SELF-ASSESSMENT: A FEISTY OR RELIABLE TOOL TO ASSESS THE ORAL PROFICIENCY OF CHINESE LEARNERS?

By

Wenyue Ma

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Teaching English to Speakers of Other Languages – Master of Arts

2018

ABSTRACT

SELF-ASSESSMENT: A FEISTY OR RELIABLE TOOL TO ASSESS THE ORAL PROFICIENCY OF CHINESE LEANERS?

By

Wenyue Ma

In this study, I took a close look at the results of oral proficiency self-assessment tests and OPIc (Oral Proficiency Interview-computer) tests taken twice by the same group of students. I did this to explore the role of self-assessment in Chinese language programs. The data were collected as part of a Language Flagship Proficiency Assessment Project. I used data from 80 college students who were studying Chinese. During the spring of two consecutive years, the students took a self-assessment (with NCSSFL-ACTFL Can-do Statements, 2015) as part of the project, and then immediately took an official ACTFL OPIc with a level of difficulty that was matched to their self-assessment outcome. I analyzed the self-assessment results on both the test and item level. In general, I investigated whether self-assessment can reliably indicate students' language gains over time, with the benchmark of true gain being (in this study) their OPIc scores. The findings revealed that most students' language trajectories were reflected by the results of the self-assessment. In addition, the accuracy rate of self-assessment was positively correlated with students' proficiency levels. After a close examination of the items that were misidentified by the students regarding the difficulty level, students tended to under-assess rather than over-assess their oral proficiency. The comparison of the scores of repeated self-assessments and OPIc tests showed that there was no significant difference in how accurately students could self-assess themselves before and after an academic year in a language program.

Keywords: self-assessment, Chinese proficiency test, oral proficiency, validity

This master thesis is dedicated to Mom and Dad.
Thank you for always supporting me.

ACKNOWLEDGEMENTS

The data used in this MA thesis was collected as part of a larger grant project funded by the National Security Education Program's Language Proficiency Flagship Initiative (grant # 2340-MSU-7-PI-093-PO1) awarded to principal investigators Paula Winke and Susan Gass. I, Melody Wenyue Ma, was a Graduate Research Assistant on the project: I served as one of the proctors who administered the test to undergraduate students at Michigan State University, and I worked with the PIs on subsequent data analyses and other research tasks. I borrowed the data from the project, having received the data as "pre-existing," and without identifiers (the names and ID numbers were removed before I received the data). I would like to thank the Flagship Project and the PIs Drs. Winke and Gass, along with the other proctors, research assistants, and key project personal who helped me with various tasks, including Dr. Emily Heidrich, the Project Manager, Dr. Angelika Kraemer, the executive director of CeLTA, the unit that hosted the grant project, and Mr. Amaresh Joshi, the project's and CeLTA's data manager, who provided me with the anonymized data in a readable file format.

In addition, I would like to express my appreciation to my parents, who always support and believe in me through my two-year master studies. Then I want to express my appreciation to my boyfriend, who keeps me company when I need, and to my friends Xiaowan Zhang, Zhonghao Wang, Hao Wang, Shinhye Lee, Myoengeun Son, Ian Solheim, Rachel Lin, and Ziyue Deng, who have offered me different kinds of help on both my studies and life,

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

To assess learners' language proficiency, two main approaches can be adopted: examinations, or self-assessment. The former is sometimes considered the only reliable way of assessing one's foreign language proficiency, whereas the latter, although drawing growing attention and having been applied in a lot of language test administrations, is often criticized (e.g., Davidson & Henning, 1985) as an improper assessment method. Most arguments against the use of self-assessment focus on the inherent unreliability of subjective evaluation. However, the question of subjective evaluation does not necessarily imply the invalidity of self-assessment. Actually, self-assessment has been widely used in diverse language learning and teaching contexts ranging from a small-scale, classroom-based language learning assessment to a high-stakes, language proficiency test. In the ACTFL (American Council on the Teaching of Foreign Languages) Oral Proficiency Interview–computer (OPIc) test (ACTFL, 2012a) and ACTFL Writing Proficiency Test (WPT) (ACTFL, 2012b), test candidates are asked to complete a self-assessment survey at the beginning of the assessment. The test form, test content, and even test result to some extent are dependent on their responses to the questions in the survey. Similarly, the Avant PLACE test, which was developed by the Center for Applied Second Language Studies (CASLS), starts with a self-evaluation where students need to assess their own language proficiency before answering the questions in the test. The results of these high-stakes tests can have a great impact on important decisions related to students, educators, and institutions, such as successful conferring a degree, admission to a program, and fulfilling job requirements. As for self-assessment in the classroom context, it can simply be a questionnaire with a couple of questions, and students may need to offer their own evaluation on how much they have learned

from the class. This feedback can be crucial to their language learning, and may provide valuable implications for language teachers.

The sustained popularity of self-assessment has prompted researchers to conduct empirical studies related to this topic focusing on various research questions, such as examining the validity of self-assessment in different language teaching and learning contexts, investigating the influences of the features of a self-assessment surveys with regard to their effectiveness or accuracy, and teasing apart test takers characteristics that may affect self-assessment results. In the current study, which is an extension of the previous studies, I focus mainly on college learners of Chinese who completed an oral proficiency test and a self-assessment survey with Can-Do statements two times before and after an academic year in a language program. The students took these measures as part of their Chinese program's participation in a Language Proficiency Flagship grant, a program that was implemented to give recommendations to the language programs at the university in regards to their established language program goals and how they were meeting them.

In this study, which was carved out from the grant project, I examined whether the results of self-assessment could be used as a reliable tool to track Chinese program students' language trajectories, and I looked at the results on both the global and item levels. I also investigated how accurately these students could self-evaluate their oral proficiency on the item level and whether they could successfully perceive the difficulty level of each NCSSFL-ACTFL Can-Do statement. In addition, I tried to tease apart the factors that potentially influence the result of self-assessment, such as students' language proficiency levels and the different criteria of determining whether the difficulty level of a task was successfully identified by the student. In the next section, I review both theory-based and empirical studies on self-assessment, including (a) the rationale of the use

of self-assessment, (b) the validity examination of self-assessment, and (c) the factors that may have an influence on the result of self-assessment.

LITERATURE REVIEW

Why Self-Assessment?

In language studies, self-assessment has been widely used and serves a variety of purposes, including placement, program evaluation, judgement of attitudes and socio-psychological differences, measure of course grade, learning diagnosis, and feedback to learners (Henning, 1987). Self-assessment has been criticized by some researchers, however, who consider the use of self-assessment as no more than a subjective and cursory self-grading done by learners themselves (Dickinson, 1987; Patri, 2002). Raymond and Gisèle (1985) held a different idea and pointed out in their article by first making a distinction between the concepts of self-assessment and self-grading, although the former, they noted, might include the latter in some cases. In addition, they suggested that self-assessment was not "an informal exercise based loosely on the student's intuition" (p. 674). Furthermore, researchers have suggested that the accuracy of self-assessment can be improved in several ways (e.g., Taras, 2001; Delgado, Guerrero, Goggin, & Ellis, 1999). I will elaborate more on this in a later section. In this section, I will mainly focus on the advantages of the use of self-assessment.

Shared assessment burden. In 1985, Raymond and Gisèle conducted a study in which they carred out several experiments leading to the use of self-assessment as a placement test. Raymond and Gisèle (1985) wrote in their article that the use of a self-assessment questionnaire could alleviate test administration burdens. Recruiting proctors is not needed in self-assessment. After all, for students, it is meaningless and unnecessary to cheat in self-assessments. Apart from that, students can even take questionnaires home and give them back to their teachers after finishing them, thus enabling teachers or test administrators to be exempted from routine procedures in examinations like establishing testing schedules and finding appropriate testing

rooms. This was in line with a claim made by Dickinson (1987) who wrote that learner participation in evaluation is beneficial because then the students share the burden of assessment with the teacher.

Beneficial effects on learning. Learners' joint efforts in the assessment process can be considered beneficial to their language development. This advantage of self-assessment was addressed in Oscarson's (1989) article, who attached great importance to learners' ability to make reliable and valid autonomous judgements of their own oral proficiency. According to him, these judgements were regarded as a crucial part of the learning process. Similarly, Cardoso (2010) claimed that self-assessment provides learners with opportunities to gain control over their learning, thus prompting them to reflect on their learning and determine whether their gains are concomitant with their efforts and goals. Cavana and Luisa (2012) investigated the effects of self-assessment on learners' learning styles and learning strategies in their pilot project by asking 17 volunteer students of English to use the electronic European Language Portfolio (eELP), which is an electronic version of language biography to assess learners' language proficiency. They found out that the use of the eELP could affect the learners' learning positively in giving them insight into their learning process, increasing their self-confidence, and helping them to set learning goals. The findings of a number of empirical studies imply that the use of development-oriented self-assessment would ultimately lead to enhanced learning productivity and learner autonomy, greater motivation, less frustration, and higher learning retention rates (Peirce, Swain, & Hart, 1993; Rivers, 2001). For example, in Rivers's study, she investigated if students could correctly assess their progress, learning styles, and strategy preferences. All learners in her study were found to have self-directed learning behaviors based on their self-assessments.

The Validity of Self-Assessment

In recent years, an increasingly growing number of researchers have investigated the validity of self-assessment (Butler & Lee, 2006; Cardoso, 2010; Delgado et al., 1999; Dolosic, Brantmeier, Strube, & Hogrebe, 2016; Kaderavek, Gillam, Ukrainetz, Justice, & Eisenberg, 2004; Malabonga, Kenyon, & Carpenter, 2005). In second language studies, examining the correlation between results of self-assessment and performances of specific skill areas is the most commonly-used approach to evaluate whether the self-assessment is valid or not. However, the results of their findings are mixed. Below I review these studies and what they have found.

Positive correlation between SA and performance. In many cases the authors of research studies correlating test scores with self-assessment scores have indicated that the two measures of the learners' performances were highly correlated (Stansfield, Gao, & Rivers, 2010; Dolosic, Brantmeier, Strube, & Hogrebe, 2016; Malabonga, Kenyon, & Carpenter, 2005). In other words, the validity of self-assessment was examined, and found to be high.

For example, Stansfield et al. (2010) investigated if self-assessment scores from 323 learners of total 8 different languages could be utilized to provide information for the National Language Service Corps (NLSC), thus enabling the Corps to make important decisions in terms of applicants' screening into the program. Specifically, the authors and the Corps wanted to examine if the score of the self-assessment could be used to accurately identify whether applicants' target language proficiency was adequate to perform their jobs. In the study, each applicant completed a two-part self-assessment composed of a serious of Can-Do statements and a simplified set of ILR (Interagency Language Roundtable: http://www.govtilr.org/Skills/ILRscale1.htm) skill level description, and the score of self-assessment is a composite score of the two parts. Researchers found that Oral Proficiency

Interview (OPI) scores received by the applicants were highly correlated with the oral self-assessment scores at a statistically significant level.

In Dolosic et al.'s (2016) study, the authors examined the relationship between self-assessment and oral production in French. They included 24 students who enrolled in a French language summer camp. Although students were not able to accurately self-assess their French proficiency skill upon arrival at the intensive language-learning summer camp (pre-test), they demonstrated great improvement in the accuracy of self-assessment at the end of the program (post-test).

In another study, Malabonga et al. (2005) investigated if self-assessment was a suitable tool to help examinees choose an appropriate starting level on the Computerized Oral Proficiency Instrument (COPI: http://www.cal.org/resource-center/publications/copi). They had 55 learners of Arabic, Chinese, or Spanish come into a laboratory setting to take two exams, the Simulated Oral Proficiency Instrument (SOPI) and then the COPI, with the order of tests randomized. Their findings revealed that self-assessment was a reliable tool for examinees to be assigned to the test tasks at appropriate difficult levels.

Negative or no significant correlation between SA and performance. However, not all studies have found positive correlations between the results of self-assessment and objective measures of language proficiency. For example, Peirce et al. (1993) examined whether self-assessment is a valid and reliable indicator of French proficiency. They had approximately 500 leaners in French immersion programs take a self-assessment and a French proficiency test successively. Results showed only weak correlations between self-assessments of language proficiency and learners' later tested performance.

Another study was conducted by Lim (2007) with learners of English. She compared the results of self-assessment on learners' oral proficiency and ratings by their tutors. She revealed that although self-assessment could be a potentially new way to assess learners' own language proficiency, some learners, especially those who were at a lower level of language proficiency, lacked the objectivity and confidence to correctly self-assess, and found it difficult to identify the weaknesses of both their own and others' language skills.

The results of the abovementioned studies are in line with the study by Brantmeier (2006) involving 71 Advanced L2 learners of Spanish participate. She investigated if the scores of self-assessments could be utilized to accurately predict learners' reading performance and subsequent reading achievement. However, the findings showed that self-assessment was not a reliable indicator for either placement purpose or subsequent performance. In another study by Brantmeier and Vanderplank (2008), the researchers investigated if pre-test self-assessment ratings of reading, as measured via both descriptive and criterion-referenced instrument, could reliably predict achievement on a computer-based test. They had 359 learners of Spanish self-assess their L2 reading abilities and then complete several tests for placement. Based on their findings, they concluded that self-assessment could provide useful, however fairly limited, reliability for reading placement purposes. In more detail, when learners' comprehension was measured via sentence completion and multiple choice items, a descriptive and criterion-referenced self-assessment can be an appropriate indicator for both reading scores and subsequent classroom performance. When it came to a measure of reading comprehension with a writing task for recalling short stories, the criterion-referenced questionnaire was not a reliable predictor.

The mixed results of the studies I have reviewed above seem to align with suggestions concerning the "high-stakes" issue of placement, and that perhaps self-assessment might not be best for placement testing: Both Bachman and Palmer (1981) and Brown and Gerhardt (2002) advocated for the use of self-assessment in non-high-stakes testing situations, such as for formative assessment, self-monitoring, or lower-stakes assessment purposes. But a general trend in the research seems to be suggesting that self-assessment might work or be an economical way to help students choose their starting point in a high-stakes computer adaptive test (Malabonga et al., 2005).

Factors Influencing the Accuracy of Self-Assessment

Apart from the issue of validity, some researchers were specifically interested in those factors which may have an influence on the validity of self-assessment, on which I would like to elaborate a little bit in this section.

Feedback. Some researchers emphasize the importance of providing feedback for learners to promote more accurate self-assessment. In general, self-assessments of language skills or abilities were found to be more accurate and reliable when learners received feedback in regards to their performance on objective measures of the targeted skills or abilities.

For example, researchers (Delgado et al., 1999) examined the following two areas: first, the extent of accuracy of bilingual students (80 bilingual Spanish and English college students) judging their language competence; and second, if providing feedback for students could influence the results of self-assessment. Findings of their research showed that feedback from the objective test improved self-assessment accuracy on both languages, but more significantly in Spanish.

The positive influence of feedback provided to learners on the accuracy of self-assessment was also verified in another study by Taras (2001). Her study was not on language learning, but rather on self-assessment of skills in higher education in general. The students were asked to prepare translation text and translation commentary, which were subsequently returned with their tutor' feedback. The tutor would withhold the students' grades until they worked through the tutors' feedback and completed self-assessment. The findings of the study showed that student receiving tutor feedback prior to self-assessment was better able to identify their own weaknesses and errors. She went further in another study (Taras, 2003) to have 17 final-year undergraduate students carry out two types of self-assessment: self-assessment prior to peer and tutors' feedback and self-assessment incorporating feedback as an integrated part. The results revealed that students were overwhelmingly in favor of the latter and did better in the latter context.

Age. While participants in most of the studies on self-assessment are university students, there is some research involving children. In these studies, researchers were particularly looking at what kind of role age was playing in influencing the results of self-assessment.

One of the studies (Kaderavek, Gillam, Ukrainetz, Justice, & Eisenberg, 2004) was conducted with 401 children whose ages ranged from 5 to 12 years old. The researcher mainly focused on learners' metacognitive ability and oral narrative production. They had the children take the Test of Narrative Language (TNL) and had them self-evaluate their ability of narration. The results of their study demonstrated that younger children could not as accurately as older children self-assess their narrative performance. In addition, the difference was statistically significant in the performances in narrative production skills between children who evaluated themselves as less competent speakers and those who evaluated themselves as more skilled

speakers. Their findings corresponded well to the results of another study by Butler and Lee (2006), who suggested that students younger than those at the fourth grade level were not good at self-assessment. Administering self-assessments to them may not be a good choice.

Gender. Another important issue related to the accuracy of self-assessment that researchers have been concerned about is gender differences. Results of a study (Pallier, 2003) showed that compared with women, men tended to consistently rate themselves higher, which implied the overestimation of their performance, and this tendency for men to express higher levels of confidence than women in self-assessment appeared to remain consistent across the age ranges. In the abovementioned study (Kaderavek et al., 2004), researchers also examined how the accuracy of self-assessment varied in relation to gender. They found that male students were more likely to overestimate their narrative skill than female students.

Language proficiency. Learners' language proficiency is another possible factor that may be playing a role in the results of self-assessment. Its influence has been investigated by some researchers. For example, Brantmeier, Vanderplank, and Strube (2012) provided details demonstrating that with the use of self-assessment, compared with students of lower level proficiency, students at the Advanced stages of proficiency were better at identifying the skills in which they were relatively better or poorer. Accordingly, other researchers (Kaderavek et al., 2004) drew a similar conclusion that in comparison. In comparison with children who had more Advanced speaking skill, children with poorer narrative skill were more likely to overestimate the performance of their narrative production.

Research Questions

To my knowledge, based on the findings from the prior literature, only a few researchers (e.g., Dolosic et al., 2016) have investigated whether students who self-assess multiple actually

assess themselves as getting better in the skills. In addition, research on items in self-assessment has been narrowly focused, so a fine grained look at the items in self-assessment is needed. In this case, the following research questions are established for the present study:

1. Can the results of self-assessment reflect students' language gain or attrition over an academic year?
2. Can students perceive the difficulty level of the questions in the self-assessment? Does this perception vary among students at different proficiency levels?
3. Can students better self-assess themselves regarding their oral proficiency after they spend an academic year learning?

METHODOLOGY

Participants

The participants in the current study were students from a large Mid-western U.S. university in a Chinese language program. The data were collected as part of a larger grant-funded project from 42 students who took the 50-statement self-assessment questionnaire and the Oral Proficiency Interview – computer (OPIc) (Language Testing International, 2012) in spring 2015 and spring 2016, and from 40 students who took the 50-statement self-assessment questionnaire and the OPIc in spring 2016 and spring 2017. Three things need to be noted here: First, those students who did not have completed sets of OPIc were not included in the study. Second, in my dataset, there are 20 students who took the test three times in three consecutive years. To make full use of the data, I randomly and evenly divided these 2 students into two groups and used, for each of these 20, only two of their three test results as aligned with their group assignment; either their 2015-2016 data, or their 2016-2017 data. This allowed me to keep the 20 students in the participant pool. Third, the data that meet either of the two following conditions were eliminated for analytic purposes: Missing data (students did not take the test) and the data of those who had a BR (below range) or a UR (unratable). Accordingly, the data of four students were excluded due to a BR (N=3) or a UR (N=1).

Materials

The materials used in the study include two components: A computer-adaptive self-assessment questionnaire and an official ACTFL OPIc. The questionnaire was developed by the PI (Winke) and her research assistants in consultation with ACTFL assessment team, which included five sets of ten Can-Do statements (50 in total) that were selected from the fuller list of NCSSFL-ACTFL Can-Do Statements (ACTFL, 2015). Each set of statements covered a range of

ACTFL levels with each statement targeting a certain level of proficiency (e.g., the first set of statements covered ACTFL levels from Novice Low (NL) to Novice High (NM) and item 3 "I can say which sports I like and don't like" targeted the level Novice Mid (NM)). Likert scales were used in the questionnaire: Participants were asked to rate how well they could perform the task described in each statement on a scale ranging from one to four: 1 ("I cannot do this yet"), 2 ("I can do this with much help"), 3 (I can do this with some help), and 4 ("Yes, I can do this well").

The items of the questionnaire administered in spring 2015 and spring 2016 were the same. However, a revised version with 15 items taken off and another 15 items added in was administered in spring 2017; this was done after an examination of the validity of the statements; the 15 items that were taken off were identified as misfitting (Tigchelaar, Bowles, Winke, & Gass, 2017). Thus, the continued use of these items would be considered problematic. Thus, my analyses in this paper are with the 35 common items used across the questionnaire administrations (see Appendix).

Procedure

The Chinese language learners were told by their instructors to take the OPIc as a course requirement, although their grades were not influenced by their performance. The self-assessment and OPIc, which lasted roughly 50 minutes, were administered by proctors within the university's language learning computer lab, which is maintained by the language programs' center on language learning. The test takers needed to complete a background questionnaire, the self-assessment, and then an official ACTFL OPIc with a level of difficulty that was matched to their self-assessment outcome (there were five OPIc forms, as will be described below). Learners first took the computer-adaptive self-assessment test with five levels, and based on the

outcome of the self-assessment, were recommended by the self-assessment algorithm output to take one of the five levels of the OPIc.

On the self-assessment, the five levels were computer-adaptive, and here I explain this more. The learners indicated on each level (that had 10 self-assessment questions) the extent to which they could do well on the Can-Do statements within that set, and if they scored high enough (80% or higher), they moved on to the next set of 10 statements. In the last set, the learners would be recommended to take the Level 5 OPIc if they indicated that they could do at least 8 tasks very well, or they would be recommended to take the Level 4 OPIc if they indicated they could not do 8 of the last 10 tasks very well. The items included in each level and the cut-off score were determined by the PI and her research assistants in consultation with ACTFL assessment expertise to ensure that the test would work the best.

Student performances on OPIc were rated by official certified ACTFL raters (as hired by Language Testing International), and students were informed of their proficiency level approximately two weeks after testing. Table 1 is summary on the number of participants completing each set of statements and the number of participants who took each level of the OPIc test.

Data Analysis

Before a further step was taken, I recoded some data for analytical purposes. First, I assigned a numeric value to each ACTFL proficiency level on a scale from 1 for Novice Low (NL) to 10 for Superior (S). Second, I recoded the responses to the items in the survey to cater to the need of this study, even though in the original survey, a Likert scale ranging from one to four was used, as designed by the original creators of the survey (see Tigchelaar, Bowles, Winke, & Gass, 2017):1 ("I cannot do this yet"), 2 ("I can do this with much help"), and 3 (I can do this

with some help) were recoded as 0; while 4 ("Yes, I can do this well") was recoded as 1. The rationale of this recoding was that a student should only move on to the next difficulty level if he or she could do most tasks in the level well ("Yes, I can do this well").

Research question one

To answer the first research question on whether the self-assessment can reflect students' language gain or attrition over an academic year, I present some descriptive data. In this study, students' language gain or attrition was measured by the comparison of the two OPIc scores they got before and after an academic year. I report the result of the self-assessment in two ways: the difficulty level of the OPIc test they took; and their responses to each individual item. My rationale behind presenting item responses as well as the difficulty level of the test is that, as mentioned earlier, the items included in the data analysis are the same 35 items having been used consistently across the three academic years, which means that there can be some discrepancies between the inferences drawn from these 35 items and from the full 50 items. The difficulty level suggested for each student could be supplementary evidence apart from the item responses.

With the data at hand, I answer the first research question by tallying and comparing the number of items in the survey assumed to be attainable by students in two years with their OPIc test scores. For example, a student received Intermediate Mid (IM; 5) and Advanced Low (AL; 7) on the OPIc tests before and after an academic year, and the student took the third and fourth difficulty level of OPIc test respectively. Among the items completed, the student indicated that there are 22 and 31 items that aligned with what the student can do, respectively, each year. This way the result of the self-assessment can be considered successfully having tracked the student's language gain over a year. However, if the student received Advanced Low (AL; 7) in the first year and Intermediate Mid (IM; 5) in the second year, everything else being the same, the result

of the self-assessment cannot be considered successful in this case. Table 1 clearly illustrates in what situation the result of self-assessment is respectively considered successful or not.

Table 1.
Criteria for Accurately and Inaccurately Self-assessing Oral Proficiency

Accurate	Inaccurate	
S1 > S2; T1≤T2; L1≥L2	S1 > S2; T1 > T2	S1 > S2; L1 > L2
S1 < S2; T1≥T2; L1≤L2	S1 < S2; T1 < T2	S1 < S2; L1 < L2
S1 = S2; T1≠T2; L1≠L2	S1 = S2; T1 = T2	S1 = S2; L1 = L2

Note. S1: a student's first-year OPIc score; S2: a student's second-year OPIc score;

T1: the number of tasks assumed to be attainable in the first year;

T2: the number of tasks assumed to be attainable in the second year;

L1: the difficulty level of OPIc test taken in the first year;

L2: the difficulty level of OPIc test taken in the second year.

Research question two

The second research question mainly dealt with the extent to which students could perceive the difficulty level that the items targeted in the survey. To have a good understanding of how well the students assessed their own language proficiency, I divided the items that they responded to into three different groups: the accurately-assessing group, the over-assessing group, and the under-assessing group. In the accurately-assessing group, students were able to correctly identify the items targeting the proficiency level lower or higher than their oral proficiency level (items with the difficulty level the same as the learner's proficiency level were all included in this group). The target difficulty level of each individual item is listed in NCSSFL-ACTFL Can-Do Statements (ACTFL, 2015), whereas students' proficiency level is shown by their OPIc test score. In the over-assessing group, students claimed to be able to

perform the task well (rate the item "4") when their oral proficiency did not reach the target difficulty level of the task. The under-assess group is the opposite, where students rated the task 1, 2, or 3 when their OPIc score is higher than the target difficulty level.

To address the criticism of the use of self-assessment from some researchers who claim that the accuracy of self-assessment can be influenced by learners' experience and proficiency (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Caputo & Dunning, 2005) and the argument that a hierarchy of item difficulty levels in the Can-Do Statement might not be perceived successfully by students (N. A. Brown, Dewey, & Cox, 2014), I adopted two ways of defining the accurately-assessing group. First, I calculated the rate of exact agreement, where a student had to precisely identify the difficulty level of the items at, below, or above his or her oral proficiency level. Second, I calculated adjacent-agreement, which is when a student incorrectly assesses items but only by one level, that is, one level below or above his or her oral proficiency level. Both of these methods are described in full by Carr (2011). By using these two different approaches to tallying the number of items and the students in each group, I have a better understanding of the extent to which the students could perceive the difficulty level of the questions in the self-assessment target.

In addition, to examine the assumption that students' oral language proficiency was playing a role in how accurately they assessed themselves, I took a closer look at the relationship between the agreement rate of self-assessment results and the students' OPIc scores by calculating the accuracy rate of the students' responses to the items in the survey and their OPIc scores. I presented the data in different proficiency level categories, which enables me to tease apart this proficiency level effect on the agreement or disagreement in terms of the match between the difficulty level of an item and students' OPIc scores. To achieve this, for each

student, I tallied the total number of the items he or she answered in total and the ones that he or she accurately, over-, and under-self-assessed respectively in two years. Accordingly, I calculated six percentages from the above noted values and the total number of items the student answered in the two surveys.

Research question three

As for the last research question, which concerns the comparison of the accuracy of self-assessment regarding learners' language development, I compared their responses to the items in the surveys with the corresponding OPIc test scores received in either the academic year 2015 and 2016 or the year 2016 and 2017. Specifically, I calculated Spearman's rank correlation coefficient based on the number of items they accurately self-assessed and their OPIc scores in two years to see the one-year studies in the language program or their language development influenced the way they assessed themselves. Similar to the second research question, I incorporated exact agreement and adjacent agreement into the analysis. Because the same group of students took the test twice, I used a paired *t* test to measure if the differences between the percentages of accurately assessing, under-, and over-assessing items were significant or not.

RESULTS

Figure 1 displays the two-year data about the number of students in each OPIc score level on the ACTFL scale. As shown in the chart, the highest proficiency level that the students reached in both years is Advanced Mid (AM), and only a very small proportion of students (N=8, 5.0%) reached the Advanced level. The oral proficiency of most students (N=123, 76.9%) clustered between Novice High (NH) and Intermediate Mid (IM). In addition, more students received a higher OPIc score in the second year, which can be reflected by the increased number of students in the levels between NH and IM combined with fewer students in the first two Novice levels in the second year.

Table 2 presents the difficulty levels students took for their OPIc tests based on their responses to the items in the self-assessment survey. It can be seen from the table that most of the students took the first two difficulty levels of the OPIc test. In other words, only a few students responded to the first 10 or 20 questions in a way that allowed them to cross the threshold of the third difficulty level of the test. Corresponding to the OPIc test results shown in Figure 1, more students took the higher levels of the test in the second year. This is clearly reflected by the fact that the majority of the students (N=61, perc.=76.3%) took the first difficulty level of the OPIc test in the first year.

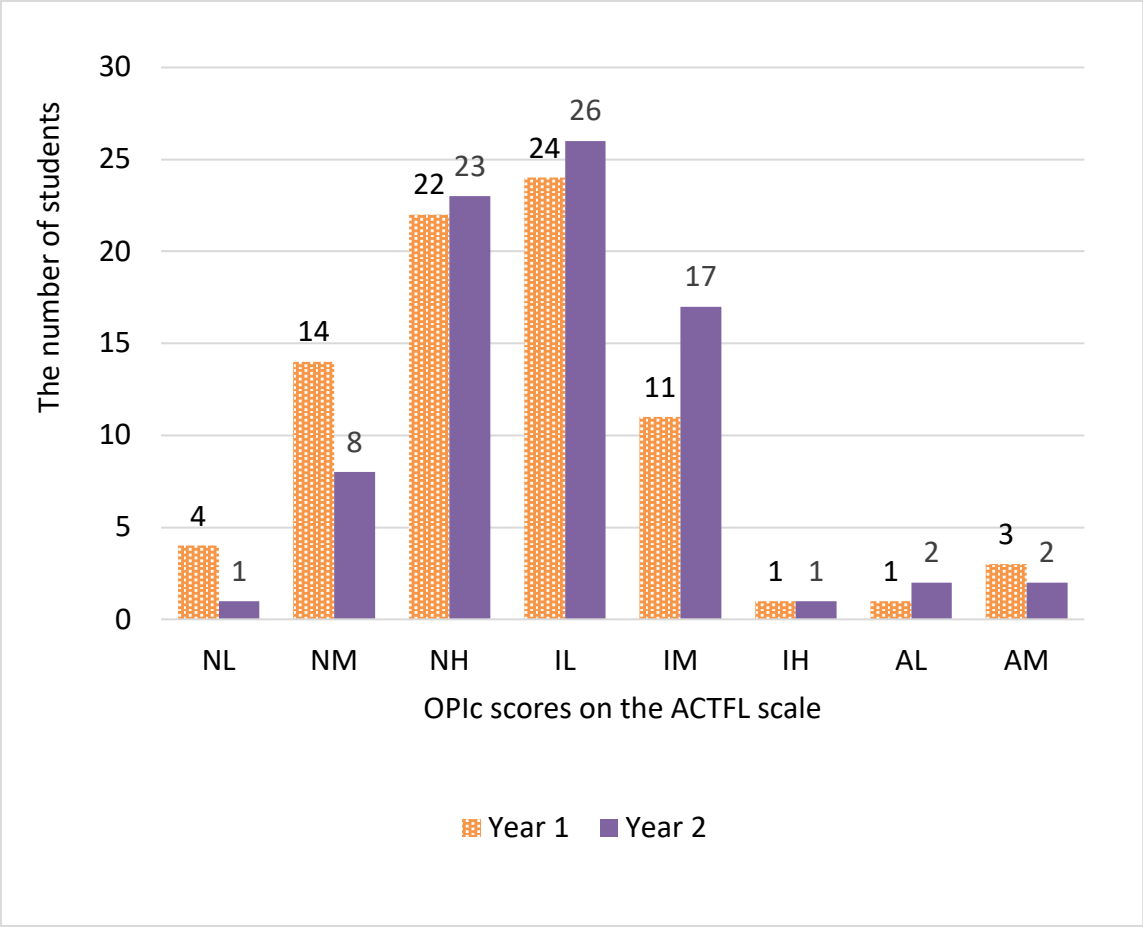

Figure 1. Descriptive statistics of the number of students and their OPIc scores in two years

Note. Novice Low (NL), Novice Mid (NM), Novice High (NH), Intermediate Low (IL), Intermediate Mid (IM), Intermediate High (IH), Advanced Low (AL), Advanced Mid (AM)

Table 2.
Number of Students Taking OPIc at Different Difficulty Levels

Difficulty level	Year 1	Year 2
1	61	55
2	14	17
3	2	4
4	2	1
5	1	3

Table 3.
Number of Students Using Self-Assessment Successfully or Unsuccessfully Tracking their Language Development Trajectories (Percentage Data in Parentheses)

Group	n	Successful	Not Successful
Language gain	35	18 (.51)	17 (.49)
Language attrition	17	10 (.59)	7 (.41)
No difference	28	21 (.75)	7 (.25)
Total	80	49 (.61)	31 (.39)

The overall descriptive statistics about how well the results of self-assessment can be used to predict students' language proficiency trajectories are presented in Table 3. Specifically, the results concern how many tasks that a student found he or she could complete with confidence in the survey and which difficulty level of OPIc test he or she took. More details about the criteria for the results either successfully or unsuccessfully exhibiting students' language gain or attrition can be found in Table 1. Both the number of students and the percentage of the number of students in each group are shown. The data in Table 3 show that 35

out of 80 students received a higher OPIc score the second time when they took the test, and among this group of students, half of them responded to the self-assessment in line with the improvement in their oral proficiency reflected in the increased OPIc scores they received in the second year. As for language attrition group, after an academic year, 17 students' OPIc scores declined by at least one level on the ACTFL scale. For 10 out of these 17 students, the results of the self-assessment are considered to have correctly predicted their language loss. In addition, 28 students received exactly the same scores in their OPIc tests before and after an academic year, and this no-change in scores was accurately reflected in 21 students' self-assessment results. The data show that the overall success rate of the results of self-assessment is moderately satisfactory (.61). Among these three groups, the success rate of the no difference group is the highest (.75%). Specifically, the students whose language proficiency levels remained the same across the two years tended to respond to the items consistently in the two years. By contrast, the students who received higher OPIc scores could not respond to the items as accurately as the other two groups in a way that reflected their language proficiency improvement.

With respect to the second research question which concerns students' perception of the difficulty level of each item in the self-assessment survey, Table 4 displays the data about the agreement and disagreement rate of self-assessment that students respectively accurately, under-, and over-assessed themselves. The results suggest that although in the exact-agreement approach, only about half of the items (53%) in the self-assessment survey were accurately identified by the students either above, below, or at their oral proficiency levels, the accuracy rate increased to a great extent when the adjacent agreement approach was adopted where for 74% of the items, the item difficulty level and students' responses corresponds accurately to their oral proficiency level. Besides that, the data indicate that among the items whose target difficulty

level mismatch students' responses regarding their oral proficiency, the difficulty levels of most of the items were shown to be below rather than above students' oral proficiency, and this contrast is even sharper in the exact agreement group (43% and 4%). In other words, students were more likely to under-assess rather than over-assess themselves in terms of how well they could complete the task.

To examine whether this tendency of students accurately or not accurately to assess themselves is related to their language proficiency levels, two scatter plots were drawn based respectively on the exact-agreement approach () and the adjacent-agreement approach (right), which presents the relationship between students' oral language proficiency levels and the agreement rate of students' responses to the items with regard to their OPIc scores. It can be seen from these two figures that the agreement rate is consistently higher for the students at Novice level (OPIc scores: 1-3) and Advanced level (OPIc scores: 7-8) than the students at Intermediate level, which is especially clearly presented by the data analyzed using the adjacent agreement approach. However, the contrast between the way the dots are scattered in these two figures is not conspicuous in the Intermediate-level score band where the agreement rate among this group of students spread out from 0 to 1. This proficiency-level related agreement rate pattern can also be well illustrated by the descriptive statistics in Table 5 where the mean agreement rate of the Novice group (.99) and the Advanced group (.80) is much higher than that of the Intermediate group (.51). As shown in the second half of the table, the contrast is especially notable when the adjacent agreement approach was adopted.

Table 4.
Summary of the Number of Items Responded by Students (Self-Assessing Accuracy and Error Rate in Parentheses)

	Exact agreement				Adjacent agreement			
Group	n	M	SD	95% CI	n	M	SD	95% CI
Accurately-assessing	816	5.10 (.53)	5.32 (.34)	[4.28, 5.92] ([.48, 58])	1062	6.64 (.74)	5.75 (.34)	[5.75, 7.53] ([.69, 79])
Under-assessing	447	2.79 (.43)	2.18 (.37)	[2.46, 3.13] ([.37, .48])	249	1.56 (.24)	2.04 (.34)	[1.24, 1.87] ([.19, .29])
Over-assessing	104	.64 (.04)	2.03 (.10)	[.16, .53] ([.01, .03])	55	.34 (.02)	1.21 (.06)	[.16, .53] ([.01, .03])
Total	1366	8.54	5.79	[7.64, 9.43]	1366	8.54	5.79	[7.64, 9.43]

Note. CI = confidence interval.

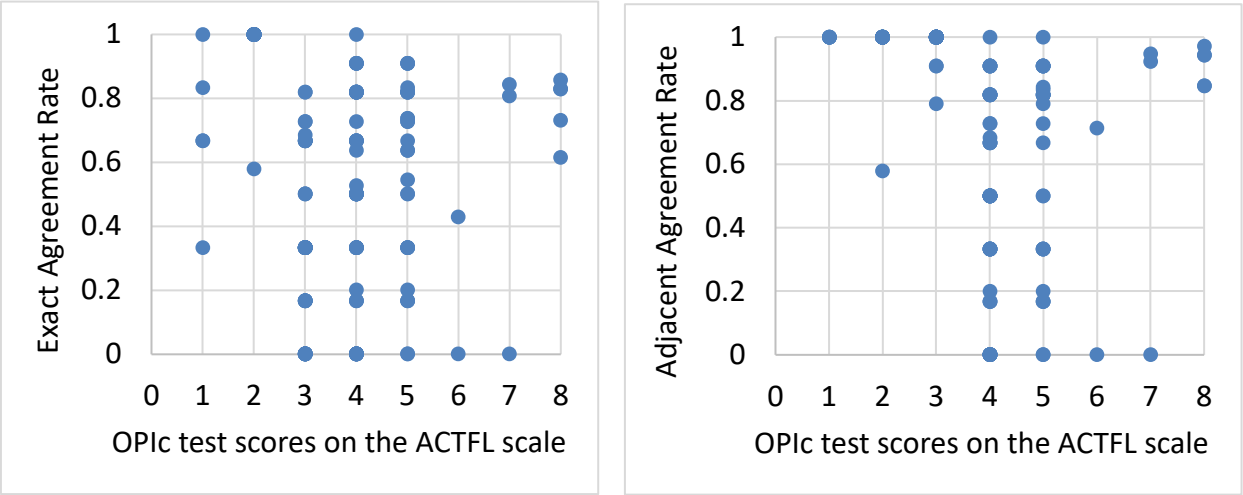

Figure 2. Two approaches presenting agreement rate of students at different oral proficiency levels. Novice Low (NL) = 1, Novice Mid (NM) = 2, Novice High (NH) = 3, Intermediate Low (IL) = 4, Intermediate Mid (IM) = 5, Intermediate High (IH) = 6, Advanced Low (AL) = 7, Advanced Mid (AM) =8.

Table 5.
Agreement Rate of Self-Assessment Responses and OPIc Scores

Agreement Rate	Oral Proficiency Level	*n*	*M (SD)*	95% CI
Exact-agreement	Novice	72	.56 (.37)	[.48, .65]
	Intermediate	80	.49 (.31)	[.42, .56]
	Advanced	8	.69 (.29)	[.49, .89]
	Total	160	.53 (.34)	[.48, .58]
Adjacent agreement	Novice	72	.99 (.06)	[.98, 1]
	Intermediate	80	.51 (.33)	[.44, .59]
	Advanced	8	.80 (.33)	[.58, 1]
	Total	160	.74 (.34)	[.69, .78]

Note. CI = confidence interval.

In addition, the data in Table 6 display the Spearman's rank correlation coefficients between the students' OPIc scores and the number of items that students accurately identified based on their own oral proficiency levels in two years. The rationale of presenting the data is to, based on what is shown in Table 5 and Figure 2, tease apart the proficiency-level effect on the extent to which students could accurately assess themselves. The agreement rates shown above in Table 5 and Figure 2 are determined by the total number of items students responded in the survey, which is already highly related to their proficiency levels, whereas the number of items that students accurately identified in the survey are less influenced by other variables. To examine whether students' language proficiency levels were playing a role in the correlation or the lack of a correlation, the data in Table 6 are presented so they correspond to the three major language proficiency levels: Novice, Intermediate, and Advanced. Table 6 shows that the coefficient of the correlation between the number of items that students correctly identified regarding their oral proficiency levels and their OPIc scores is different for students at the different oral proficiency levels. While there seemingly exists no, or very weak if any,

correlation between the OPIc scores and the number of items that students at Intermediate level accurately assessed themselves, the correlation is moderate to strong among the students at Novice and Advanced levels, and this finding is consistent before and after an academic year. For Advanced students, the number of items that they correctly identified based on their oral proficiency levels tended to be positively correlated with their OPIc scores, although this result is not statistically significant probably due to the small sample size. However, a statistically significantly negative correlation ($r_{year\ 1} = -.52, p < .001$; $r_{year\ 2} = -.66, p < .001$) was found among the Novice students, which indicates that the students at Novice Low (NL) tended to accurately identify more items in the self-assessment survey regarding their oral proficiency than the ones at Novice High (NH).

With respect to the third research question, which relates to whether students could better assess themselves after spending an academic year studying the target language, Table 7 display both the two-year data on self-assessment agreement in two ways where the agreement rate and the number of correctly self-assessing items are both presented. Similar to Table 5, the results are shown in both the exact-agreement and the adjacent-agreement approaches. Interestingly, it can be seen that the two-year data present different results using different standards to measure the agreement. Although the results are not statistically significant, the agreement rate drawn from students' responses to the items when they first took the test is slightly higher than that in the second year, whereas students tended to respond to fewer items with accuracy in the first year. This finding was consistent no matter which agreement approach was adopted.

Table 6.
Correlations between the Number of Accurately Self-assessing Items and OPIc Scores (Exact-agreement Approach)

Proficiency Level	Year 1		Year 1	
	n	Spearman's rho	*n*	Spearman's rho
Novice	40	-.52***	32	-.66***
Intermediate	36	.11	44	.17
Advanced	4	.54	4	.89
Total	160	.06	160	.19

Note. **p* < .05; ***p* < .01; ****p* < .001.

Table 7.
Summary of the Contrast between Self-assessment Agreement in Two Years

	Agreement Rate				No. of Accurately Self-assessing Items			
Approach	*n*	*M (SD)*	95% CI	*p*	*n*	*M (SD)*	95% CI	*p*
Exact Agreement								
Year 1	80	.56 (.33)	[.49, .63]	.13	80	4.93 (4.80)	[3.87, 5.98]	.33
Year 2	80	.51 (.34)	[.43, .58]		80	5.28 (5.82)	[4.00. 6.55]	
Adjacent Agreement								
Year 1	80	.77 (.32)	[.70, .84]	.18	80	6.40 (5.12)	[5.27, 7.52]	.30
Year 2	80	.72 (.35)	[.64, .79]		80	6.88 (6.34)	[5.48, 8.27]	

DISCUSSION

One of the major findings of this study is that the accuracy rate of the self-assessment is related to students' language proficiency levels, which is not surprising based on the findings of previous research studies (e.g., Brantmeier et al., 2012; Kaderavek et al., 2004). However, different patterns of accuracy seemed to exist in the three major proficiency levels. Among the Advanced-level students, the more proficient the students were, the higher the self-assessment accuracy rate. While among the Novice-level students, the lower the students' language proficiency levels are, the higher the self-assessment accurate rate was. As for the students at Intermediate level, no clear correlation could be found. This proficiency-level-related pattern in the accuracy rate was even more notable when the adjacent-agreement approach was adopted. The overall pattern taking all three levels of proficiency into consideration is that the accuracy rate increases, there is an accelerating agreement in self-assessment and proficiency—the relationship between the two variables (self-assessment and proficiency level) is not linear as proficiency increases. Rather it appears in this dataset to be parabolic relationship that is relatively high at the lowest level of proficiency. The correlation coefficient gets larger when students are mostly at Intermediate level of proficiency and it is even higher when they are at high proficiency level. This parabolic relationship calls for test methods that can test for a non-linear relationship between two variables. However, so far, apart from a few studies, researchers of most studies on self-assessment have used a simple correlation, which tests a linear relationship between two variables only, and I see a parabolic pattern between the self-assessment results and students' proficiency levels: Scatter plots are needed in this type of research to better illustrate. Based on what I could find, only the studies conducted by Brown et al. (2014) and Dolosic et al. (2016) used scatter plots to present their findings. Most of the

researchers of the studies on self-asessment, such as Brantmeier et al. (2012), Lim (2007), Roever and Powers, (2005), and Delgado et al. (1999) used only correlations to examine the relationship between self-assessment accuracy and the focused variables. Multivariate regression or some other statistical methods to test a parabolic relationships are needed for further research.

The different accuracy patterns that exist in these proficiency levels may be related to several factors according to the findings of the previous research studies. The accuracy of self-assessment can be influenced by learners' language proficiency levels, and the result of the current study align with the findings of the studies by Brantmeier et al. (2012) and Kaderavek et al. (2004) acknowledged that Advanced language learners were better at identifying the tasks that were beyond or within their capabilities. The sample size for the Advanced group in this study is far from satisfactory (n= 8), so a larger sample size of Advanced learners is needed to draw a more conclusive result. So far, most the authors of studies on self-assessment have focused mainly on learners enrolled in language programs, who are mostly at Novice or Intermediate levels. Not many researchers have conducted studies to investigate Advanced learners' performance in self-assessment, and this work has to be done to present more evidence for a more generalizable finding. In this study, I attempted to do so but did not succeed because it turned out that most of the students included in the current study were still Novice and Intermediate level learners. Therefore, most of them only had the opportunity to respond to the first set of self-assessment items (only 6 common items that were used in the survey across the three years were incorporated in this study). Accordingly, the analysis fell predominantly on the items targeting lower proficiency levels. A fine-grained investigation on more item responses that target higher proficiency levels would be valuable for future research.

To try to answer the questions why language proficiency may influence the accuracy of self-assessment, and why the accuracy rate seemed polarize Intermediate students in the current study, previous studies can help. The first question may be explained by the fact that the accuracy of self-assessment can be affected by whether students have past experiences with the task being asked about. In the study done by Tigchelaar et al. (2017), among the 50 Can-Do Statements that were used in the self-assessment survey, 5 out of the 15 misfitting items that did not fit the Rasch modeling were found to be experience-dependent. In other words, students who did not have similar experience described in the task were not likely to correctly evaluate how well they could complete the task. How much the experience that students have with the language is highly related to their proficiency. Reasonably, the more Advanced the students, the more likely they have the experiences that is specified in the task.

As for the second question regarding why the accuracy rate appeared to be scattered chaotically, the answer may have something to do with the self-assessment itself. Among the items in the self-assessment survey, some of them are specified with concrete and detailed descriptions (" I can say which sports I like and I don't like" or "I can list my favorite free-time activities"), whereas some of them are described in a more abstract and general way ("I can schedule an appointment" or "I can talk about my favorite music"). For students at low proficiency level (NL or NM), it might be easier for them to correctly identify the tasks that are within or beyond their capabilities because their language proficiency may only allow them to complete those very basic, simple, and concrete tasks, such as listing or naming a few words. However, for higher proficiency level students, when encountering the tasks that are described without many specific details, different students may adopt different interpretations of the same task description, and the difficulty level of the same item may vary among the students at the

same proficiency level. It is likely that students' responses to an item might not only be related to their proficiency levels, but also related to the imagined difficulty level of the task answered by them. The findings of the previous study by Butler and Lee (2006) may offer some insight into how to control this variable: compared with an off-task self-assessment where students were asked to self-evaluate their performance in a general and somewhat decontextualized manner, an on-task self-assessment where students could attend to the language themselves was shown to generate more accurate responses regarding their language proficiency levels.

Another main focus of this study is the extent to which the students could perceive the difficulty level of each individual item in the self-assessment survey. The results show that students tended to under-assess themselves rather than overestimate their oral language proficiency, which was not in line with the findings of most previous studies on self-assessment (Kaderavek et al., 2004; Stansfield et al., 2010; Dolosic et al., 2016). The results of these studies revealed that learners, especially those at lower proficiency levels, tended to over-assess their language proficiency. This discrepancy shown between the findings of the current study and the previous research might be explained by the different formats of the self-assessment, the population differences, and the different methodology for data analysis adopted in these studies.

In the study conducted by Kaderavek and her colleagues (2004), the learners were children between 5 and 12 years old of age. Instead of having the learners self-assessing their own language proficiency level based on concrete task descriptions, the researchers asked them to evaluate how well they could tell a good story on a Likert scale ranging from 1 to 5 with five different faces representing very sad (1), somewhat sad (2), neutral (3), somewhat happy (4), and very happy (5). Unlike the current study where a criterion-referenced instrument was used in the self-assessment, the questionnaire adopted in this study was more general, and it is possible

that without a clear benchmark for each point on the scale, students could not differentiate one from another well. Pinpointing this issue with this type of self-assessment, Brantmeier (2006) suggested that a more contextualized descriptive and criterion-referenced instrument might be more appropriate and beneficial for self-assessment purposes. What's more, in the same study, Kaderavek and her colleagues already noted that the age of learners may have an impact on how accurately they could self-evaluate their language performance – the younger the learners, the less accurately they could evaluate their proficiency. Thus to some extent, the designs of these two studies are not comparable to each other considering the different target population and the different formats employed in the self-assessment surveys. In another two studies, even though the researchers used Can-Do Statements as in the current study, the methods of scoring the self-assessments are not identical. Specifically, in the current study, although a Likert scale ranging from 1 to 4 was used in the self-assessment survey, a binary scoring system was adopted to by the administrators of the test to decide whether which OPIc test level a student should take. I kept using this scoring system in my study due to the same reason: students would move on to the next set of ten questions only if they responded to eight out of the ten questions with 4 ("Yes, I can do this well"). In comparison, in the study conducted by Dolosic et al. (2016), similarly, they used a Likert scale ranging from 1 to 5 with detailed descriptions of the tasks, but they analyzed the data using all 5 data points, which is different from the current study. In the study done by Stansfield et al. (2010), when the learners responded to the items in the Can-Do Statements, they only needed to either accept or reject an affirmative statement. Based on the information shown above, it is easy to tell that the methodology of these three studies may share some features in common, but we need to be cautious when we evaluate whether the results are comparable to each other.

It is possible that the way that the data were re-coded in the current study exaggerated the extent to which the students under-assessed themselves. For example, a student's 3 ("I can do it with little help") might not be different from another student's 4 ("I can do it very well"), but these two responses were indeed interpreted as a 'difference' by the scoring method used in this study. Similarly, one student differentiate the extent to which he or she could complete a task by responding to one item with 1 ('I cannot do this yet') and another item with 2 ('I can do this with much help'), but this discrepancy was not captured by the data analysis. Indeed, the students' responses to each task were not completely reflected in the results. Even though this coding was used in this study, the underlying raw scores are still available for analyses and future research. What kind of factors might result in the inconsistency exhibited in the responses to items in the self-assessment survey by the students at the same language proficiency level and are there any effective ways to alleviate this inconsistency? These might be the future research questions of empirical studies on self-assessment.

CONCLUSION

This study investigated the extent to which the college students learning Chinese could accurately self-assess their oral language proficiency based on the descriptions specified in the Can-Do statements and whether their ability to self-assess was related to their language proficiency levels. The results revealed that Advanced-level students were more likely to be able to successfully identify the difficulty level of a task that is within or beyond his or her capability when they are compared with their lower-proficiency counterparts, although the sample size was too small to draw a conclusive finding. In addition, among the Novice level students, students whose language proficiency is closer to beginners (Novice Low) tended to do a better job at self-assessing their proficiency in comparison to those whose proficiency is at Novice Mid or Novice High. Great inconsistency of accuracy rate was shown among the students at Intermediate level, more accurately evaluate their oral proficiency. Despite the various patterns displayed among students at different proficiency levels, the overall tendency is that the accuracy rate of self-assessment was shown to increase as students' proficiency increases.

In addition, regarding whether the students' responses to the items in the self-assessment survey could successfully predict their language trajectories, it was found that the accuracy rate was moderately higher than merely by chance. In other words, the language gain or loss that students experienced after an academic year in a language program was somewhat reflected in their responses to the statements in the self-assessment.

Apart from that, students in this study were found to be more likely to under-estimate rather than over-estimate their oral proficiency, which was not in line with the findings of previous research. After a close examination and a careful comparison of these studies, it was found that the result of self-assessment could be influenced by the different formats that the self-

assessment survey employed and the different scoring systems adopted. The results of both the current study and previous research revealed that compared with decontextualized or general descriptions without much information given for learners as a reference, a more informative and contextualized instrument was shown to generate more accurate response by learners.

APPENDIX

APPENDIX

35 Common Items in the Can-Do Statement Used from Spring 2015 to Spring 2017

ACTFL OPIc Level	Can-Do Statements	ACTFL Levels
1	I can say which sports I like and don't like.	*NM*
1	I can list my favorite free-time activities and those I don't like.	*NM*
1	I can talk about my school or where I work.	*NM*
1	I can talk about my room or office and what I have in it.	*NM*
1	I can answer questions about where I'm going or where I went.	*NM*
1	I can present information about something I learned in a class or at work.	*NM*
2	I can describe a school or workplace.	*IL*
2	I can schedule an appointment.	*IM*
2	I can talk about my family history.	*IH*
2	I can explain why I was late to class or absent from work and arrange to make up the lost time.	*AL*
2	I can tell a friend how I'm going to replace an item that I borrowed and broke/lost.	*AL*
3	I can give some information about activities I did.	*IM*
3	I can talk about my favorite music, movies, and sports.	*IM*
3	I can arrange for a make-up exam or reschedule an appointment.	*IM*
3	I can ask for and follow directions to get from one place to another.	*IH*
3	I can return an item I have purchased to a store.	*IH*
3	I can present an overview about my school, community, or workplace.	*AL*
3	I can compare different jobs and study programs in a conversation with a peer.	*AL*
3	I can discuss future plans, such as where I want to live and what I will be doing in the next few years.	*AM*
4	I can present ideas about something I have learned, such as a historical event, a famous person, or a current environmental issue.	*IH*
4	I can explain how life has changed since I was a child and respond to questions on the topic.	*AL*

ACTFL OPIc Level	Can-Do Statements	ACTFL Levels
4	I can discuss what is currently going on in another community or country.	*AL*
4	I can provide a rationale for the importance of certain classes, subjects, or training programs.	*AL*
4	I can talk about present challenges in my school or work life, such as paying for classes or dealing with difficult colleagues.	*AM*
4	I can give a presentation about cultural influences on society.	*AH*
4	I can participate in conversations on social or cultural questions relevant to speakers of this language.	*AH*
5	I can interview for a job or service opportunity related to my field of expertise.	*AL*
5	I can present an explanation for a social or community project or policy.	*AL*
5	I can present reasons for or against a position on a political or social issue.	*AL*
5	I can give a clear and detailed story about childhood memories, such as what happened during vacations or memorable events and answer questions about my story.	*AM*
5	I can exchange general information about my community, such as demographic information and points of interests.	*AM*
5	I can exchange factual information about social and environmental questions, such as retirement, recycling, or pollution.	*AM*
5	I can exchange complex information about my academic studies, such as why I chose the field, course requirements, projects, internship opportunities, and new advances in my field.	*AH*
5	I can provide a balance of explanations and examples on a complex topic.	*S*
5	I can explain, participate actively and react to others appropriately in academic debates, providing some facts and rationales to back up my statements.	*S*

REFERENCES

REFERENCES

ACTFL. (2012a). ACTFL OPIc familiarization manual. Retrieved March 15, 2018, from https://www.languagetesting.com/pub/media/wysiwyg/manuals/actfl-fam-manual-opic.pdf

ACTFL. (2012b). Writing Proficiency TEST familiarization manual. Retrieved March 15, 2018, from https://www.languagetesting.com/pub/media/wysiwyg/ACTFL-Writing-Proficiency-Test-WPT-Familiarization-Manual-.pdf

ACTFL. (2015). NCSSFL-ACTFL can-do statements. Retrieved December 11, 2017, from http://www.actfl.org/global_statements

Bachman, L. F., & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language Learning*, *31*(1), 67–86. https://doi.org/10.1111/j.1467-1770.1981.tb01373.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

Brown, K. G., & Gerhardt, M. W. (2002). Formative evaluation: An integrative practice model and case study. *Personnel Psychology*, *55*, 951–983. https://doi.org/10.1111/j.1744-6570.2002.tb00137.x

Brown, N. A., Dewey, D. P., & Cox, T. L. (2014). Assessing the validity of can-do statements in retrospective (Then-Now) self-assessment. *Foreign Language Annals*, *47*(2), 261–285. https://doi.org/10.1111/flan.12082

Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessment among Korean elementary school students studying English. *The Modern Language Journal*, *90*(4), 506–518. https://doi.org/10.1111/j.1540-4781.2006.00463.x

Caputo, D., & Dunning, D. (2005). What you don't know: The role played by errors of omission in imperfect self-assessments. *Journal of Experimental Social Psychology*, *41*(5), 488–505. https://doi.org/10.1016/j.jesp.2004.09.006

Cardoso, C. W. (2010). Self-assessment : Indispensable tools for successful learning. *New Routes*, *42*, 24–26.

Carr, N. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Cavana, P., & Luisa, M. (2012). Autonomy and self-assessment of individual learning styles using the European Language Portfolio (ELP). *Language Learning in Higher Education*, *1*(1), 211–228. https://doi.org/10.1515/cercles-2011-0014

Davidson, F., & Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. *Language Testing*, *2*(2), 164–179. https://doi.org/10.1177/026553228500200205

Delgado, P., Guerrero, G., Goggin, J. P., & Ellis, B. B. (1999). Self-assessment of linguistic skills by bilingual Hispanics. *Hispanic Journal of Behavioral Sciences*, *21*(1), 31–46. https://doi.org/10.1177/0739986399211003

Dolosic, H. N., Brantmeier, C., Strube, M., & Hogrebe, M. C. (2016). Living language: Self-assessment, oral production, and domestic immersion. *Foreign Language Annals*, *49*(2), 302–316. https://doi.org/10.1111/flan.12191

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

Kaderavek, J. N., Gillam, R. B., Ukrainetz, T. a., Justice, L. M., & Eisenberg, S. N. (2004). School-age children's self-assessment of oral narrative production. *Communication Disorders Quarterly*, *26*(1), 37–48. https://doi.org/10.1177/15257401040260010401

Language Testing International. (2012). ACTFL speaking assessment: The oral proficiency interview - computer® (OPIc). Retrieved December 11, 2017, from https://www.languagetesting.com/oral-proficiency-interview-by-computer-opic

Lim, H. (2007). A study of self- and peer-assessment of learners' oral proficiency. *CamLing*, 169–176.

Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). *Self-assessment, preparation and response time on a computerized oral proficiency test*. *Language Testing* (Vol. 22). https://doi.org/10.1191/0265532205lt297oa

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

Pallier, G. (2003). Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles*, *48*(5–6), 265–276. https://doi.org/10.1023/A:1022877405718

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, *19*(2), 109–131.

Peirce, B. N., Swain, M., & Hart, D. (1993). Self-assessment, French immersion, and locus of control. *Appl. Linguist.*, *14*(1), 25–42. Retrieved from http://ezproxy.usherbrooke.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true

&db=fcs&AN=4936928&site=ehost-live

Raymond, L., & Gisèle, P. (1985). Self-assessment as a second language placement instrument, *19*(4), 673–687.

Rivers, W. P. (2001). Autonomy at all costs: An ethnography of metacognitive self-assessment and self-management among experienced language learners., *85*(2), 279–290.

Roever, C., & Powers, D. E. (2005). Effects of language of administration on a self-assessment of language skills. *Monograph Series*, (February).

Stansfield, C. W., Gao, J., & Rivers, W. P. (2010). A concurrent validity study of self-assessments and the federal interagency language roundtable oral proficiency interview. *Russian Language Journal/Russkii Yazyk*, *60*, 301–317. Retrieved from http://search.proquest.com/docview/1430171560?accountid=13042%5Cnhttp://oxfordsfx.hosted.exlibrisgroup.com/oxford?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=article&sid=ProQ:ProQ%3Allbashell&atitle=A+Concurrent+Validity+Study+of+Se

Taras, M. (2001). The use of tutor feedback and student self- assessment in summative assessment tasks : Towards transparency for students and for tutors. *Assessment & Evaluation in Higher Education*, *26*(6). https://doi.org/10.1080/0260293012009392

Taras, M. (2003). To feedback or not to feedback in student self- assessment to feedback or not to feedback in student, *28*(5). https://doi.org/10.1080/02602930301678

Tigchelaar, M., Bowles, R. P., Winke, P., & Gass, S. (2017). Assessing the validity of ACTFL Can-Do Statements for spoken proficiency: A Rasch analysis, *50*, 584–600. https://doi.org/10.1111/flan.12286