basics, 7*, Retrieved from www.ncsall.net/index.php?id=736

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.

Hills, A. (2005). *Foolproof guide to statistics using SPSS*. Frenchs Fores, New South Wales: Pearson Education Australia.

Mathews, J., Van Horne, A., & Van Horne, R. (2006). *Promoting success of multilevel ESL classes: What teachers and administrators can do*. Retrieved from www.cal.org/caela/esl_resources/briefs/multilevel.html

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

Stansfield, C., Gao, J., & Rivers, W. (2010, October). *A concurrent validity study of self assessments and the federal interagency language roundtable oral proficiency interview.* Paper presented at the East Coast Organization of Language Testers (ECOL), 9th Annual Conference, Washington, DC.

Surface, E. A., Nelson, K., DuVernet, A., & Thornhill, D. (2012, March).*Perceived versus actual productive proficiency measures: A meta-analytic investigation of the relationship between self-reported and assessed speaking and writing.* Paper presented at the American Association for Applied Linguistics 2012 Annual Conference, Boston, MA.

Surface, E. A., DuVernet, A., Nelson, K., McDaniel, M. & Thornhill, D. (2011, March). *Perceived versus actual proficiency measures: A meta-analytic investigation of listening and reading assessments.* Paper presented at the American Association for Applied Linguistics 2011 Annual Conference, Chicago, IL.

SWA Consulting Inc. (2005, March). *SOF language transformation strategy needs assessment project: Final project report* (Technical Report #20040606). Raleigh, NC: Author.

Wrigley, H. S., & Guth, G. J. A. (1992). *Bringing literacy to life: Issues and options in adult ESL literacy*. San Mateo, CA: Aguirre International.

ABOUT SWA CONSULTING INC.

SWA Consulting Inc. (formerly Surface, Ward, and Associates) provides analytics and evidence-based solutions for clients using the principles and methods of industrial/organizational (I/O) psychology. Since 1997, SWA has advised and assisted corporate, non-profit and governmental clients on:

- Training and development
- Performance measurement and management
- Organizational effectiveness
- Test development and validation
- Program/training evaluation
- Work/job analysis
- Needs assessment
- Selection system design
- Study and analysis related to human capital issues
- Metric development and data collection
- Advanced data analysis

One specific practice area is analytics, research, and consulting on foreign language and culture in work contexts. In this area, SWA has conducted numerous projects, including language assessment validation and psychometric research; evaluations of language training, training tools, and job aids; language and culture focused needs assessments and job analysis; and advanced analysis of language research data.

Based in Raleigh, NC, and led by Drs. Eric A. Surface and Stephen J. Ward, SWA now employs close to twenty I/O professionals at the masters and PhD levels. SWA professionals are committed to providing clients the best data and analysis upon which to make evidence-based decisions. Taking a scientist-practitioner perspective, SWA professionals conduct model-based, evidence-driven research and consulting to provide the best answers and solutions to enhance our clients' mission and business objectives. SWA has competencies in measurement, data collection, analytics, data modeling, systematic reviews, validation, and evaluation.

For more information about SWA, our projects, and our capabilities, please visit our website (www.swa-consulting.com) or contact Dr. Eric A. Surface (esurface@swa-consulting.com) or Dr. Stephen J. Ward (sward@swa-consulting.com).

The SWA Consulting Inc. team members contributed to this report (listed in alphabetical order):

Ms. Sarah Bienkowski
Ms. Amanda Deane
Dr. Amanda Gissel
Dr. Reanna Poncheri Harman
Dr. Eric A. Surface
Dr. Aaron Watson