Assessing the Validity of Can-Do Statements in Retrospective (Then-Now) Self-Assessment

N. Anthony Brown  
Brigham Young University

Dan P. Dewey  
Brigham Young University

Troy L. Cox  
Brigham Young University

Abstract: In this study, the authors evaluated the strengths and limitations of a self-assessment based on ACTFL Can-Do statements (ACTFL, 2013) as a tool for measuring linguistic gains over an internship abroad in Russia. They assessed its reliability, determined how its items mapped with the ACTFL scale, and measured the degree to which students’ self-evaluations matched oral proficiency interview (OPI) test results (i.e., predictive validity). Data revealed a high level of reliability. Furthermore, self-assessment items ascended in the order of difficulty expected (i.e., Superior items were the most difficult, followed by Advanced), but differences between the means for items representing the ACTFL levels were not statistically significant. Finally, while students demonstrated significant gains from pre- to posttests on both the OPI and the self-assessment, correlations between these measures were only moderate.

Key words: learning environment, oral proficiency, Russian as a second language, self-evaluation, study abroad
Introduction

Several years ago, a former student related an experience he had after completing an intensive 12-week Russian course in preparation for spending 2 years in Russia. Just prior to boarding his plane, he called his parents to say goodbye and to impress them with his newly acquired foreign language skills. Quite sincerely, he informed them that he still had some work to do in order to become fluent in Russian but that he was almost there. To this student’s credit, after returning to the United States 2 years later and completing another 2 years of coursework in Russian, he tested Superior on the ACTFL Oral Proficiency Interview (OPI) and since then has proven remarkably adept at learning other foreign languages. Nevertheless, the question arises as to the criteria by which individuals self-assess their language proficiency at any given time.

In the case of the aforementioned student, he assessed himself at the Superior level after only 12 weeks of study, but he still needed another 4 years of studying Russian to reach that level. Certainly, lack of exposure to established proficiency criteria partly explained such a self-assessment. Research has also suggested that the unskilled have a propensity for overrating their abilities (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008). Regardless of the factors contributing to this student’s linguistic self-perception when he spoke with his parents, his knowledge and skills proved to be insufficient at best after setting foot in the target language country less than 24 hours later. Alternatively, more accurate self-assessment can heighten individuals’ awareness of skills that they can and cannot carry out effectively and guide their judgment in setting learning goals. Recognizing the need to implement a more effective self-assessment procedure at their home institution, the researchers developed a survey instrument based on the National Council of State Supervisors of Languages-ACTFL (NCSSFL-ACTFL) Can-Do Statements (ACTFL, 2013) that was designed to compare self-assessed language abilities both prior and subsequent to participating in a 12-week internship program in Moscow. This article presents pre- and post–study abroad data that were gathered using both the Can-Do self-assessment instrument and official OPIs, and also addresses the strengths and limitations of the self-assessment instrument—in particular, the process of linguistic self-assessment. Three research questions guided the collection and analysis of the data as well as the subsequent interpretation of the findings:

1. What is the reliability of the retrospective Can-Do self-assessment instrument used in this study?
2. To what extent do the survey items ascend in a hierarchy of difficulty based on the ACTFL speaking proficiency guidelines?
3. What is the predictive validity of the self-assessment items in determining an OPI score?

Literature Review

Internships, Study Abroad, and Other Experiential Language Learning

Recent literature on the nature and role of experiential language learning informed this research. Experiential language learning is learning that allows a student to acquire and use a foreign language both in and out of class in a variety of formal and informal settings. The number of U.S. students studying abroad has steadily increased over the past decade, reaching more than 340,000 in 2011–2012 (Institute of International Education, 2013, n.p.). Furthermore, in the decade between 2000 and 2010, the number of students who traveled abroad for a credit-earning internship program ballooned from 1,700 to 16,400, with another 8,700 who worked abroad on a noncredit basis (Simon, 2013, n.p.). Perhaps this growth could most accurately be described as one of augmenting existing study abroad programs and thereby providing students “work-study abroad” (Simon, 2013, n.p.).
In addition to these increases, enrollment in domestic intensive study abroad programs, such as those at Beloit College and Middlebury College, continues to grow. In fact, ongoing demand for enrollment in Middlebury Language Schools led to the creation of a West Coast campus to accommodate the dramatic increase in student applications (Duran, 2008). Other experiential learning programs, such as on-campus language-specific housing options, where speakers of the same foreign language agree to speak only in that language so as to improve their proficiency and depth of cultural knowledge in a domestic setting, have also drawn interest (Bown, Dewey, Martinsen, & Baker, 2011; Dewey, Bown, Baker, & Martinsen, 2011; Martinsen, Baker, Dewey, Bown, & Johnson, 2010). In short, interest in and demand for experiential language learning opportunities both at home and abroad have grown considerably over the past decade as students, faculty, and administrators increasingly view foreign languages as an indispensable component of a global community.

Self-Assessment

The literature on self-assessment has typically highlighted several advantages: (1) it is cost-effective and relatively easy to design, administer, and score; (2) it can promote greater learner awareness and self-regulation; and (3) it can motivate students by adding variety to, as well as increased participation in, the assessment process (Dickinson, 1987; Oscarson, 1997; Ross, 1998, 2006). Critics have suggested that self-assessment is not appropriate because learners are not capable of accurately gauging their own abilities, and some asserted that it can lead to lower standards, rewarding students who overestimate their abilities (Ross, 2006). In spite of these criticisms, self-assessment has been used regularly either as the sole measure of language development or as a complement to other measures in research on study abroad (see Badstübner & Ecke, 2009; Carlson, Burn, Ussem, & Yachimovicz, 1990; Dewey, 2004; Dyson, 1988; Magnan & Back, 2007; Meara, 1994; Opper, Teichler, & Carlson, 1990; Teichler & Maiworm, 1997), internships abroad (see Feldman & Bolino, 2000; Gillespie, Braskamp, & Braskamp, 1999; Steinberg, 2002; van’t Klooster, van Wijk, Go, & van Rekom, 2008; Waryszak, 2000), intensive domestic immersion (see Dewey, 2004; Savchenko, 2011), and foreign language residences (see Bown et al., 2011; Martinsen et al., 2010).

To determine the validity of self-assessment as a proxy for more objective measures, researchers in a variety of fields, including math, science, first and second language reading and writing, and medicine, have administered both self-assessments and objective measures and have found that learners are largely able to make good judgments of their own abilities, but the accuracy of these judgments improves as learners reach higher levels of achievement in the domain being assessed (Falchikov & Boud, 1989). In fact, it appears that learners are more aware of their areas of difficulty than they are of what is easier for them (Burson, Larrick, & Klayman, 2006). Researchers have also found that accuracy can vary depending on one’s field, with self-estimates of abilities in the hard sciences being the most accurate (Falchikov & Boud, 1989). Meta-analyses of such studies typically have indicated that correlations between self-assessments and objective measures have found that learners are largely able to make good judgments of their own abilities, but the accuracy of these judgments improves as learners reach higher levels of achievement in the domain being assessed (Falchikov & Boud, 1989). In fact, it appears that learners are more aware of their areas of difficulty than they are of what is easier for them (Burson, Larrick, & Klayman, 2006). Researchers have also found that accuracy can vary depending on one’s field, with self-estimates of abilities in the hard sciences being the most accurate (Falchikov & Boud, 1989). Meta-analyses of such studies typically have indicated that correlations between self-assessments and objective measures often range between r =0.60 and r =0.80, although some certainly are considerably lower and others even higher (Falchikov & Boud, 1989; Oscarson, 1997; Ross, 1998). While one might consider correlations in this range too low to merit direct substitution of a self-assessment for another standardized measure, as Oscarson (1997) pointed out, these correlations are “about the same magnitude as those obtained between different sub-sections in a major language test battery” (p. 179). In other words, subtests purporting to measure aspects of the same construct are as closely related (i.e., correlated) as self-assessment results.
are with more objective measures of language proficiency, suggesting some overlap in terms of the constructs being measured. In the 1980s, researchers conducting large-scale studies of second language learners, including students enrolled in higher education institutions (Clark, 1981) and instructors in K–12 settings (e.g., Hilton, Grandy, Kline, & Liskin-Gasparro, 1985) across the United States, deemed these correlations high enough to merit using self-assessment as a primary tool for gathering information on proficiency.

Some factors that can influence the accuracy of learners' self-assessments include academic record, peer-group and parental expectations, career aspirations, lack of training in self-assessment, cultural background, and self-management skills (see the following for more extensive reviews of the findings in these areas: Falchikov & Boud, 1989; Oscarson, 1997; Ross, 1998). Awareness of these variables and their possible influence can allow educators to make adjustments to improve both the accuracy and interpretation of self-assessment data.

Researchers have found that self-assessments prove most useful when they are tied to tasks that learners are likely to encounter and can imagine themselves experiencing. As Oscarson (1997) pointed out, “Self-assessments are more accurate when based on task content closely tied to students’ situations as potential users of the language in question.” Furthermore, “The evidence is that it is easier for learners to assess their ability in relation to concrete descriptions of more narrowly defined linguistic situations” (p. 183). Providing concrete descriptors with specific examples can help learners more accurately evaluate their abilities. Recent efforts to implement self-assessment in second language education have involved using Can-Do statements connected with tasks that are often the focus of language curricula and that learners ought to expect to encounter in authentic real-world situations. Learners respond to statements such as, “I can give directions from one location to another within my neighborhood,” using options such as the following: (1) “I can do this with no difficulty at all,” (2) “I can do this with little difficulty,” (3) “I can do this with some difficulty,” or (4) “I cannot do this at all.” Such Can-Do statements have been based on the Common European Framework of Reference for Languages/CEFR (Engelhardt & Pfingsthorn, 2013), the ACTFL Speaking Proficiency Guidelines (Dewey, Bown, & Eggett, 2012), and the Interagency Language Roundtable Language Proficiency Skill Level Descriptions/ILR scale (Stansfield, Gao, & Rivers, 2010). Bandura (2006) noted that in order to measure a learner’s self-perceived capability, “items should be phrased in terms of can do” (p. 308). In fact, he asserted that valid self-assessment requires such statements.1 Early indications are that these Can-Do self-assessments can be used effectively to evaluate and facilitate learners’ linguistic progress.

Then-Now Self-Assessment
When it is used to evaluate changes resulting from immersion programs such as internships or study abroad, self-assessment typically consists of having students evaluate their abilities to perform specific linguistic tasks two times: once at the beginning of their experience and once again at the end (see Dyson, 1988; Meara, 1994; Teichler & Maiworm, 1997). This pretest-posttest self-assessment design has been a common feature in educational research for many years, but it has been replaced in recent decades in many studies by a post + retrospective pretest method, in which participants assess their skills only at the end of their study period. In that single end-of-program survey, students retrospectively evaluate their abilities prior to their learning experience (this retrospection is often labeled “Then”) and then provide an additional rating of their abilities following instruction (often labeled “Now”). Even though the instrument is administered in one sitting, the results are analyzed as if it were administered in a traditional pretest-posttest design in which gain
is measured through the difference of the means of the Then questions and those of the Now questions. In an extensive review of research on the use of post-retrospective (i.e., Then-Now) surveys, Lam and Bengo (2003) concluded, “More than three decades of research on post-retrospective method has unequivocally supported this approach over the traditional pretest-posttest approach to measuring change” (p. 78).

A major disadvantage of the traditional pretest-posttest approach is that learners experience a perspective shift between pre- and posttesting because their standard of measurement at posttesting tends to be different from pretesting due to increased experience with the tasks being self-assessed. Learners with little or no experience with a task can often be fairly confident well before they are expected to perform the task, but as the day of reckoning draws closer and they begin to feel the reality approach, their degree of confidence consistently drops (Gilovich, Kerr, & Medvec, 1993); and when they actually have experience with attempting the task, their evaluations can change again. The more experience and proficiency learners have, the more accurately they assess (Caputo & Dunning, 2005; Ehrlinger et al., 2008). Addressing these types of response shift, Lam and Bengo (2003) and Rohs and Langone (1997) argued that a Then-Now approach to self-assessment is more accurate than the traditional pre- and post-self-assessment technique because students completing Then-Now self-assessments are “evaluating themselves with the same standard of measurement or level of understanding on both their posttest responses (how they feel now) and how they felt before the program (then)” (Rohs & Langone, 1997, p. 156). Prior to the learning experience, they may overestimate or underestimate their abilities due to a lack of experience, but following the experience, they know better what the tasks entail on which they are asked to rate themselves.

Based on a review of 40 years of research on post-retrospective self-assessment, Hill and Betz (2005) listed two important benefits of this approach: first, it is a good means of promoting self-efficacy, and second, it is capable of “describing change as experienced subjectively by … participants” (p. 514). Hill and Betz continued, “If the aim is to understand how participants feel about program effectiveness and their personal growth or skill acquisition, the retrospective test provides a more direct assessment of these factors” (p. 514).

**Measurement Principles**

Statistical analyses of surveys often utilize a classical test theory paradigm. This practice can be problematic as the resulting data regularly violate the basic parametric statistical assumption that data are interval in nature. For this reason, it is important to address several key issues related to quantitative analysis of results, including the conversion of raw test scores to measures used in statistical analyses, characteristics of interval data, Rasch scaling, measurement invariance, diagnosing rating scales, and analyzing reliability.

**Conversion of Raw Scores to Measures**

One criticism of scoring in the human sciences is that the data are presumed to be interval when that presumption has not been tested empirically. Stevens (1946), in his seminal work on types of measurement scales, noted that most of the scales used by researchers in the social sciences were actually ordinal, and while using parametric statistics “illegally” can yield results that lead to interesting insights, the means and standard deviations computed on the data are in “error to the extent that the successive intervals are unequal in size” (p. 679). The tendency to assign numbers to objects and then treat the numbers as interval data in doing statistical tests still persists (Bond & Fox, 2007; Crocker & Algina, 1986; Raykov & Marcoulides, 2010). One way to ensure that the data truly meet interval criterion is to convert the raw scores to measures. Raw scores are the observed counts in their original state with no statistical adjustment
(Bond & Fox, 2007). Measures are derived by assigning numerals to objects based on rules (Stevens, 1946). For the measures to be interval, the rules require that the numerals are assigned in a linear manner based on a scale (Crocker & Algina, 1986).

Characteristics of Interval Data
In classical test theory, an instrument is designed to measure the amount of a trait that someone possesses. Each question is designed to demonstrate the extent to which a person possesses that trait; analogously, a series of hurdles on a track represents an opportunity for a person to demonstrate jumping ability. In classical test theory, a person’s ability is estimated by a score based on the total number of items answered correctly (Brown, 1996). So if 10 people went through a course with 10 hurdles, and the first person cleared 8 hurdles while the second only cleared 4, a score of 80% would be assigned for the first person and 40% for the second. An item’s difficulty is calculated by dividing the number of examinees who answer the item correctly by the total number of examinees (Bachman, 2004). So a hurdle that is 2 feet high is expected to be easier than a hurdle that is 4 feet high. If 9 of the 10 people cleared the 2-foot hurdle, the ratio would be 9/10 for an item difficulty of 0.9, and if only 1 of the 10 cleared the 4-foot hurdle, 0.1 would be the item difficulty. Both person ability and item difficulty are dependent on the people who took the survey and the items that were included (Crocker & Algina, 1986). What is more, even if the items are intended to measure the same construct at one ability level (e.g., all the hurdles are supposed to be 3 feet tall), there is no guarantee—and it is in fact unlikely—that the resulting item and person measures have the properties of interval data (Bond & Fox, 2007). Note that in using this hurdle analogy, it has been possible to reference the height of the hurdles through the standard English measurement of feet. With classical test theory, no such standard test-external rulers exist to measure latent traits. If jumping were a latent trait, one would be left to describe the behavior using primitive measurements, such as cubits (i.e., the tip of the finger to the elbow).

The property of equal-intervalness requires that space between any two adjacent scores be equidistant (Stevens, 1946). Furthermore, the same distance between two points should demonstrate the same increase in ability regardless of where that space falls. So if one created a hypothetical self-assessment survey with 25 questions and a student completed it prior to participating in an internship with a score of 10, and completed it post-internship with a score of 15, it would be assumed that the student increased in ability by five points. To be interval data, that difference of five would need to progress in a linear manner, meaning that difference of five points would reflect the same amount of growth regardless if it were from 1 to 6 or from 18 to 23. Not all learning is linear. Some learning is characterized by initial steep gains in ability that flatten out as the person becomes more adept, resulting in a construct that is nonlinear (see Figure 1). To be interval data, that ability increase of five should have the same meaning throughout the scale—that is, from 0 to 25. The issue of equal intervals becomes even more pronounced with survey instruments that use Likert-type
scales (e.g., five-point rating scales ranging from “strongly disagree” to “strongly agree”). Rarely do ordinal categories reflecting affective characteristic progress in an interval manner. While most social scientists acknowledge that the data from their measurement instruments are not truly interval in this sense, they still use parametric statistics (Wright & Linacre, 1989). Furthermore, although some may argue that parametric statistics are robust enough to use with ordinal data (Knapp, 1990; Norman, 2010), the use of Rasch scaling can make the criticism a moot point.

**Rasch Scaling**

Instead of interpreting person ability and item-difficulty estimates from the total points of a test score or survey instrument, Rasch scaling uses the logit scale (Baylor et al., 2011). Logits (or log odds ratios) are the natural logarithm of odds ratios of success and can be converted to and from probabilities. For example, if someone has a 0.6 probability of answering an item correctly, then that person would have a 0.4 probability of answering it incorrectly. Therefore the odds of answering the item correctly are 0.6 divided by 0.4, or 1.5 to 1. By being transformed to a log odds ratio, the measures are now interval data and have additive properties (Linacre, 1991). Those logits can then be transformed back into probabilities. Georg Rasch, the Danish mathematician who developed the measurement model, described the principle as follows:

A person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one. (Rasch, 1960, p. 117)

In Rasch scaling, person ability and item difficulty are measured conjointly so that an examinee with a person ability estimate of a given value will have a 0.50 probability of answering an item with a difficulty parameter of that same value correctly (Linacre, 1991). So if an examinee has a person ability estimate logit of 1.00 and a prompt has an item difficulty parameter logit of 1.00, the probability of the examinee responding to that prompt correctly is 50-50, or odds of 1 to 1. In contrast, if an examinee has a person ability estimate of 1.00 and the prompt has an item difficulty parameter of −1.00 for a distance of two logits between person ability and item difficulty, then the probability of the examinee responding to that prompt correctly is 0.88, or odds of 7.3 to 1.

Besides yielding interval data, another advantage of Rasch scaling is that the parameter estimates for both persons and items have the quality of measurement invariance (Engelhard, 2008). That is, when measuring a unitary construct, person ability estimates are the same regardless of the items that are presented to the examinees, and item ability estimates are the same regardless of who responds to them. A helpful metaphor is to envision “the logit scale as a type of ‘ruler’ for latent traits, with the units on the ruler being logits instead of inches” (Baylor et al., 2011, p. 245). In essence, instead of using primitive measure descriptions to describe the heights of hurdles, the use of “calibrated” hurdles yields an instrument that will provide consistent person ability estimates from whoever uses it. Conversely, if there is a group of individuals with known jumping ability, one can observe their performance on the hurdles to calculate the item difficulty parameters and obtain evidence if the intended difficulty levels corresponded empirically with the actual performance.

Finally, Rasch scaling provides more tools for determining the reliability of the measurement instrument. A rating scale diagnostic is possible by evaluating the following: (1) category frequencies, (2) average logit measures, (3) threshold estimates, (4) category probability curves, and (5) fit statistics (Bond & Fox, 2007). As reliability is defined as the ratio of the true variance to the observed variance (Crocker & Algina, 1986), Rasch reliability reports the relative
reproducibility of results by including the error variance of the model in its calculation. When the reliability is close to 1.0, it indicates that the observed variance of the object being measured (person or item) is close or nearly equivalent to the true (and immeasurable) variance. Therefore, as person reliability approaches 1, the differences in scores are due more to differences in examinee ability rather than to measurement error.

**Methods**

**Background**

Since 2007, students at Brigham Young University (BYU) have interned in Moscow with a number of prestigious nongovernment organizations, political and economic think tanks, hospitals, law firms, businesses, investment banks, consulting firms, news media organizations, and the like. Prior to going abroad, students complete an OPI administered by ACTFL, or as of winter 2013, a computerized ACTFL OPI (OPIc). Shortly before returning to the United States, they complete a post-OPIc in order to ascertain the degree to which their language skills improved while participating in the program. In order to clarify what constitutes a speaker at the Advanced and Superior levels, participants receive a copy of the ACTFL Oral Proficiency Guidelines followed by a careful explanation (ACTFL, 2012a). With such criteria in hand and having received their pre-OPI ratings, students are able to both pinpoint their strengths and weaknesses and target specific areas on which to focus.

In addition to pursuing full-time internships, students attend advanced foreign language courses twice a week (6 hours total) in which they analyze and discuss readings dealing with global issues (e.g., climate change), review grammar topics, and address language-related questions that arise at their respective internships. Consistent contact with and feedback from a native speaker trained in teaching Russian as a foreign language lends structure to their otherwise informal study of the language on the job.

In its infancy, the program spanned only the spring/summer semester in order to stay within the 90-day visa window; however, as demand grew, program dates likewise expanded to include winter and fall semesters. With such growth also came increased accountability, which prompted the development of a survey instrument to promote student self-awareness and self-reflection on their experiences.

**Context**

Most students in upper-division Russian courses at BYU have spent more than 2 years living in a Russian-speaking country. Upon returning, many opt to continue their study of Russian and matriculate directly into third-year advanced grammar courses. Unlike the gradual attrition that normally occurs in university foreign language programs over the course of 4 years, enrollments at BYU swell at the third year and stay consistently high through the fourth year. However, even though most students in upper-division courses have lived in a Russian-speaking country, they lack experience using the target language in a professional capacity.

When educators at BYU were considering possible ways to structure a program that would give students professional language opportunities, the question arose of whether to partner with providers on an individual basis or with an in-country institution that would handle the logistics of placing students in internships, not to mention facilitate housing and visa support. The latter proved far more advantageous than relying on long-distance contact limited by both time zone difference and access to top-level management. In addition to streamlining logistical aspects of the program, partnering with an in-country institution enabled students to enroll in advanced-level Russian language courses designed specifically for learners of Russian as a second language. Aside from choosing to partner with an
institutions of higher education in Russia, the decision was made to focus on Moscow rather than outlying cities in order to take advantage of the immense influx of capital into one city that translates into increased work opportunities for students. In 2006, this initiative was formalized through an agreement between Brigham Young University and the Russian Academy of National Economy and Public Administration.

For this study, everyone that participated in this program was contacted and asked to complete the Can-Do statement Then-Now self-assessment survey. As of winter semester 2014, 68 students had participated in the program, of which 36 (27 male, 9 female) responded to the entire survey and completed a pre- and post-OPI/OPIc.

**Self-Assessment Survey Instrument**

During summer 2013, the authors developed a survey that asked students to respond to Then-Now statements regarding their language abilities on a five-point categorical scale to indicate their degree of confidence. The statements were culled from the major subheadings of the NCSSFL-ACTFL Can-Do Statements (ACTFL, 2013) that reflect the ACTFL proficiency Interpersonal and Presentational Communication standards (see Table 1 for Interpersonal Standards). The survey framed tasks in the form of Can-Do statements and required learners to provide both pre- and post-program estimates of their abilities, (e.g., “I could exchange detailed information on topics within and beyond my fields of interest.”). As these participants already had extensive language experience, their self-assessed levels ranged from Intermediate High to Distinguished. A scale including the following five categories accompanied each task:

1. Could not do this even with extensive preparation.
2. Unsure as to whether I could or could not do this.
3. Could do this with extensive preparation.
4. Could do this with minimal preparation.
5. Could do this without any preparation.

In addition to asking questions about specific linguistic tasks, the survey inquired about changes relating to personal development, e.g., “My internship helped me develop increased self-confidence”; changes in terms of academic commitment, e.g., “My internship served to enhance my interest in academic study”; changes in terms of intercultural development, e.g., “My internship helped me better understand my own cultural values and biases”; and changes in terms of career development, e.g., “My internship helped me acquire skill sets that influenced my career path” (Dwyer & Peters, 2004). Previous participants received an e-mail invitation to participate in a Qualtrics survey (see http://qualtrics.com/ for information on this survey tool) in regard to their experience on the Moscow internship program.

**Findings**

The goals of this study included evaluating the reliability of a Then-Now self-assessment instrument, determining the extent to which the Can-Do statements reflected the hierarchy of the proficiency scale empirically with item-difficulty parameters, and examining the predictive validity of using the self-assessment instrument for participants to self-report their OPI level.

**Reliability of Self-Assessment Instrument**

The first research question examined the reliability of the self-assessment instrument, which comprised Can-Do statements (ACTFL, 2013) spanning the proficiency range of ACTFL Intermediate High to Distinguished levels. To answer this question, researchers applied the Rasch measurement to the results of the survey using the rating scale model for polytomous data. A diagnosis of the functionality of the scale is followed by a reliability analysis of the test scores from the use of the scale.
<table>
<thead>
<tr>
<th>ACTFL Standard</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice Low</td>
<td>I can communicate on some very familiar topics using single words and phrases that I have practiced and memorized.</td>
</tr>
<tr>
<td>Novice Mid</td>
<td>I can communicate on very familiar topics using a variety of words and phrases that I have practiced and memorized.</td>
</tr>
<tr>
<td>Novice High</td>
<td>I can communicate and exchange information about familiar topics using phrases and simple sentences, sometimes supported by memorized language. I can usually handle short social interactions in everyday situations by asking and answering simple questions.</td>
</tr>
<tr>
<td>Intermediate Low</td>
<td>I can participate in conversations on a number of familiar topics using simple sentences. I can handle short social interactions in everyday situations by asking and answering simple questions.</td>
</tr>
<tr>
<td>Intermediate Mid</td>
<td>I can participate in conversations on familiar topics using sentences and series of sentences. I can handle short social interactions in everyday situations by asking and answering a variety of questions. I can usually say what I want to say about myself and my everyday life.</td>
</tr>
<tr>
<td>Intermediate High</td>
<td>I can participate with ease and confidence in conversations on familiar topics. I can usually talk about events and experiences in various time frames. I can usually describe people, places, and things. I can handle social interactions in everyday situations, sometimes even when there is an unexpected complication.</td>
</tr>
<tr>
<td>Advanced Low</td>
<td>I can participate in conversations about familiar topics that go beyond my everyday life. I can talk in an organized way and with some detail about events and experiences in various time frames. I can describe people, places, and things in an organized way and with some detail. I can handle a familiar situation with an unexpected complication.</td>
</tr>
<tr>
<td>Advanced Mid</td>
<td>I can express myself fully not only on familiar topics but also on some concrete social, academic, and professional topics. I can talk in detail and in an organized way about events and experiences in various time frames. I can confidently handle routine situations with an unexpected complication. I can share my point of view in discussions on some complex issues.</td>
</tr>
<tr>
<td>Advanced High</td>
<td>I can express myself freely and spontaneously, and for the most part accurately, on concrete topics and on most complex issues. I can usually support my opinion and develop hypotheses on topics of particular interest or personal expertise.</td>
</tr>
<tr>
<td>Superior</td>
<td>I can communicate with ease, accuracy, and fluency. I can participate fully and effectively in discussions on a variety of topics in formal and informal settings. I can discuss at length complex issues by structuring arguments and developing hypotheses.</td>
</tr>
<tr>
<td>Distinguished</td>
<td>I can communicate reflectively on a wide range of global issues and highly abstract concepts in a culturally sophisticated manner.</td>
</tr>
</tbody>
</table>
Scale Diagnosis

The five-level scale categories (one through five) functioned well within established guidelines (Linacre, 2002). The absolute frequency of each category had a minimum of 10, though categories one and two had a combined relative frequency of only 7% and thus could be candidates for combination to a single category. The average measures for each category increased monotonically without exception, as did the threshold estimates. The threshold estimates had the minimum recommendation of 1.4 logits between each category, thus indicating that each category showed distinction. Furthermore, for the scale to be treated as interval data, it was desirable for the thresholds to be regularly spaced (see Figure 2)—a criterion met by the survey data. An examination of the category probability distributions indicated that each category functioned well (see Table 2) and that none of the outfit mean squares exceeded 2.0. One note of interest: The two most frequent categories were four (43% of responses) and five (29% of responses), indicating that 72% of the time, students self-assessed that they could accomplish the Can-Do tasks with minimal to no preparation. This finding suggests that students perceived the tasks to be fairly easy. Despite this observation, the category descriptions of the scale functioned within the expected parameters, and there was no need to make adjustments to the categories in order to evaluate the reliability of the instrument.

Reliability Analysis

One advantage of a Rasch measurement analysis is that the facets can be compared on a vertical scale showing the link between persons, items, and the rating scale. This vertical scale helps one visualize separation reliability. Figure 3 illustrates the logits in the first column, the “Then” person ability estimate in the second, the “Now” person ability estimate in the third, the item difficulty in the fourth, and the rating scale in the fifth. The person labels are coded as the OPI/OPIc score (six = Intermediate High,
TABLE 2

Human-Rated Holistic Speaking Level Rating Category Statistics

<table>
<thead>
<tr>
<th>Category</th>
<th>Absolute Frequency</th>
<th>Relative Frequency</th>
<th>Average Measure</th>
<th>Outfit</th>
<th>Threshold</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>1%</td>
<td>−2.98</td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>149</td>
<td>6%</td>
<td>−0.30</td>
<td>1.61</td>
<td>−4.13</td>
<td>0.26</td>
</tr>
<tr>
<td>3</td>
<td>544</td>
<td>22%</td>
<td>0.60</td>
<td>0.80</td>
<td>−1.24</td>
<td>0.11</td>
</tr>
<tr>
<td>4</td>
<td>1070</td>
<td>43%</td>
<td>2.80</td>
<td>0.86</td>
<td>1.13</td>
<td>0.06</td>
</tr>
<tr>
<td>5</td>
<td>720</td>
<td>29%</td>
<td>4.91</td>
<td>1.02</td>
<td>4.24</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Categories:
1. Could not do this even with extensive preparation.
2. Unsure as to whether I could or could not do this.
3. Could do this with extensive preparation.
4. Could do this with minimal preparation.
5. Could do this without any preparation.

FIGURE 3

Self-Assessment Vertical Scale

<table>
<thead>
<tr>
<th>#</th>
<th>“Then” Person Ability Estimate</th>
<th>“Now” Person Ability Estimate</th>
<th>Item Difficulty Parameter</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>9-ROB 9-SPE 9-ANT 7-NEV 7-ANT</td>
<td>8-MED 7-RAV 7-RAV 7-RAV 7-RAV</td>
<td>3 5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6-RAY 8-RAV 8-RAV 8-RAV 8-RAV</td>
<td>6-RAY 8-RAV 8-RAV 8-RAV 8-RAV</td>
<td>3 4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>8-RAV 8-RAV 8-RAV 8-RAV 8-RAV</td>
<td>8-RAV 8-RAV 8-RAV 8-RAV 8-RAV</td>
<td>3 3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>7-PER 7-GOD 9-ROB 7-RAV 7-RAV</td>
<td>7-PER 7-GOD 9-ROB 7-RAV 7-RAV</td>
<td>3 1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>8-FAR 8-FAR 8-FAR 8-FAR 8-FAR</td>
<td>8-FAR 8-FAR 8-FAR 8-FAR 8-FAR</td>
<td>3 3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>7-PER 7-GOD 9-ROB 7-RAV 7-RAV</td>
<td>7-PER 7-GOD 9-ROB 7-RAV 7-RAV</td>
<td>3 3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>9-ROB 9-SPE 9-ANT 7-NEV 7-ANT</td>
<td>8-MED 7-RAV 7-RAV 7-RAV 7-RAV</td>
<td>3 5</td>
<td>5</td>
</tr>
<tr>
<td>0</td>
<td>8-ROB 8-ROB 8-ROB 8-ROB 8-ROB</td>
<td>8-ROB 8-ROB 8-ROB 8-ROB 8-ROB</td>
<td>3 3</td>
<td>3</td>
</tr>
<tr>
<td>-1</td>
<td>7-WAH 7-WAH 7-WAH 7-WAH 7-WAH</td>
<td>7-WAH 7-WAH 7-WAH 7-WAH 7-WAH</td>
<td>3 2</td>
<td>2</td>
</tr>
<tr>
<td>-2</td>
<td>6-HEP 6-HEP 6-HEP 6-HEP 6-HEP</td>
<td>6-HEP 6-HEP 6-HEP 6-HEP 6-HEP</td>
<td>3 1</td>
<td>1</td>
</tr>
</tbody>
</table>
seven = Advanced Low, etc.) and randomized initials. Accordingly, in Figure 3, the person label 9-NEW in the Then column represents a student with a pre-internship OPI/OPIc of Advanced High and a pre-internship self-assessment logit of 1. That same student in the Now column has the label 10-NEW, thus indicating a post-internship OPI/OPIc of Superior and a post-internship self-assessment logit of just over 3. The item labels are coded as intended OPI level and question number, so the item labeled AM5 is the fifth survey question that reflects an Advanced Mid self-assessment task. The italicized letters along the vertical axes represent the mean (M), one standard deviation from the mean (S) and two standard deviations from the mean (T) for the Then-Now and Items columns. The horizontal axis indicates the 50% probability threshold, so a person with a logit of 1 (i.e., 9-NEW) had a 50% probability of self-assessing his/her ability to accomplish tasks AM5, AH3, and AH4 in category four (i.e., “can do this with minimal preparation”).

From the reliability, one can calculate how many statistically separate groups existed for both persons and items. Figure 2 shows that the person ability estimates ranged from −3 to 7 on the scale, with a mean of 2.68. The analysis found that the separation reliability among the students was 0.96, with a separation strata index of 5.57, thus suggesting that estimated person ability parameters were indicative of reliable differences in the students’ self-perceptions of their abilities in five distinct levels. The item ability estimates ranged from −3 to 2 on the scale, with a mean of 0. The item separation reliability was 0.95, with a separation strata index of 4.52. Such findings indicate that the items were reliably different from each other and represented four distinct difficulty levels.

Thus, to summarize the findings in regard to the first research question addressing the reliability of the self-assessment instrument: The five categorical levels for each of the Can-Do statements functioned well, although the students perceived the self-assessment items to be quite easy. The reliability estimates of both persons and items approached the upper limit of 1, indicating that the instrument was internally reliable.

**Hierarchy of Item Difficulty Levels**

The second research question asks whether the difficulty of the Can-Do statements ascended in the hierarchical order in which they were predicted. For example, it was predicted that the Intermediate tasks would be perceived to be easier than the Advanced, the Advanced easier than the Superior, and so forth. To answer this question, the item logits from the Rasch analysis were grouped by their intended ACTFL levels and an ANOVA was run. An analysis of the descriptive statistics (see Table 3) shows that the Intermediate items (mean = −0.83, sd = 1.14, 95% CI [−2.02, 0.36]) were perceived to be the easiest, the Advanced (mean = −0.02, sd = 1.00, 95% CI [−0.50, 0.46]) were next, followed by the Superior (mean = 0.38, sd = 0.98, 95% CI [−0.65, 1.41]), and finally the Distinguished terms were perceived as the most difficult (mean = 0.60, sd = 0.57, 95% CI [−0.11, 1.30]) (see Figure 4). An independent-measures ANOVA found that the differences were not statistically significant (F = 2.36, df = 3, p = 0.09).

Revisiting Figure 2, one sees that the self-assessment items were clustered together and that students felt that most of the Can-Do statements could be accomplished successfully with minimal or no preparation. This could be indicative of a failure of the students to understand the intended difficulty of the tasks or an overinflated sense among students of what they could actually accomplish. Regardless, the data suggest that the survey items ascended in a hierarchy of difficulty levels based on the ACTFL Oral Proficiency Guidelines (ACTFL, 2012a) but not to a statistically significant degree.

**Predictive Validity of Self-Assessment Items for OPIs/OPIcs**

The third research question addressed the predictive validity of the self-assessment
instrument for OPI scores. Answering this question involved a three-step analysis: (1) examining if OPI scores changed from the pretest to the posttest, (2) determining whether students perceived a difference in their ability by looking at the person ability estimates from the Then questions to the Now questions, and (3) assessing how well the students’ perceived abilities correlated with their OPI scores.

<table>
<thead>
<tr>
<th>Intended OPI Level</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>95% Confidence Interval for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>Intermediate High</td>
<td>6</td>
<td>-0.83</td>
<td>1.14</td>
<td>0.46</td>
<td>-2.02</td>
</tr>
<tr>
<td>Advanced Low</td>
<td>19</td>
<td>-0.02</td>
<td>1.00</td>
<td>0.23</td>
<td>-0.50</td>
</tr>
<tr>
<td>Mid</td>
<td>8</td>
<td>-0.32</td>
<td>0.97</td>
<td>0.34</td>
<td>-1.13</td>
</tr>
<tr>
<td>High</td>
<td>6</td>
<td>-0.26</td>
<td>1.17</td>
<td>0.48</td>
<td>-1.48</td>
</tr>
<tr>
<td>Superior —</td>
<td>6</td>
<td>0.38</td>
<td>0.98</td>
<td>0.40</td>
<td>-0.65</td>
</tr>
<tr>
<td>Distinguished —</td>
<td>5</td>
<td>0.60</td>
<td>0.57</td>
<td>0.25</td>
<td>-0.11</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
<td>0.00</td>
<td>1.04</td>
<td>0.17</td>
<td>-0.35</td>
</tr>
</tbody>
</table>
Pre-Internship vs. Post-Internship OPIs
To determine the extent of language gain during the internship, pre- and post-internship OPIs were administered. For the pre-internship OPI, the mean score was 7.42 (sd = 1.01), with the median at Advanced Low and mode at Advanced Mid (see Figure 5). For the post-OPI, the mean was 8.45 (sd = 1.09), with the median at Advanced Mid and the mode at Advanced High. As OPI scores do not have the characteristics of interval-level data and the data were not normally distributed, the nonparametric Wilcoxon matched-pairs signed ranks test was conducted with the ratings from the pre- and post-internship OPIs. Results of the analysis indicate significant gain from pre- to post-internship OPIs (Z = −5.57, p < 0.001), with 41 of the 68 subjects scoring higher on the posttest. There were 12 instances in which subjects had the same rating on the pre- and posttest and only two instances in which a student scored lower on the post-internship OPI than on the pre-internship OPI. In those two instances, neither subject crossed a major threshold.

Then vs. Now Self-Assessments
To determine the extent of perceived language gain during the internship, the Then-Now instrument was administered at the end of the experience. The mean person ability estimates of the Then statements were compared to the mean person ability estimates of the Now statements (see Figure 2). As logits are interval data and the data were normally distributed (see Figure 6), a paired-samples t test was used. The difference of the means was −1.88 (sd = 1.64, 95% CI [−2.43, −1.33]), resulting in t = −7.00, df = 36, p < 0.001. The students therefore perceived an increase in language gain using the self-assessment instrument.

Then-Now Self-Assessments and OPIs/OPIcs
To determine the extent to which Then vs. Now self-assessments correlate with OPIs/
OPIcs, a correlation was run between the Then person ability estimates and the pre-internship OPIs/OPIcs and between the Now person ability estimates and post-internship OPIs/OPIcs. The correlation between the Then and the pre-OPI/OPIc was 0.27 \((n = 37, 95\% \text{ CI } [-0.05, 0.56])\), indicating a small to medium effect size. The correlation between the Now and the post-OPI/OPIc was 0.21 \((n = 37, 95\% \text{ CI } [-0.13, 0.49])\), indicating a small to medium effect size. Figure 7 illustrates in scatter chart format a slight trend, but the effect size was just that—slight. To see if the gain in the person ability estimates correlated with gain in OPI scores, a correlation between those variables was run and yielded a correlation of 0.21 \((n = 37, 95\% \text{ CI } [-0.13, 0.49])\)—a small to medium effect size (see Figure 8). Thus, in answer to the third question, both OPIs and Then-Nows showed gain but the effect was small, perhaps owing to the relatively small sample size.

**Discussion**

To determine the strengths and limitations of self-assessment as a tool for evaluating linguistic gains over an experiential language learning setting (internship in Russia), the reliability of the instrument was evaluated, the item difficulties were mapped with the ACTFL scale, and the degree to which students’ self-evaluations matched OPI test results (i.e., predictive validity) was explored. The data revealed a high level of reliability (i.e., the scale functioned within the expected parameters, and there was no need to make adjustments to the categories in order to evaluate the reliability of the instrument); furthermore, there were reliable differences in students’ self-perceptions of their own abilities and questionnaire items proved to be reliably different from each other. Next, data confirmed that the self-assessment items ascended in the order of difficulty expected (Superior items were most difficult, followed by Advanced, and so forth), but differences among the means for
items representing the ACTFL levels were not statistically significant. Finally, while students demonstrated statistically significant gains from pre- to posttests on both the OPI and the self-assessment, correlations between these measures were relatively low.

Given the results presented here, the relative merits and limitations of the self-assessment can be considered. First, it is encouraging that the level of reliability was so high and that the response scale seemed to function well in terms of differentiating individuals and items from each other. As Gronlund (2003) stated, “Unless the results are generalizable over similar samples of tasks, time periods, and raters, we are not likely to have confidence in them” (p. 212). In terms of this self-assessment, a high level of confidence in the consistency of the results is possible. Furthermore, as illustrated in Figure 1, the thresholds for the scale were regularly spaced, indicating a fairly distinct hierarchy in terms of difficulty for the numeric responses. Such regular spacing suggests that a 1 represents the lowest level of difficulty, followed, by a 2, and so forth.

Next, because the data fit the Rasch model, the raw scores were transformed successfully into interval values, thus meeting the assumptions needed to conduct parametric analyses appropriate only for interval data, such as the ANOVA used to determine whether there were significant differences between the means for the items from each of the ACTFL levels. These interval values (logits) further permitted more meaningful comparisons between items and individuals, thus allowing increased confidence in measuring how far apart two individuals were in terms of their perceived abilities, how much more difficult one item was than another, etc.

It is encouraging that the self-assessment items tended to align with the ACTFL categories with which they were associated.
(i.e., they followed the expected hierarchy). However, the mean scores for items at each ACTFL level were not significantly different from each other, indicating that this hierarchy may not be as straightforward as one would expect. Figure 2 illustrates how some Can-Do item results were lower or higher in difficulty than expected (e.g., SU3, which had a logit value of lower than $-1$, indicating it was easier than some Intermediate and Advanced items and below average overall). Figure 2 also depicts an overall tendency for students to rate themselves at an excessively high level, in particular when it came to Superior and Distinguished tasks. The figure maps item difficulty with person ability and demonstrates that item difficulty tended to be lower than person ability. In other words, learners perceived items to be easier than they ought to be. Two questions therefore arise: (1) Are some Can-Do items truly easier or more difficult than other items aimed at the same ACTFL level or sublevel? and (2) Are the mismatches described here due to students’ failure to accurately comprehend the nature and difficulty of the tasks on which students were being asked to rate themselves?

Regarding possible mismatches in difficulty between items at the same level, the necessary alignment between function, text type, content, and accuracy to fulfill the task may not have been communicated clearly to participants. For example, supporting an opinion represents a Superior function, but only when discussing abstract topics with extended discourse. Thus, while an Intermediate speaker could offer a series of sentences with an opinion on where to eat dinner because of price and location, the Superior speaker could speak at length on the effect that globalization and franchised chains have had on locally owned restaurants. This failure to account for the necessary alignment in tasks could have made some tasks appear easier than they were intended to be. In addition, the design of the instrument could have contributed to the students’ tendency
to inflate their self-assessments. To keep the length of the survey down to 36 statements, the items were based on the subheadings rather than the specific examples in the Can-Do statement, which may have overgeneralized the nature of the tasks.

In addition, instructors may believe that they have a familiarity with the proficiency guidelines, but their cursory understanding may cause them to provide student feedback that is higher than warranted. This feedback could, in turn, encourage students to over-inflate their self-assessments. For example, an instructor who knows “bits and pieces” of information about the proficiency scales may think that success on discrete features is evidence that his or her students are at a higher level than they truly are (ACTFL, personal communication, March 4, 2014). If the instructors share that feedback with the students, the students may in turn self-assess too high.

Another source of misconception concerning the scale comes from conflating performance and proficiency. For example, if a student completes the Superior function of supporting an opinion, yet the response has been practiced, with vocabulary and content that has been explicitly taught, the instructor may mistakenly inform the students that they are at the Superior level when in fact there is insufficient evidence on which to make such a judgment. In order to assist language educators in their efforts to provide students realistic feedback, ACTFL developed the ACTFL 2012 Performance Descriptors for Language Learners, which explicitly differentiate proficiency and performance (ACTFL, 2012b). Indeed, misinformed feedback from instructors can hamper the ability of students to accurately self-assess.

As far as students’ possible failure to accurately comprehend the nature and difficulty of the Can-Do tasks, there is evidence that learners regularly fall into this trap when self-assessing (Oscarson, 1997). Strong-Krause (2000) suggested that the more specific and closer the self-assessment is to a real-world task, the more accurate the self-assessment will likely be. In her re-search, she paired self-assessments of difficulty with scenarios that learners were asked to imagine themselves acting out, including descriptions of who the interlocutor would be, several specific things they would need to say to the interlocutor, and what they were expected to accomplish linguistically. In other research, Strong-Krause (personal communication, January 29, 2014) also found that learners with limited experience in an area tend to overestimate their abilities on more difficult tasks more than those with no experience at all. In other words, their limited experience with the tasks can produce confidence, even though the ability they develop through that experience may be minimal. It is possible that learners in this study had just enough experience with what they perceived to be higher-level tasks to develop greater confidence in their abilities to accomplish them.

One possible explanation for students rating some items as much easier or harder than expected even though they were actually at the same ACTFL level is that learners rated individual items relative to other items: that is, if participants thought that they could not do some of the items at all and thus considered these items to be beyond their abilities, they rated themselves higher on the ones that they felt they could do. Promoting more accurate self-assessment when learners overestimate their abilities can be accomplished by providing prompt feedback (Lichtenstein & Fischoff, 1980) and getting learners to think of the reasons why their estimates might be wrong (Koriat, Lichtenstein, & Fischoff, 1980). People tend to be more accurate in their self-evaluations of ability when they experience counterevidence (evidence that suggests it would be difficult to achieve the task), usually as they are asked to attempt or imagine themselves attempting a specific task or as they are asked to provide the specific tools that would be necessary for a task, such as the vocabulary, the structures, the pragmatic knowledge, etc. (Koriat et al., 1980).

Regardless of the lack of a clear-cut hierarchy and the apparent inflated
estimates, the Can-Do items provided valuable data that can be informative to experiential language programs. This self-assessment data can provide an overall picture of the tasks about which individuals or groups of individuals feel most and/or least confident and can suggest some general patterns in terms of task difficulty at the various ACTFL levels. This understanding can inform second language teaching by helping instructors to know which tasks may require additional instruction, etc.

Introducing a self-assessment instrument into a program has the effect of enhancing rather than diminishing the role of the OPI. Indeed, such an approach provides formative checks for learners prior to taking an OPI/c and clear objectives that assist students when instructors articulate language-learning goals. It is noteworthy that learners showed significant gains on both measures, suggesting that self-assessment may be useful for seeing group performance in terms of gains over time. Furthermore, the self-assessments provided insights into individuals’ confidence in their abilities to perform specific tasks, which the OPI does not provide. The OPI is a holistic score based on an external evaluation, whereas the self-assessment is a collection of a range of self-estimates that can be summed up into one collective score or broken down into discrete items spread across the range of difficulty represented by the ACTFL scale.

Stansfield et al. (2010) evaluated the validity of similar Can-Do statements based on ILR descriptions similar to the Can-Do statements in this survey. Their correlations were higher than this (moderate to high effect sizes), but they used raw scores rather than logits, and the timing of their self-assessments was different. They administered the Can-Do statements in order to screen applicants and determine whether to administer the OPI. In their study of predictive validity, they concluded that applicants to National Language Service Corps programs “can make reasonably effective judgments about their own language skills,” given that “self-assessment scores ... exhibited significant positive relationships with OPI scores” (Stansfield et al., 2010, p. 313). They pointed out that their correlations were well above those found between standardized tests used for university admission and first-year GPA. They argued that an assessment that predicts performance on the criterion measure to a similar or better degree ought to be acceptable.

This study focused strictly on self-assessment administered at a single time (retrospective self-assessment following internship abroad). The literature on self-assessment suggests that it is a good tool for developing learner autonomy and self-awareness, but the aim of this study was not to test this claim. Rather, the focus was on determining the value of the aforementioned Then-Now Can-Do self-assessment for understanding student performance over the internship abroad experience. Because the correlations between self-assessments and the OPI were not high enough to use the self-assessments to predict OPI scores, it seems that additional training involving multiple self-assessments and reflections designed to make learners more aware of their abilities and to facilitate more accurate self-assessment is in order.

The current study focused strictly on university-level learners approaching or at the Advanced level or above. There are many other possible audiences for Can-Do statements, and the growing body of literature in this area suggests great potential. For example, the Can-Do statements making up the CEFR have been used widely in a range of settings and for a variety of assessment purposes. One of the original aims of the CEFR and the subsequent European Language Portfolio was to facilitate learner reflection and autonomy (Council of Europe, 2001, 2004), and this has consequently led to the use of self-assessment in portfolios and other measures (Little, 2005). Similarly, the ACTFL World Readiness Standards for Learning Languages (http://www.actfl.org/publications/all/world-readiness-standards-learning-languages) and the 21st Century Skill Map (http://www.actfl.org/sites/
to examine how self-assessment changes with beginning learners who may be studying in the classroom. Future studies ought to employ Can-Do self-assessments (perhaps modified versions that contain specific examples of tasks and are more focused) on a regular basis throughout an experience abroad. Adding ongoing assessment would allow researchers to track whether learners become more accurate in their self-assessments over time.

This study of Then-Now Can-Do self-assessment was limited to one setting (internship abroad). This same assessment approach could be used in a variety of other settings, including classroom instruction (e.g., for having students reflect and evaluate progress on various tasks over the course of instruction) in addition to internships and other out-of-class experiential learning. Doing so will allow language educators to better understand how Then-Now and other types of Can-Do self-assessments (i.e., more regular reflective self-assessments spaced over time) can be used to promote self-awareness and reflection and to better understand what learners feel they are learning during their experiences.

The primary issues that come to the forefront are learner awareness, experience, and training. Learners typically have limited experience with Advanced or higher tasks prior to experiential learning, such as study abroad and internships. It may therefore be difficult for them to make judgments regarding their own capability to perform these tasks. Language educators and program directors typically assume that learners ought to have opportunities to at least observe Advanced or higher language usage while engaged in experiential and classroom learning. If this is the case, such experience ought to inform learners and help them make more accurate judgments of their own abilities. Given that the data showed that students tended to overestimate their own abilities at these higher-level tasks, even after an internship abroad, several questions arise. Are students actually engaging in higher-level tasks that would allow them to better
understand their own abilities? If not, are they at least observing such tasks? If they are encountering Advanced or higher experiences, are they processing them sufficiently and identifying their own strengths and gaps in terms of their ability to perform these tasks well?

One additional area for further research is the role of feedback in self-assessment. Once a learner self-assesses, what evidence does he or she receive regarding the accuracy of that assessment? As Lichtenstein and Fischoff (1980) found, prompt feedback helps promote more accurate self-assessment. The 21st Century Skill Map suggests that learners use a digital self-assessment and portfolio to track their progress over time. Instructors could easily view these self-assessments and provide ongoing feedback to help learners know where they are under- or overestimating their own abilities. Furthermore, computer-based tools could provide instant feedback to learners to indicate whether their performance on objective measures (questions graded by the computer) match their own self-assessments of their performance on the same measures. Diagnostic feedback could be used to help learners know where gaps exist and therefore gain a more accurate understanding of their abilities. Students could even listen to audio recordings of their own performance after they have self-assessed and then have another opportunity to assess their own speaking performance. These examples illustrate how providing feedback (i.e., evidence regarding performance) could help learners develop their self-assessment skills. The effectiveness of various types of feedback should be carefully evaluated before drawing any conclusions.

Finally, future research could evaluate the accuracy and impact of self-assessment over time. For example, when learners begin self-assessing at the Novice level, do they become more accurate over time? Does regular self-evaluation promote increased linguistic development over the course of learning? In addition, self-assessment has been used successfully with children (Hasselgreen, 2005). Research to determine differences between adults and children in terms of the accuracy and effects of self-assessment would be helpful. Are children and adults equally accurate in their assessment of their own abilities? Do they develop similarly in their accuracy as they assess themselves over time? These and other questions remain unanswered in this research, but they certainly merit further consideration.

Notes
1. Sometimes Can-Do statements are reversed by creating statements such as “I cannot do X.” Other wording, such as “It is difficult for me to do X” may also be used. We are unaware of research comparing the effects of such wording, but it is possible that such slight changes, in particular moves away from Bandura’s (2006) suggested “can do,” could change the results and lead to measurement of a different construct.
2. Note that the standards progress from survival language skills produced with memorized material (i.e., Novice Level) through the highest skills that require rhetorical skills and cultural knowledge (Distinguished level), which only a small percentage of native speakers ever attain.
3. Statistical analyses were conducted using Winsteps software.

References


Badstübner, T., & Ecke, P. (2009). Student expectations, motivations, target language use, and perceived learning progress in a summer study abroad program in Germany. Die Unterrichtspraxis/Teaching German, 42, 41–49.


Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation,* 70, 857.

Submitted February 13, 2014

Accepted March 2, 2014
学霸图书馆
www.xuebalib.com

本文献由“学霸图书馆-文献云下载”收集自网络，仅供学习交流使用。

学霸图书馆（www.xuebalib.com）是一个“整合众多图书馆数据库资源，
提供一站式文献检索和下载服务”的24小时在线不限IP图书馆。
图书馆致力于便利、促进学习与科研，提供最强文献下载服务。

图书馆导航：
图书馆首页 文献云下载 图书馆入口 外文数据库大全 疑难文献辅助工具