2018-04-01

Performance Self-Appraisal Calibration of ESL Students on a Proficiency Reading Test

Jodi Mikolajcik Petersen
Brigham Young University

Performance Self-Appraisal Calibration of ESL Students on a Proficiency Reading Test

Jodi Mikolajcik Petersen

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Troy L. Cox, Chair
Dan D. Dewey
Jennifer Bown

Department of Linguistics and English Language

Brigham Young University

ABSTRACT

Performance Self-Appraisal Calibration of ESL Students
on a Proficiency Reading Test

Jodi Mikolajcik Petersen
Department of Linguistics and English Language, BYU
Master of Arts

Self-assessment as a placement measure or accurate assessment of skill has been scrutinized in previous research. Findings have shown a general human tendency towards overconfidence in performance (Kruger & Dunning, 1999). This study looks at performance self-appraisals in an ESL population, with participants from varying cultural backgrounds. Performance self-appraisal calibration is a measure of the relationship between an examinee's perceived skill (or confidence) and their actual skill (or ability) on a given exam item (Phakiti, 2016). Being well-calibrated is an indication that test takers know their strengths and weaknesses and thus the difference between confidence and ability is minimal, whereas poorly calibrated examinees may be oblivious to their weaknesses. While some research has explored self-appraisal calibration in first language (Hassmén & Hunt, 1994; Gutierrez & Schraw, 2015; Stankov & Lee, 2014) and foreign language contexts (Bastola, 2016; Phakiti, 2016), the language research has been limited to the performance of native language speakers on norm-referenced tests.

It still needs to be determined how test takers would perform on a criterion-referenced exam with items of differing difficulty parameters administered to examinees from different language backgrounds. To that end, a proficiency-based criterion-referenced reading comprehension test was administered to 96 ESL students with 8 different language backgrounds. To measure confidence, a pre- and post-test questionnaire was administered in addition to a confidence slider bar that was embedded into each test item. We investigated correlations between cultural background and item difficulty on the students' self-appraisal calibrations. Our results showed that ESL students were overconfident in their self-calibrations, and their overconfidence was more pronounced as item difficulty increased. There were significant differences based on native language background. Implications will be discussed.

Keywords: self-assessment, self-appraisal calibration, assessment, confidence, reading comprehension, cultural differences

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

Introduction

Benjamin Franklin (1750) said, “There are three things extremely hard: steel, a diamond, and to know one’s self.” Similarly, Charles Darwin (1871) stated, “Ignorance more frequently begets confidence than does knowledge.” Given these two insights into the human tendency to be oblivious to weaknesses, past studies have found that people are generally overconfident (Burson, 2012; Ehrlinger, Johnson, Dunning, 7 Kruger, 2008; Kruger & Dunning, 1999; Mahmood, 2016; Moore & Healy, 2008; Stankov & Lee, 2014) even in the context of education and assessment (Bastola, 2016, Brantmeier, 2005; Brantmeier, 2006; Brantmeier & Vanderplank, 2008; Hassmén & Hunt, 1994; Phakiti, 2016, Ross, 1998, Stone, 2000). This study will investigate how these previous findings hold up when ESL (English as a Second Language) students are asked to evaluate their confidence of their knowledge on a criterion-referenced reading comprehension test using performance self-appraisals. It is hypothesized that an ESL population will also show a general trend towards overconfidence, but that we will see differences in calibration among cultural groups and on items of varying difficulties. It is also theorized that confidence will decrease and test takers will become better calibrated on post-test surveys if they are allowed to appraise their confidence before, during, and after testing. A brief review of previous work in this area will be discussed below.

Literature Review

Self-assessment, self-appraisal confidence and calibration

Self-assessment refers to any involvement of students in making judgments about their work. This differs from a *performance self-appraisal*, a subset of self-assessment, which is specific to judgments made on test performance (Phakiti, 2016). *Self-appraisal confidence* has

been defined as the learner's perception of the likely outcome of his performance on a test (Bastola, 2016; Phakiti, 2016). In order to measure this, test takers are asked to self-report their confidence regarding and immediately following a response to an exam item. *Self-appraisal calibration*, then, is a measure of the relationship between examinees' perceived skill (or confidence) and their actual skill (or ability) (Phakiti, 2016). In other words, it is what a test taker thinks he knows as opposed to what he truly knows.

In a 2016 study, Phakiti made the distinction between two different types of appraisal confidence: *single-case appraisal confidence* and *relative frequency appraisal confidence*. Single-case appraisal confidence is the reported confidence on a single test item and relative frequency appraisal confidence refers to the reported confidence on the test as a whole. Single-case appraisal confidence judgments are embedded into the test and appear after every test item; they reveal what the test taker believes to have been his performance on that item. On the other hand, relative frequency appraisals are given either before or after the test and ask the test taker to assess their ability on the test overall. Here a test taker makes a judgment on the number of questions they believe to have answered correctly. When test takers make a relative-frequency appraisal, other test factors indirectly affect their judgment, such as test instructions, test environment, and time constraints; because these contribute to test anxiety and test takers may focus on these factors rather than their actual knowledge (Kleitman & Stankov, 2001).

When a test taker is considered "well-calibrated" it means they can accurately judge their ability. Being well-calibrated is an indication that test takers know their strengths and weaknesses, and thus the difference between the confidence and ability is minimal, whereas poorly calibrated, or miscalibrated examinees may be oblivious to their weaknesses. Test takers are considered "overconfident" if average confidence (percentage) minus items answered

correctly on the test (total %) is a positive value. “Underconfident” test takers receive a negative calibration score. Individuals are considered perfectly calibrated if that score is equal to zero; the closer they are to zero, the more accurate their self-appraisal calibration. Results are typically displayed in comparison with a 45º line which is often referred to as a “unity line” (see Figure 1) (Phakiti, 2016).

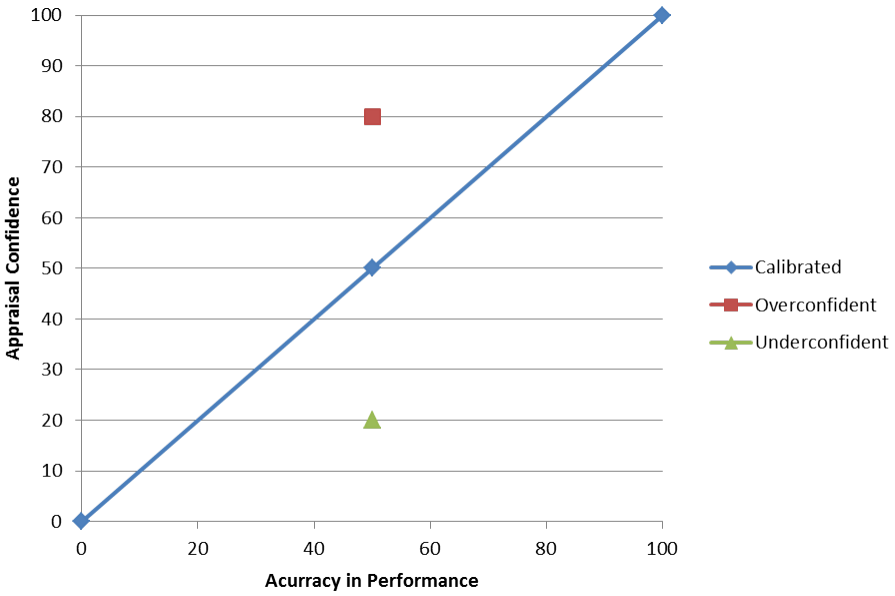

Figure 1 Calibration

While calibration can be looked at as a single instance where ability is compared with confidence, there are a few other methods of measuring a student’s calibration. Calibration could refer to improved accuracy in self-assessment over time (Boud, Lawson, & Thompson, 2013; Boud, Lawson, & Thompson, 2015). Another study looks at how student self-assessment can become calibrated by performing item response theory (IRT) and adjusting for student error (Labutov & Studer, 2016). However, for the purposes of this study, calibration will be defined as the distance between perceived and demonstrated levels of understanding and capability (Alexander, 2013).

In order to self-assess correctly, learners and test takers must be willing to recognize all aspects of their knowledge and overcome the assumption that they are above average (Sitzmann, Ely, Brown, & Bauer; 2010). This is especially difficult for individuals who lack particular cognitive and meta-cognitive skills and may suffer from the Dunning-Kruger effect. According to the authors, "People who are unskilled…suffer a dual burden: Not only do these people reach erroneous conclusions and make unfortunate choices, but their incompetence robs them of the metacognitive ability to realize it" (Kruger & Dunning, 1999, page 1121). Previous studies have found that the majority of test takers are overconfident in their appraisals in first language tests (Burson, 2012; Ehrlinger, Johnson, Dunning, & Kruger, 2008; Hassmén & Hunt, 1994; Kruger & Dunning, 1999; Mahmood, 2016; Sitzmann et al., 2010) and foreign language contexts (Bastola, 2016; Phakiti, 2016; Summers, in press).

Accuracy of self-assessment in language testing depends greatly on the skills being assessed (Blanche, 1988; Brantmeier, Vanderplank, & Strube, 2012). Many studies have found that individuals are more accurate when self-assessing their reading skills than when assessing other skills (Brantmeier, 2006; Brantmeier & Vanderplank, 2008; Brantmeier & Vanderplank, & Strube, 2012; LeBlanc & Painchard, 1985; Ross, 1998; Stankov & Lee, 2008; Wan-a-rom, 2010). Ross (1998) composed a meta-analysis of 60 studies which compared self-assessment of the four skill areas and found that language learners rated themselves lowest in speaking and highest in reading. He asserts that this is because learners in a foreign language context have greater exposure to reading, especially via technology, and reading is a skill that is usually developed prior to listening and speaking (Ross, 1998). In most of the previous second-language studies performed regarding self-assessment and reading, the subjects were taken from universities where they have had extensive exposure to reading, as well. Ross (1998) argues that

self-assessment of reading is more valid than self-assessment of the other language skills; this may be especially true when performed prior to the reading task (Brantmeier & Vanderplank, 2008).

Item Difficulty and Proficiency Testing

As participants are generally unaware of the performance of their peers, it is task difficulty that drives perception (Burson, 2012).Test takers use this perception of item difficulty in order to assess their performance. Therefore, the difficulty of the item greatly determines the ability of the participant to make well-calibrated self-appraisal judgments (Stankov & Lee, 2014). Ideally, on a criterion-referenced test, the test takers should be most confident on items whose difficulty aligns with or is below their proficiency level. As difficulty increases, confidence should decrease if the test taker is aware that the item is testing beyond their proficiency level. However, overconfidence tends to be greater on items whose difficulty is greater than the proficiency level of the test taker, which is referred to as the *hard/easy effect* (Stankov & Lee, 2014). In addition, test takers will often underestimate their performance when the task is easy or when their own proficiency is great. This is because there is always an error component in judgment and "it is easier to underestimate than overestimate your score on a test when you get everything right. As a result, people underestimate their performance when it is high" (Moore & Healy, 2008, p. 9).

Proficiency also influences self-appraisal calibration in reading. Ross (1998) noted that beginning students had a tendency toward overconfidence and advanced students were usually underconfident. In an EFL context, Bastola (2016) found that low to moderate performers overstated their performance and thus their performance appraisals were not calibrated with actual performance. In other words, low level test takers thought they knew more than they

actually did. High performers in Bastola's study were much better calibrated. Brantmeier et al. have performed several studies with learners of Spanish (2006, 2008, 2012). They also have found that low performers were highly inaccurate in their self-assessments. Beginners may not yet fully understand what it means to be considered an "excellent" reader and thus exaggerate their identification as one. It is only for advanced learners of the language that self-assessment can become an accurate predictor of performance. Advanced students are better at assessing their skills. As Kruger and Dunning (1999) have noted, "…one way to make people recognize their incompetence is to make them competent" (p. 1131).

Not only are item difficulty and proficiency factors in self-appraisal calibration, but also the type of assessment being given. A main characteristic of self-assessment is the involvement of students in making judgments about their work and to what extent it matches standards or criteria. Therefore, students must be able to understand and identify the criteria that apply and make judgments based on these criteria (Wan-a-rom, 2010).

A proficiency test is a criterion-referenced test based on ACTFL guidelines, and, on a proficiency test, test takers are compared to a set of standards as opposed to their peers. Most language tests measuring the receptive skills of reading and listening have been norm-referenced in that they were designed to compare test takers to each other (Clifford, 2016). Thus the items on these norm-referenced tests (NRTs) are all a similar difficulty level. Making norm-referenced comparisons is much more natural for test takers. When they compare themselves to others, it is easier to see how they rank compared to their peers than to compare themselves to external criteria, which they may not fully understand, as is done in criterion-referenced testing. On a criterion-referenced test (CRT), items are based on different criteria and therefore each test items can be linked to an intended difficulty level. These are items are meant to relate to language use

in real-world situations and success on a particular level requires sustained performance at that level (Clifford, 2016).

Thus, the item difficulty largely relies upon the criteria it is meant to represent and if it not fully understood by the test takers, they are more likely to misinterpret the difficulty of the task. Summers (in press) conducted a study where students made self-assessments based on ACTFL can-do statements for speaking and writing and their perceived ability to complete the tasks described in the statements. Sixty-one percent of participants rated themselves as Superior in speaking and 48% rated themselves Superior in writing, when, in reality, none of them achieved that level on their corresponding placement tests. It was theorized that students may not fully understand the criteria that they are measuring themselves against and that the use of more specific can-do statements would enable test takers to become more accurate in their self-appraisals. Additionally, Summer's study showed that can-do statements as predictors of placement might not function as expected because students could be thinking of a single incident of success rather than sustained performance over time. Similarly, Burson (2012) found that participants were, in fact, better calibrated on the assessment of relatable tasks (i.e. juggling) that they deemed difficult. This portrays the significance of understanding the criteria by which one is being measured to make accurate judgements. Therefore, research ought to look at how test takers appraise their confidence on items in a criterion-referenced test.

Cultural and L1 Backgrounds in Self-appraisals

There is much conjecture that some groups are better calibrated than others. For example, females were shown to be somewhat better calibrated than males on assessments testing cognitive, mathematical, and verbal ability (Hassmén & Hunt, 1994; Pallier et al., 2002). Another study found that 15-year-old students from Singapore tend to be well-calibrated on

items testing mathematical ability (Morony et al., 2012). However, as of 2014, relatively little was known about cultural differences in overconfidence. In a study performed by Stankov and Lee (2014), participants were divided into 9 world regions and asked to rate their confidence as they answered items on a measure of fluid intelligence, The Five-item Number Series Test in their native language. It was found that East Asian participants had the highest confidence, but also the highest performance out of the regions. Though all groups were overconfident in their self-appraisals overall, those of Anglo origin were the closest to being well-calibrated. South East Asians scored the lowest on accuracy, but had confidence scores similar to the East Asian region. Overconfidence was more pronounced in the lower-scoring regions, including South East Asia, all regions of Africa, Latin America. European, Anglo, and East Asians were the top scorers in this study and their overconfidence was less pronounced. Stankov & Lee (2014) determined that cultural differences in confidence exist, but to a smaller extent than anticipated. It is yet to be tested whether these results would stand in a study designed to assess confidence on a test in the subject's second language.

Better calibration in East Asians may be connected to high competition in the educational systems (Stankov & Lee, 2014). Overconfidence may also be linked to other survival mechanisms and a preservation of self-esteem. When people perceive their abilities as better than they actually are, they may be cushioning the blow of failure and buoying up their self-esteem to continue functioning as effective members of society (Stankov & Lee, 2014). Therefore, it is not too far-reaching to consider that countries and individuals with lower cognitive ability exhibit overconfidence to protect the individual's well-being and dignity (Stankov & Lee, 2014). Using an ESL context would allow us to compare different cultural groups in their performance self-appraisal calibration.

Timing of self-appraisals

The timing of self-appraisals, whether made before, during, or after answering questions, also affects the accuracy with which test takers appraise themselves. Hassmén and Hunt (1994) used single case self-appraisals in a multiple-choice test similar to the SAT and found that self-assessments made immediately after selecting a response on a test item were more accurate than asking test takers to self-assess prior to responding. Furthermore, when students are allowed to self-assess immediately following a response, they are accessing usable knowledge, which is a combination of a person's knowledge and an assessment of their knowledge. This knowledge is then used to make decisions and solve problems.

Further findings have indicated that the use of a descriptive and criterion-referenced questionnaire as a pre-test self-assessment was a reliable predictor of performance on computer-based and classroom-based testing (Brantmeier & Vanderplank, 2008). This questionnaire provides the self-assessor with more detailed examples of the criteria being measured and what the individual is expected to do with the language. This instrument, as it becomes more extensively validated, could be an important tool for advanced placement, but it is still most effective with advanced language learners, usually those entering at the university level. Brantmeier and Vanderplank (2008) additionally posited that the use of criterion-referenced items on such a test would improve self-appraisal calibration. As both pre-test criterion-referenced self-assessment and single case self-appraisals have been validated separately, it stands to reason that an instrument that combined these two methods would be the most reliable form of self-assessment. This type of instrument has yet to be studied to any great extent.

Research Questions

Due to a general lack in studies available involving participants in an English as a second language context, this study seeks to examine a few facets of self-appraisal calibration that have been performed in other contexts, mainly EFL and first language, to see if the results hold true among this population. The questions we pose in this study are:

(1) At which ACTFL reading passage and question levels are test takers better calibrated?
(2) What is the relationship between L1 background and the tendency to be overconfident?
(3) How do ESL students' pre-test confidence level, mean confidence level, and post-test confidence level compare with their actual score?

The ESL participants will allow us a unique perspective regarding how culture plays into calibration, as will be addressed in research question #2.

Method

Setting and Participants

The research instrument was administered in an intensive English program (IEP) in the western United States where English as a Second Language (ESL) is taught to students seeking to improve their English ability in a non-credit seeking program. Participants were new students admitted to the IEP for the summer semester of whom 96 out of 99 gave consent to have their data used in this study. The participants' age ranged from 17 to 63 years old (M = 26.4, SD = 9.3), with 53% male and 47% female. The students' proficiency encompassed ACTFL proficiency levels Novice to Advanced. Previous experience with English language study varied between participants and was not recorded as part of this investigation. Students came from a variety of countries and L1 backgrounds as portrayed in Figure 2.

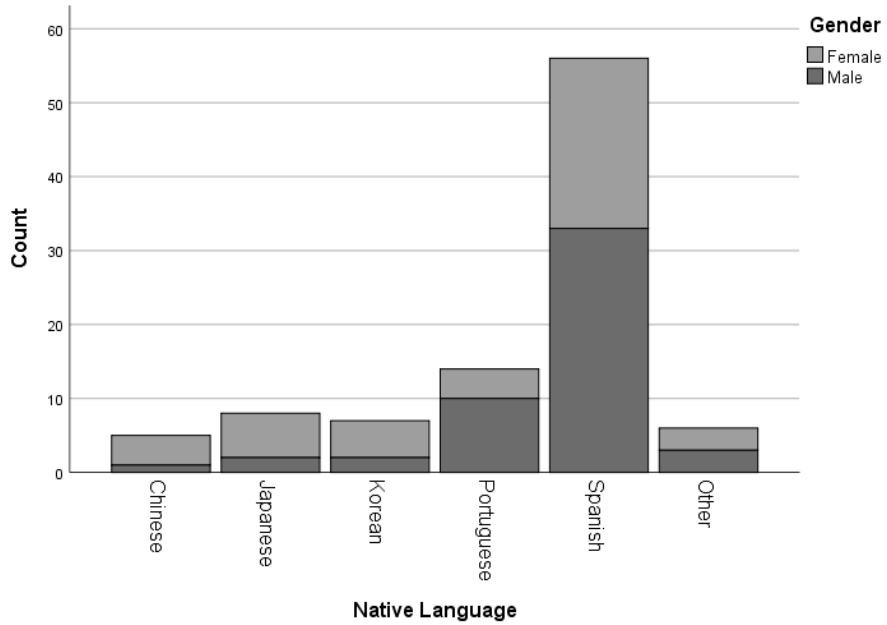

Figure 2 Language background and gender of participants

Research Instrument

This section will discuss the instruments created and used in order to address the research questions, 1) a can-do statement survey presented before and after the test and 2) an English Reading Test with an accompanying appraisal confidence slider bar.

Can-Do Statement Survey

Immediately before the English Reading Test, a *Can-Do Statement Survey,* based on the ACTFL "can-do" statements was administered as a pre- and post-test self-assessment measure. These statements were adapted from the ACTFL Can-Do Reading section (2015) and the 13 statements spanned Novice-Mid to Superior levels and were presented as "can-do" statements where each statement began with the phrase "*I can…"* followed by the task (see Appendix). Students were asked to rate how confident they were that each statement was an accurate description of their ability on a scale of 0-100.

In order to facilitate accurate self-assessment of novice learners, the can-do statement survey was translated into the top five L1 populations anticipated for summer semester admission to the IEP: Spanish, Portuguese, Chinese, Japanese, and Korean. The translations were offered to each student because the target skill area was reading and the can-do statements themselves were not meant to be a test of their reading ability, but rather an opportunity for the students to make confidence judgments. However, not all novice learners were able to benefit from the translation if their L1 was not one of the top five languages. There was only one instance where this was the case, but as a result, the students who consisted of the "Other" language group were not compared to the other L1 groups when answering research question 2.

The survey was presented before the test and the same survey was again given immediately following the test. The purpose of the reproduction of the survey was to measure any changes in confidence expressed by the students before and after they had taken the test. The survey presented post-test had one additional question: an optional, open-ended qualitative response where students could indicate if their survey question responses changed after taking the test and why.

English Reading Test

The English Reading Test was a criterion-referenced test (CRT) that assessed reading comprehension and was comprised of 30 items. Each item had a reading passage and a single accompanying comprehension question. The reading passages encompassed three different ACTFL proficiency levels: the first 15 items were Intermediate, the following ten were Advanced, and the final five items were Superior. The test was created using a database of copyrighted and validated items from a previous study (Clifford & Cox, 2012).

Appraisal Confidence Slider Bar

Following every reading passage and comprehension question, there was an embedded self-appraisal slider bar. The question preceding the slider bar stated, "How confident are you in your answer choice?" The slider bar ranged from 0 to 100 with the labels, *very unconfident, unconfident, somewhat unconfident, somewhat confident, confident,* and *very confident* spaced evenly across the slider bar (see Figure 3). The cursor was always first presented as set in the middle of the bar (at a value equal to 50) and participants were not able to progress to the next question without first answering the self-appraisal confidence question. As a result, 50 was not a valid answer choice. This proved helpful as many of the students did not notice the slider bar at first due to its low placement on the screen and the upper placement of the submit button.

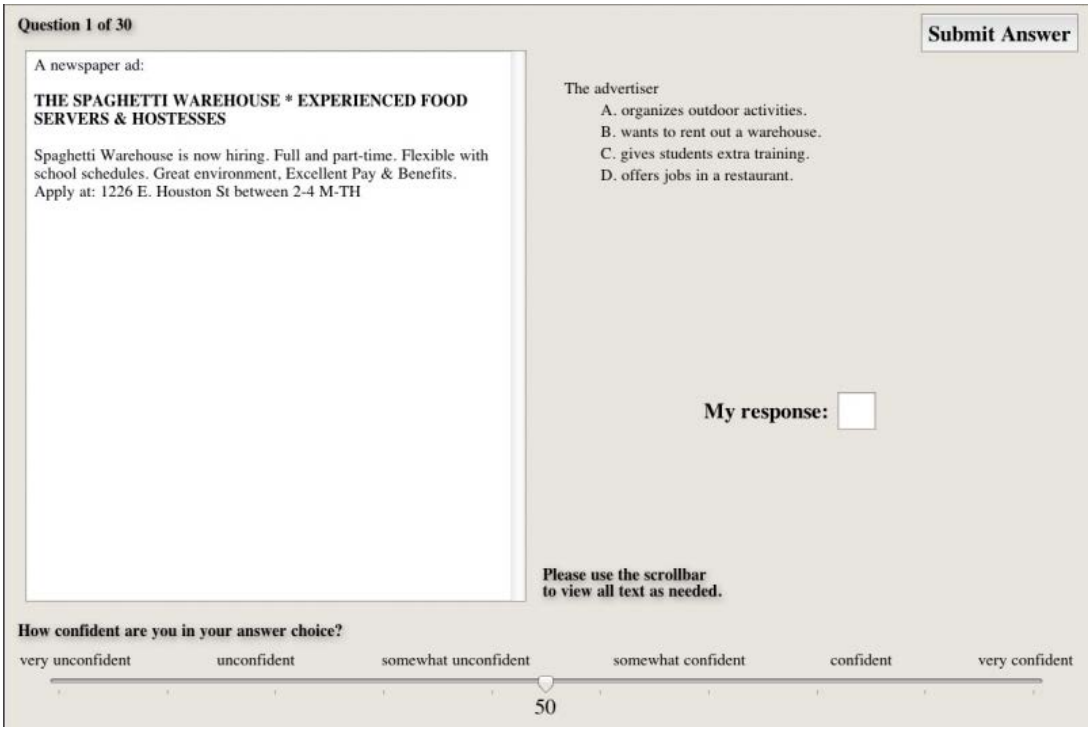

Figure 3 Example of Confidence Slider Bar Placement

Test Administration

The English Reading Test instrument developed was added to the initial placement battery for new students admitted to the ESL program. The test was administered in a climate controlled computer lab which participants were familiar with, as they had previously completed a portion of their placement testing at that location. The English Reading Test was administered on the final day of placement testing. Prior to the English Reading Test, students were scheduled to participate in an oral interview at different times. As a result, students entered the testing environment individually and began and finished the test at different times. Due to this, it was not possible to give proctoring instructions to the whole group. Rather, as each participant entered, they were given brief instructions regarding the nature of the test and were offered a translation of the survey portion of the test. The researcher and computer lab assistants were available to answer questions and resolve any problems students encountered.

On average, participants took 75 minutes to complete the test with the accompanying surveys. The test was administered via in-house testing software and retrieved electronically after the completion of the test. In addition to answering the comprehension questions and self-appraisal calibration survey questions, participants were asked to provide basic personal information, such as age, gender, country of origin, and first language.

Data Analysis

To answer the research questions, the calibration (difference between appraisal confidence and actual score) was calculated. If the difference between appraisal confidence and actual score was a positive value, the participant would be considered overconfident, and when the difference is negative, the participant is underconfident. For example, if the student the student had a confidence score of 90 and a test score of 70, their calibration score would be 20. A

series of statistical analyses were then used to answer the research questions. To investigate the first question: whether test takers are more calibrated with items whose difficulty levels align with student ability level, a one-way repeated measures ANOVA was used. For the second question that addressed the relationship between L1 background and tendencies to be over or underconfident, a one-way ANOVA was used with the dependent variable being calibration and the independent variable, the L1 of the participant. Finally, to answer the third question the timing of the self-assessment, three repeated measures paired t tests were used (pre-test confidence level, mean confidence level during the test, and post-test confidence level compared with their actual score). Confidence scores were aggregated in order to compare the different difficulty levels and timing measures, as well as to portray overall tendencies of each L1 background.

Results

Research question 1: At which ACTFL reading passage and question levels are test takers better calibrated?

As student proficiency was estimated to be in the Novice to Intermediate levels, it was expected that test takers would be best calibrated at the intermediate item difficulty level than at item difficulty levels that were beyond their ability. In looking at the test taker calibration, the three major item difficulty levels were compared: intermediate, advanced, and superior. The descriptive statistics are displayed in Table 1. A one-way ANOVA was used to compare the ability of test takers with their calibration at the different item difficulty levels, with calibration operating as a dependent variable and item difficulty level as the independent variable. There was a significant effect on calibration, [F (2, 190) = 38.347 and $p < .00$] as portrayed in Figure 4.

Table 1

Descriptive Statistics of Test Items at Three Proficiency Levels

Item Difficulty	*M* (*SD*)	N	95% *CI* Lower Bound	95% *CI* Upper Bound
Intermediate	1.59 (19.14)	15	-2.29	5.47
Advanced	19.36 (20.51)	10	15.20	23.52
Superior	23.88 (29.91)	5	17.82	29.94

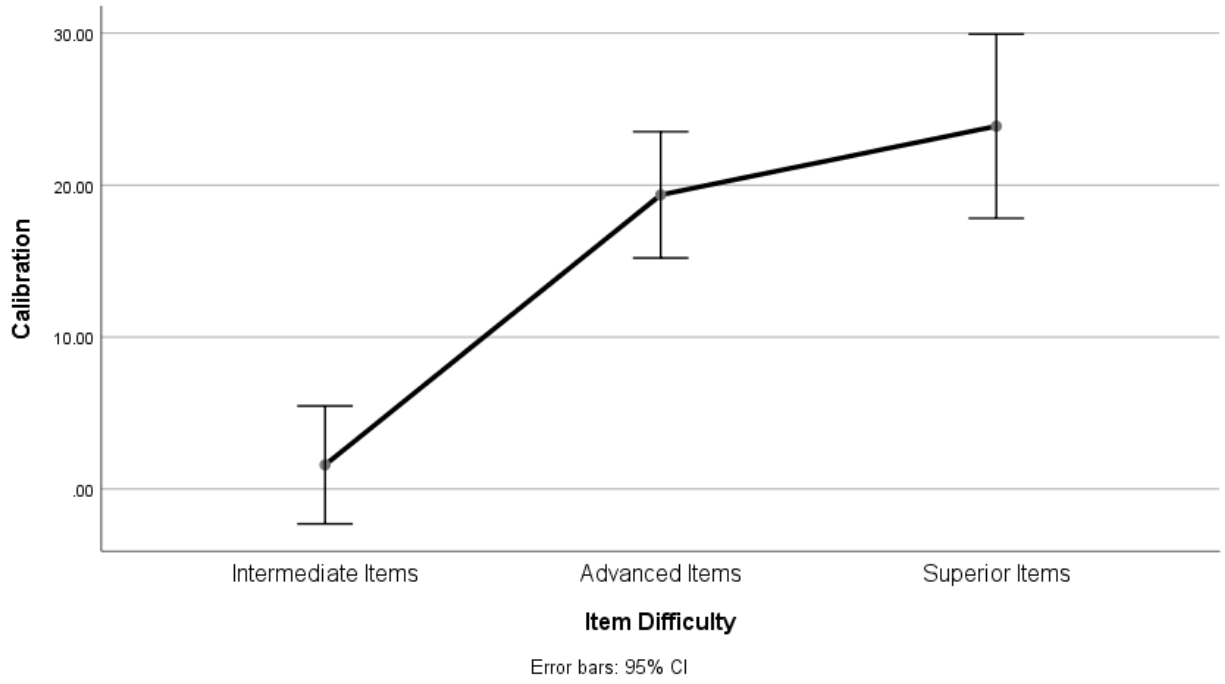

Figure 4 One-way ANOVA comparing calibration with item difficulty

Comparisons using Bonferroni's contrasts found statistically significant differences between the intermediate and advanced items (mean difference = 17.77 , 95% *CI(confidence interval)* = [12.86, 22.69] , $p < .001$, Cohen's $d = 0.90$) and the intermediate and superior items (mean difference = 22.29, 95% *CI* = [14.89, 29.68], $p < .001$, Cohen's $d = 0.89$), showing that the intended item difficulty level had a large effect on how calibrated test takers were at the intermediate level. There was no statistically significant difference between calibration at the advanced and superior item level (mean difference = 4.52, 95% *CI*= [2.57, 11.60, $p = .37$,

Cohen's $d = 0.18$). Students were more accurate at their ability level (Intermediate) than they were on items that were above their level.

Research question 2: What is the relationship between L1 background and tendencies to be over or underconfident?

Since there is a scarcity of research regarding cultural differences and self-appraisal calibration, researchers adopted the null hypothesis that L1 background would not have a significant effect on confidence, but that all test takers would be generally overconfident. A one-way between-subjects ANOVA was conducted to compare the effect of L1 background on confidence. There was a significant effect of language background on confidence [$F = 3.90$ (5, 90), $p = 0.003$]. Descriptive statistics are displayed in Table 2. Using Bonferroni's contrast as the post-hoc we found that the only statistically significant difference was between speakers of Spanish and Japanese (mean difference = 24.54 with 95% CI between 2.33 and 46.74 and p <0.02) with a medium effect size of $d = 1.65$.

Table 2

Descriptive Statistics of Test Taker Confidence by L1 Background

L1 Background	*N*	*M* (*SD*)	Std. Error	95% CI	
				Lower Bound	Upper Bound
Spanish	56	78.96 (17.08)	2.28	74.38	83.53
Korean	7	71.54 (11.49)	4.34	60.92	82.17
Portuguese	14	64.51 (24.74)	6.61	50.23	78.79
Chinese	5	59.95 (30.92)	13.83	21.56	98.34
Japanese	8	54.42 (12.26)	4.33	44.17	64.66

However, being confident does not necessarily signify miscalibration. Calibration is the comparison of confidence with ability. A test taker may be confident because they are able and

still be considered well-calibrated since their confidence aligns with their ability. Therefore, it was necessary to also examine the effect of language background on calibration.

Descriptive statistics of participants calibration grouped by language background are displayed in Table 3. The only group that was underconfident in their ability comprised speakers of Japanese with a mean calibration of -3.50, whereas Spanish speakers were the most overconfident with a calibration score of 16.46. This information is depicted in Figure 5.

Table 3

Descriptive Statistics of Calibration Among Different L1 Backgrounds

				95% *CI*	
L1 Background	*N*	*M* (*SD*)	Std. Error	Lower Bound	Upper Bound
Spanish	56	16.46 (11.92)	1.59	13.26	19.65
Korean	7	13.45 (20.68)	7.82	-5.68	32.57
Chinese	5	9.95 (28.78)	12.87	-25.79	45.68
Portuguese	14	3.55 (18.48)	4.94	-7.11	14.22
Japanese	8	-3.50 (10.65)	3.77	-12.41	5.41

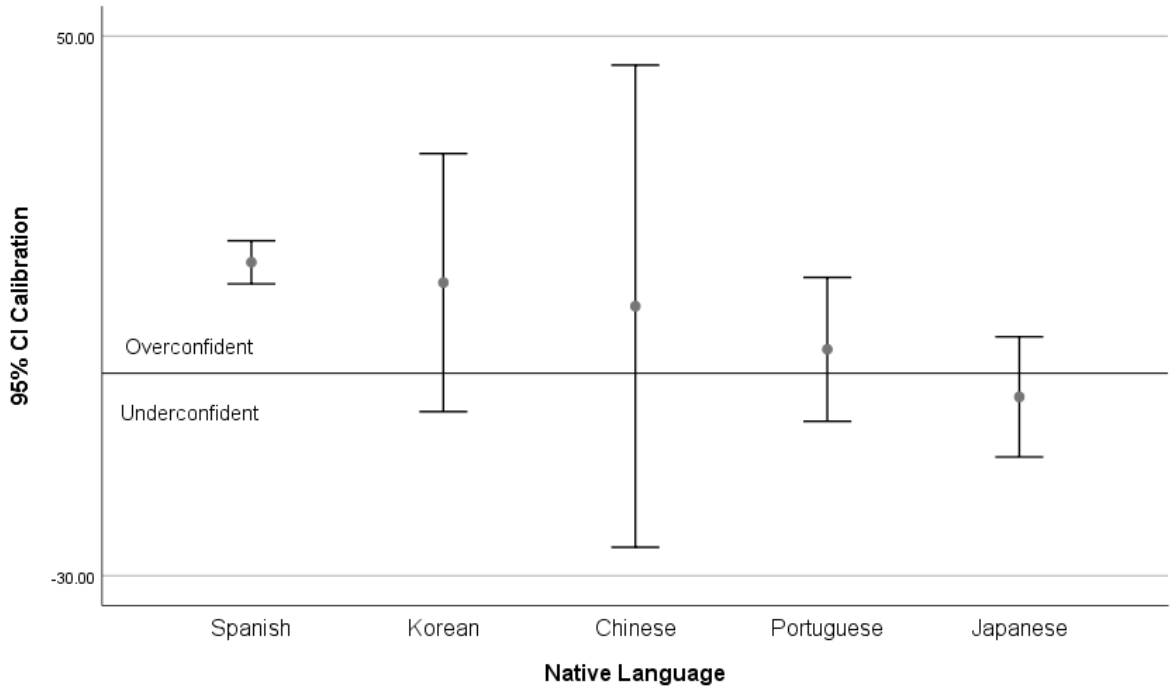

Figure 5 One-way ANOVA comparing calibration with L1 Backgrounds

Another one-way between-subjects ANOVA was conducted to compare the effect of L1 background on test taker calibration and was found to be statistically significant, but with low variability [$F = 4.01$ (5, 90), $p = 0.002$]. Using Bonferroni's contrasts as a post-hoc analysis, we found that, consistent with the results found on confidence, the only statistically significant differences the effect of L1 background on calibration was between Japanese and Spanish speakers (mean difference = 19.96 with 95% *CI* between 1.90 and 38.01 and $p < 0.02$), where Japanese speakers were underconfident and Spanish speakers were overconfident. There was a medium effect size of $d = 1.77$.

Research question 3: How do ESL students' pre-test confidence level, mean confidence level, and post-test confidence level compare with their actual score?

Three paired sample *t-tests* were conducted to compare test taker ability with confidence levels as it was measured pre-test, during the test, and post-test. Paired samples statistics are displayed in Table 4 and this information is displayed graphically in Figure 6.

Table 4

Paired Samples Statistics: Ability vs. Confidence at Various Times

	M (*SD*)	*N*	Std. Error Mean
Ability	60.90 (15.37)	96	1.57
Pre-test	65.33 (18.57)	96	1.90
During (Item Level)	72.16 (20.91)	96	2.13
Post-test	66.20 (18.90)	96	1.93

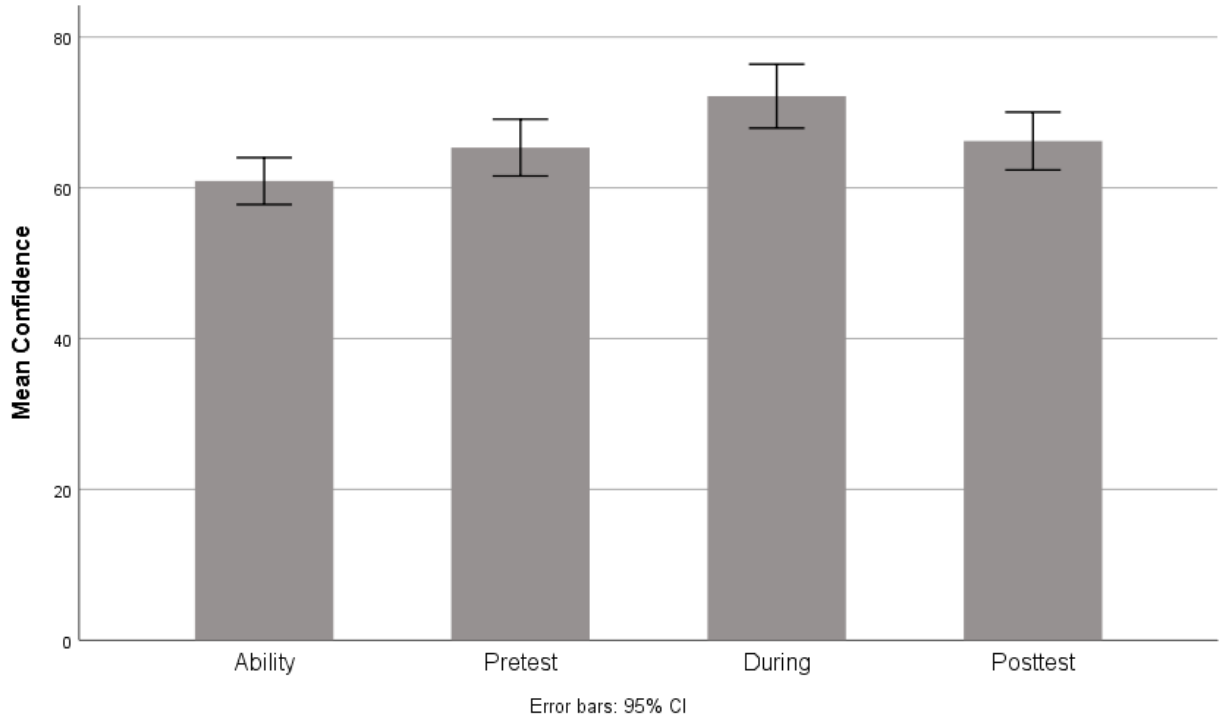

Figure 6 Mean ability and confidence pre-test, during the test, and post-test

There was no significant difference between ability and confidence on the pre-test survey (p = .089), showing that pretest confidence most closely aligned with ability. There was a significant difference in the scores for confidence during the test (M = 72.16 SD = 20.91) and ability (M = 60.90, SD = 15.37); t = 6.47, p < .001. The effect size was medium, d = 0.61. There was also a significant difference in the scores for confidence post-test (M = 66.20, SD = 18.91) and ability (M = 60.90, SD = 15.37); t = 2.04, p = .04. This effect size was small, d = 0.31.

It was also of interest to the researchers to see how confidence changed over time (pre, during, and post) and if these changes were statistically significant, thus another set of paired samples t tests was conducted. The only groups that showed statistically significant differences were the pre-test confidence totals and the during-test confidence levles (t = 2. 21, p = .03).

Researchers expected to see changes over time, with participants becoming less confident throughout the course of the test as a reflection of their becoming more aware of deficits in their knowledge as they encountered tasks that reflected the can-do statements in the pre- and post-test

surveys. Though the data did not reflect this change, test takers were given the opportunity to respond to an optional question at the end of the post-test survey: "If your answers are different than before, explain why you changed them". Out of the 96 participants, 50% chose to respond. 3 participants (6.3% of respondents) reported that their survey responses at the end of the test were similar to their responses at the beginning of the test or that they did not remember. An example of this type of response was "I do not remember my answers before the test, but those are similar now". The remaining responses where students indicated a change in their pre-test and post-test survey responses have been grouped into categories with example responses as portrayed in Table 5. The majority of test takers felt like they had changed their answers significantly enough to comment on that change, with over half claiming that their confidence had decreased upon completing the test. Only 11 participants mentioned feeling more confident in their ability post-test.

Table 5

Student Survey Responses of Reported Change Post-test

Categories	N	%	Student Response Examples Verbatim
Reported Confidence Decreased	7	14.6%	"um....maybe I had over confidence before the test…I'll have to study harder than yesterday" "I changed some answers because sometimes you think you know it, but when you see it (articles, etc.), you realize that you still have room to grow and learn more"
Reported Confidence Increased	11	22.9%	"Because placing myself in real situation [*sic*] makes me realize how easy is for me to understand texts according to the topic" "Somethings [*sic*] that I saw in the text, make me sure about me [*sic*] skills reading in English."
Identified Areas for Improvement	24	50%	"After the test, I understand that I need to learn more words and improve my vocabulary." "My answers was changed [*sic*] because I figure out that I need help to improve my skills, and I need to become more academic. I didn't recognize a lot of words and subjects."
Reported Confidence Different	3	6.3%	"After I take a test, I got what the questions really are. And then I know more what level I am."

Discussion

In answer to the first research question regarding the effect of item difficulty level on calibration, we found that test takers were best calibrated on intermediate items and significantly overconfident on the advanced and superior item levels. This finding aligned with expectations set forth by the literature. Because students' proficiency was generally in the novice to intermediate ACTFL level range, the test takers were more closely calibrated on the intermediate items because those items aligned with their ability. In keeping with previous findings, test takers were significantly overconfident on items above their proficiency level, which has been referred to as the hard/easy effect (Stankov & Lee, 2014). In this regard, ESL test takers are not unique and are generally overconfident when they really should not be, as was hypothesized.

For future research, there would be value in letting students see their scores in comparison with their confidence for each test item. By showing them a personal report, test takers would be able to know on which test items they were miscalibrated. This would give us a better perspective on how confidence compared to ability on the individual level and also look for trends in overconfidence with particular items. Perhaps future research can also look at how ESL test takers' self-appraisals change after metacognitive strategy instruction. A similar test could be administered to compare differences in student self-appraisals before and after strategy instruction. In addition to comparing differences between individual students before and after strategy training sessions, research could also look at differences between students who receive training and those who do not.

The second research question addressed differences in confidence among L1 backgrounds. There is still much that is unanswered in regards to cultural differences and self-assessment. What can be ascertained from this study is that speakers of Spanish had a tendency

towards overconfidence while Japanese speakers had a tendency towards underconfidence and were closest to being perfectly calibrated. Since the English-language based studies examining confidence were primarily performed in the U.S. and the findings showed tendencies toward overconfidence, it stands to reason that the speakers of other Western languages were also overconfident. Speakers of Korean and Chinese were also overconfident, differing from their Japanese neighbors, who were the only group in this study to exhibit underconfidence. This finding with Japanese students is supported by Stankov and Lee's (2014) suggestion that lower confidence may be due to high competition in the educational system. Perhaps, Japanese speakers in particular differ from their other East Asian neighbors since Korean and Chinese speakers were overconfident in this study. However, our sample size was small and may not reflect this language background's true tendencies. It may be, given a large enough sample, that Japanese speakers may tend towards overconfidence overall much like the other language backgrounds.

Ultimately, more research will be required to make any definite claims about particular tendencies of cultures concerning calibration on reading comprehension tests. This study could be replicated, but with a much larger sample size and controlling for language background. Alternatively, this test could still be used as part of the initial placement battery over several semesters, in order to gain this larger sample size and get a better understanding of how different cultures appraise their confidence and ability.

The final question this study sought to answer had to do with the timing of the measurement of confidence. Responses to the pre-test survey were closest to student ability; however, this was not a statistically significant difference. Confidence was greatest at the item level during the test and slightly elevated on the post-test survey responses. A gradual decrease

in confidence during the testing process was hypothesized, operating under the assumption that each test item would help students reassess their knowledge and recognize any deficits. The post-test survey was meant to be a way for students to reexamine their responses on the can-do statements, having now attempted the tasks they assessed themselves on. However, post-test answers were practically the same as the pre-test ones, despite the students' open-ended responses that indicated a change. However, with small effect sizes, no real conclusions can be made about when test takers will be better calibrated in their confidence judgments. It seems, as indicated by the literature, that students do not fully understand the criteria they are using to measure themselves. It is possible that the ACTFL can-do statements used in this study are too general and, for that reason, students were overgenerous in their judgments. It would be beneficial in future research to provide more explicit examples with their accompanying can-do statements so students would be able to see the task that corresponds with the statement before rating their confidence. This might, then, improve ESL test takers' performance self-appraisal calibration.

Implications in ESL instruction and testing

It has been established that the opportunity to make self-assessments, in general, benefits learners (Geeslin, 2003; Hassmén & Hunt, 1994; Leach, 2012; Oscarson, 1989). Through self-assessment, learners can become trained in evaluation, have a raised level of awareness, gain control in a learning environment, and continue learning after a course is finished (Oscarson, 1989). Students develop metacognitive skills and learner autonomy (Brantmeier, Vanderplank & Strube, 2012), and being well calibrated on a self-assessment tool is predictive of future learning success (Phakiti, 2016).

Therefore, by embedding the option to self-assess at the item level – as was done in this study – test developers and educators can gain insights into how self-appraisals can be beneficial in teaching test-taking strategy use. Because students who are well-calibrated at assessing their performance success use better test-taking strategies (Stone, 2000), training students in calibration would help them perform better on exams.

Performance self-appraisals also have implications on high stakes tests that discourage guessing and where points are deducted for incorrect responses, but not for leaving the item unanswered. Hássmen and Hunt (1994) found that overall performance improved and that students were less likely to leave items unanswered when they were given the opportunity to make performance self-appraisals after each test question. They also found that giving students the option to self-assess makes the multiple-choice test more accurate in measuring the usable knowledge of the test taker (Hassmén & Hunt, 1994). ESL students may benefit from training in this particular self-assessment measure as it could help them evaluate their responses on high stakes tests, such as the TOEFL.

Although accurate self-assessment still remains a difficult task (Kruger & Dunning, 1999), if teachers train students to pause and evaluate their confidence and knowledge on the item level, students will be able to better implement test-taking strategies. For example, if students are able to recognize that they do not have the knowledge required by a particular item, they can engage strategies to help them eliminate distractors. On the other hand, if students gauge that their confidence is high, they can feel certain about moving on to the next test item and not wasting time on one that does not warrant their additional attention, thus allowing them to focus on areas where they are weaker. Therefore, ESL instructors should seriously consider using performance self-appraisals on multiple-choice assessments even in classroom testing.

Teachers can promote awareness by giving students the opportunity to make self-appraisals often and regularly. If training in self-appraisals is done frequently, students may be able to better understand specific criteria and become better calibrated (Brantmeier, 2005). In fact, students seem to enjoy being given the responsibility to self-assess and have indicated its usefulness when it is built into a course (Brantmeier, Vanderplank & Strube, 2012). Self-assessment of reading ability has been found to be positively correlated with enjoyment in that skill area, and, as enjoyment increases, reading comprehension performance improves (Brantmeier, 2005). By building self-assessment into a reading course, this may help produce lifelong L2 readers because students are reflecting on strategies and abilities often.

Limitations

Though it is effective in answering the questions posed by the researchers, this study is not without its limitations. The most obvious limitation is the sample size of the different language groups. Given that this test was developed to be administered at the time of the initial placement battery, participants were limited to the new, incoming students. The semester that it was administered had a particularly large number of Spanish speaking students and smaller numbers of the other language groups typically found in this IEP. Because of this restriction, it was difficult to see significant differences between the language groups. In this study, the Japanese test takers were, on the whole, underconfident, but that may not hold true if the sample size were larger.

There were also limitations with how the test was administered. Again, due to the nature of the initial placement battery, test takers entered the testing area one by one upon completing another portion of the placement battery. As a result, they were only given brief instructions regarding how to indicate their confidence, which became confusing for some students who had

never encountered a survey question like it before. Translations were also only available to the five major language groups. Those included in the "Other" group, had to be aided in English when and if they had questions. Though this was not a concern at the time because very few students requested a translation at all, the reliability of the results found for the "Other" group should be evaluated.

There was also no time limit assigned to this test and students may have performed differently had they been given a time limit. For the purposes of this study, we did not want to introduce the variable of time as a possible distraction from the intended purpose – to measure confidence on the item level- or as a potential source of anxiety. However, it can be argued that students may have performed better on items whose difficulty was greater than the students' skill because they could spend as much time as they wanted in answering, and given an infinite amount of time, the probability of answering an item correctly increases. In retrospect, post-test calibration may have improved if students were allowed to see their pre-test survey answers as they responded to those same items post-test. Many students indicated that they felt less confident and indicated some areas of weakness on the optional post-test survey item regarding changes they made to their survey responses, but this was not reflected in the data. This discrepancy may have been avoided had the students been able to compare their responses.

As mentioned earlier, confidence scores were averaged across sections and the test as a whole in order to answer the research questions. However, anytime data are simplified, such as collapsing the confidence data to an average score, information is lost. The data presented in this study do not necessarily reflect the confidence and calibration of the individual test taker.

References

Alexander, P. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction 24*(1), 1-3. doi: 10.1016/j.learninstruc.2012.10.003

Bastola, M. N. (2016). Do learners know "what they know" in EFL reading? *Journal of NELTA*, *21*(1–2), 61–73.

Blanche, P. (1988). Self-assessment of foreign language skills: Implications for teachers and researchers. *RELC Journal*, *19*(1), 75–93. https://doi.org/10.12968/sece.2012.9.141

Boud, D., Lawson, R., & Thompson, D. (2013). Does student engagement in self-assessment calibrate their judgment over time? *Assessment and Evaluation in Higher Education, 38*(8), 941-956.

Boud, D., Lawson, R., & Thompson, D. (2015). The calibration of student judgment through self-assessment: Disruptive effects of assessment patterns. *Higher Education Research and Development, 34*(1), 45-59. doi: 10.1080/07294360.2014.934328

Brantmeier, C. (2005). Nonlinguistic variables in advanced second language reading: Learners' self-assessment and enjoyment. *Foreign Language Annals*, *38*(4), 494–504. https://doi.org/10.1111/j.1944-9720.2005.tb02516.x

Brantmeier, C. (2006). Advanced L2 learners and reading placement: Self-assessment, CBT, and subsequent performance. *System*, *34*(1), 15–35. https://doi.org/10.1016/j.system.2005.08.004

Brantmeier, C., & Vanderplank, R. (2008). Descriptive and criterion-referenced self-assessment with L2 readers. *System*, *36*(3), 456–477. https://doi.org/10.1016/j.system.2008.03.001

Brantmeier, C., Vanderplank, R., & Strube, M. (2012). What about me? Individual self-assessment by skill and level of language instruction. *System*, *40*(1), 144–160. https://doi.org/10.1016/j.system.2012.01.003

Burson, K. A. (2012). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *12*(4), 437–475. doi:10.1023/A:1009084430926

Clifford, R. (2016). Rationale for criterion-referenced proficiency testing. *Foreign Language Annals, 49* (2), 224-234. doi: 10.1111/flan.12201

Cox, T. & Clifford, R. (2014). Empirical validation of listening proficiency guidelines. *Foreign Language Annals, 47*(3), 379-403.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*(1), 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002

Geeslin, K. L. (2003). Student self-assessment in the foreign language classroom: The place of authentic assessment instruments in the Spanish language classroom. *Hispania*, *86*(4), 857–868.

Gutierrez, A. P., & Schraw, G. (2015). Effects of strategy training and incentives on students' performance, confidence, and calibration. *The Journal of Experimental Education*, *83*(3), 386–404. https://doi.org/10.1080/00220973.2014.907230

Hassmén, P., & Hunt, D. T. (1994). Human self-assessment in multiple-choice testing. *Journal of Educational Measurement*, *31*(2), 149–160.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–34. https://doi.org/10.1037/0022-3514.77.6.1121

Labutov, I. & Studer, C. (2016). Calibrated self-assessment. *Cornell University*.

Leach, L. (2012). Optional self-assessment: some tensions and dilemmas. *Assessment & Evaluation in Higher Education*, *37*(2), 137–147. https://doi.org/10.1080/02602938.2010.515013

LeBlanc, R., Painchaud, G., 1985. Self-assessment as a second language placement instrument. *Educational Psychology Review, 12*(4), 437–475. doi:10.1023/A:1009084430926

Mahmood, K. (2016). Do people overestimate their information literacy skills ? A systematic review of empirical evidence on the Dunning-Kruger Effect. *Communications in Information Literacy*, *10*(December), 199–213. https://doi.org/10.7548/cil.v10i2.385

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Morony, S., et al. (2012). Predicting achievement: Confidence vs self- efficacy, anxiety and self-concept in Confucian and European countries. *International Journal of Educational Research*, 58, 79-96. doi:10.1016/j.ijer.2012.11.002

Moskal, B. M. (2010). Self-assessments : What are their valid uses?, *Academy of Management Learning & Education. 9*(2), 314–320.

Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing*, *6*, 1–13. https://doi.org/10.1177/026553228900600103

Pallier, G. et al. (2002). The role of individual differences in the accuracy of confidence judgments. *Journal of General Psychology, 129*(3), 257-299.

Phakiti, A. (2016). Test takers' performance appraisals, appraisal calibration, and cognitive and metacognitive strategy use. *Language Assessment Quarterly*, *13*(2), 75–108. https://doi.org/10.1080/15434303.2016.1154555

Ross, S. (1998). Self-assessment in second language testing : A meta-analysis and analysis of experiential factors. *Language Testing*, *15*(1), 1–20. https://doi.org/10.1191/026553298666994244

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management*, *9*(2), 169–191. https://doi.org/10.5465/AMLE.2010.51428542

Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology, 100*(4), 961– 976. https://doi.org/10.1037/ a0012546

Stankov, L., & Lee, J. (2014). Overconfidence across world regions. *Journal of Cross-Cultural Psychology*, *45*(5), 821–837. https://doi.org/10.1177/0022022114527345

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *TESOL Quarterly, 19*(4), 673-687

Summers, M. (2017). Investigating the use of the ACTFL can-do statements in an intensive English program for student placement. Manuscript submitted for publication

Wan-a-rom, U. (2010). Self-assessment of word knowledge with graded readers : A preliminary study. *Reading in a Foreign Language*, *22*(2), 323–338.

Appendix

Pre-test and Post-test Survey adapted from ACTFL Can-do Statements

Name:
Country of Origin:
First Language:
Gender:
Age:
Consider the following statements and indicate how confident you are that you could complete the following tasks in English:

1. I can recognize items on a grocery list in English.
2. I can understand basic familiar information from a newspaper ad in English.
3. I can understand a text message from a friend in English.
4. I can understand what I need to fill out on an application form in English.
5. I can understand some information on job postings in English.
6. I can understand the main idea of a summary of a historical figure's accomplishments in English.
7. I can understand information about an upcoming activity on a flyer in English.
8. I can follow instructions to make an online purchase in English.
9. I can read a description about a candidate to make a voting decision in English.
10. I can read an article about how technology has changed in the past 20 years in English.
11. I can follow the plot in a short story in English.
12. I can understand the author's opinion in a persuasive essay in English.
13. I can make inferences about the author's purpose from a text in English.