Explicit Instruction in Second Language Self-Assessment: Exploring the Potential for Improving Calibration Through Training

Rachael Reynolds

Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

Part of the Arts and Humanities Commons

BYU ScholarsArchive Citation


This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.
Explicit Instruction in Second Language Self-Assessment:

Exploring the Potential for Improving

Calibration Through Training

Rachael Reynolds

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

Master of Arts

Troy Cox, Chair
Jennifer Bown
Laura Catharine Smith

Center for Language Studies
Brigham Young University

Copyright © 2022 Rachael Reynolds
All Rights Reserved
ABSTRACT

Explicit Instruction in Second Language Self-Assessment:
Exploring the Potential for Improving Calibration Through Training

Rachael Reynolds
Center for Language Studies, BYU
Master of Arts

This study explores the effect of an explicit training module on improving language learner ability to accurately self-assess second-language proficiency. There was a total of 409 intermediate and advanced level participants across six languages: French, Italian, Japanese, Portuguese, Russian and Spanish. Most of the participants had extensive immersion backgrounds. The Language Ability Self Evaluation Resource (LASER) was used to measure the perceived proficiency of the participants (self-assessed language ability). The Oral Proficiency Interview – Computer (OPIc) was used to measure the actual proficiency of the participants (rater-assessed language ability). The participants were randomly assigned to either the control or experimental group. Both groups started and ended the self-assessment the same way, with the experimental group receiving the training directly before the self-assessment questions. The training module was designed to include clarification, instruction, modeling, practice and immediate feedback while only adding 10 minutes to the overall assessment time. The limited time was dictated by the overall desire for maximum usability within already busy curricula and by already busy students. All participants completed the OPIc within 30 days of taking the LASER. The results were unexpected, revealing that those learners who received the training were more likely to over-assess their own proficiency level than learners who received no training at all. Possible reasons for this outcome are explored and possible ways to improve the assessment tool are discussed.

Keywords: self-assessment, second language acquisition, proficiency training, LASER, OPIc, over-assessing, proficiency, calibration
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE PAGE</td>
<td>i</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>DESCRIPTION OF THESIS STRUCTURE AND CONTENT</td>
<td>viii</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Factors Affecting Self-Assessment Calibration Accuracy</td>
<td>1</td>
</tr>
<tr>
<td>Methods for Improving Calibration Accuracy</td>
<td>2</td>
</tr>
<tr>
<td>Literature Review</td>
<td>4</td>
</tr>
<tr>
<td>Criteria and Purpose of the Self-assessment</td>
<td>5</td>
</tr>
<tr>
<td>Instruction, Modeling, and Practice</td>
<td>6</td>
</tr>
<tr>
<td>Feedback</td>
<td>7</td>
</tr>
<tr>
<td>Time on Training</td>
<td>8</td>
</tr>
<tr>
<td>Motivation for the Current Study</td>
<td>9</td>
</tr>
<tr>
<td>Methodology</td>
<td>10</td>
</tr>
<tr>
<td>Participants</td>
<td>10</td>
</tr>
<tr>
<td>Materials</td>
<td>13</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Implications for Practitioners</td>
<td>45</td>
</tr>
<tr>
<td>Future Research</td>
<td>46</td>
</tr>
<tr>
<td>References</td>
<td>48</td>
</tr>
<tr>
<td>APPENDIX A The Training Module Slides and Scripts</td>
<td>52</td>
</tr>
<tr>
<td>APPENDIX B Practice and Feedback module</td>
<td>59</td>
</tr>
<tr>
<td>APPENDIX C Can-Do Statements Considerations</td>
<td>64</td>
</tr>
<tr>
<td>Example change to S3</td>
<td>65</td>
</tr>
<tr>
<td>APPENDIX D Additional Comparisons and Discussion</td>
<td>66</td>
</tr>
<tr>
<td>Gender Across Groups</td>
<td>66</td>
</tr>
<tr>
<td>Survey Language Between Groups</td>
<td>67</td>
</tr>
<tr>
<td>Immersion Hours between Groups</td>
<td>68</td>
</tr>
<tr>
<td>Age at the Start of Language Acquisition Between Groups</td>
<td>69</td>
</tr>
<tr>
<td>The Effect of Survey Language and Training on Calibration</td>
<td>69</td>
</tr>
<tr>
<td>The Effect of Immersion and Training on Calibration</td>
<td>71</td>
</tr>
<tr>
<td>The Effect of Initial Language Acquisition Age and Training on Calibration</td>
<td>72</td>
</tr>
<tr>
<td>The Effects of Years Studying Language in School and Training on Calibration</td>
<td>73</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1 *Descriptive Statistics of Duration for All Groups.* .......................................................... 14

Table 2 *OPIc Levels of the Population* .......................................................................................... 21

Table 3 *Breakdown of OPIc Level by Group* ................................................................................ 23

Table 4 *Estimated Marginal Means of Each Group* ...................................................................... 24

Table 5 *Post Hoc Comparison Between the OPIc Levels of All Groups* ........................................ 24

Table 6 *Frequencies of Self-Assessment by ACTFL Sublevel of Entire Population* ............... 27

Table 7 *Frequencies of Self-Assessment by ACTFL Level across the Population* ...................... 28

Table 8 *Estimate Marginal Means of Self-assessments Ratings from All Three Groups* .......... 29

Table 9 *Post-Hoc Comparisons of Self-assessment Ratings* ....................................................... 30

Table 10 *Frequencies of Speaker Calibration by Sublevel Between Experimental and Control Groups* ............................................................................................................. 31

Table 11 *Results of Independent T-tests Comparing Calibration by Proficiency Level* .......... 33
LIST OF FIGURES

Figure 1  Participants by Target Language  ................................................................. 11
Figure 2  Approximation of Immersion Hours  .............................................................. 12
Figure 3  Age at the Start of Language Learning for Participants Reporting No Years of Language in School  ................................................................................................................ 13
Figure 4  Confidence Interval Between the Control and Experimental Groups Regarding Duration ................................................................. 15
Figure 5  Gender Split Between the Control and Experimental Groups ....................... 17
Figure 6  Visualization of LASER Participants and Test Procedure  ................................ 18
Figure 7  Scatterplot of Proficiency Corelated to when LASER was Taken by Language .... 26
Figure 8  Plot of Calibration by OPIc Rating Compared Between Experimental and Control groups .......................................................................................................................... 32
Figure 9  Plots of Calibration by OPIc Score between Groups ......................................... 34
Figure 10  Test Duration in Minutes for the Preliminary, Experimental and Control Groups .... 38
Figure 11  Participants Self-reporting Use of Specific Proficiency Levels in Their Native Language ................................................................................................................................. 39
DESCRIPTION OF THESIS STRUCTURE AND CONTENT

This thesis, *Explicit Instruction in Second Language Self-Assessment: Exploring the Potential for Improving Calibration through Training*, is written to satisfy both the requirements for a journal ready publication as well as the traditional thesis format. The initial pages satisfy the requirements for submission to the university whereas the following pages reflect the requirements of submission to a journal.

The thesis begins with a review and evaluation of the previous literature relevant to explicit training in self-assessment followed by a discussion of the methods used to test an explicit training module and collect the data. The methodology also includes a more detailed description of the test, the participants, the training module, and the procedures of the test. The results section contains a detailed account of the collected data followed by the discussion and conclusion section which connects these findings to previous research. This section also includes a discussion of what has been achieved through this research, both significant results as well as limitations, and posits implications for future research and pedagogical implementation.

Appendix A contains both screen shots of the slides and the script for the proficiency training module. Appendix B contains screen shots and description of the practice and feedback module. Appendix C lists the 24 Can-Do statements used in the LASER as well as a categorization of potential limitations of each. Finally, Appendix D contains additional results that were not relevant to the research questions but may be informative to the reader.
**Introduction**

Self-assessment has been used to provide both formative and summative feedback in world-language classrooms for many years. It is valuable to both the teacher–who needs to find quick and effective ways to assign grades and ascertain student needs–as well as the student–who benefits from greater insight and, arguably, greater control over their own journey towards proficiency (Engelhardt & Pfingsthorn, 2013; LeBlanc & Painchaud, 1985). However, research has shown that learners often inaccurately measure their own communicative abilities when self-assessing. It follows that this inconsistency lessens the usefulness of self-assessment in the language classroom. However, researchers suggest that self-assessment is appropriate and useful under the right circumstances (Jamrus & Razali, 2019; Li & Zhang, 2020; Ross, 2006). That is, circumstances that encourage calibration. Calibration in language self-assessment is when a learner’s actual proficiency aligns with their perceived proficiency. Factors affecting self-assessment calibration and recommendations to improve calibration are discussed below.

**Factors Affecting Self-Assessment Calibration Accuracy**

Many researchers search for ways to overcome the fallibilities of the various types of self-assessments. Studies have shown that lack of calibration in self-assessment correlates with several different aspects of the learner and the test. Among these considerations, two that consistently play a role in the test validity are learner proficiency level and testing purpose.

Learner proficiency level in the target language has been shown to play a major role in learner self-assessment calibration. Middle level learners are likely to overestimate what they can do (Boud & Falchikov, 1989; Brown et al., 2014, Ma & Winke, 2019; Piñana, 2009) while both lower- and higher-level learners usually assess themselves more accurately or sometimes underestimate their abilities (Brown et al., 2014, Dlaska & Krekel, 2008; Ma & Winke, 2019).
The goal of a particular assessment has also been shown to affect the validity of self-assessments warranting caution and encouraging that the benefits of self-assessment should not be overgeneralized. In particular, self-assessments should only be used for low-stakes testing (e.g., assessments not associated with a grade) as these tests provided little to no incentive for students to purposefully over- or under-estimate their abilities (LeBlanc & Painchaud, 1985; Summers et al., 2019). One such low stakes assessment category is placement tests which offer an incentive for students to rate accurately to be placed in an appropriate level where they will feel comfortable (LeBlanc & Painchaud, 1985; Summers et al., 2019).

Methods for Improving Calibration Accuracy

Within the broad category of self-assessment, research has found several promising methods for improving learner calibration accuracy to be more in line with the standards presented in recognized language proficiency guidelines. Two of these methods are using well-constructed Can-Do statements and explicit training. Can-Do statements are positive statements (beginning with the words “I can...”) that, in language learning, describe possible functions the learner can do in their target language or culture. (See Appendix C for a list of Can-Do statements used in this study). They are currently receiving attention as researchers explore their capacity to improve calibration between student perception and rater expertise. (Cox et al., 2018, Little 2007, Tigchelaar et al., 2017). Many world-language organizations, including US-based ACTFL and its European counterpart the Common European Framework of Reference for Languages (CEFR), have developed Can-Do statement frameworks. These frameworks consist of lists of Can-Do statement that include progressively difficult language functions based on the proficiency standards of the organization. The increasingly common development of Can-Do statement frameworks reflects the increasing value that this type of assessment is being given in
the language learning community. This value is not unfounded, as well-designed Can-Do statements have been shown to be an effective way to self-assess learner proficiency (Engelhardt, & Pfingsthorn, 2013; Shleykina, 2020; Tigchelaar et al., 2017). However, Can-Do statements come with their own weaknesses and are difficult for learners to answer if the function they describe is ambiguous, if they contain more than one function, or if the statement requires experience (Tigchelaar et al., 2017). As such creating effective Can-Do statements requires careful consideration. And despite progress in the creation of Can-Do based self-assessments, many may not facilitate self-assessment calibration, despite being consistently reliable. That is, although the assessment results are consistent with how the learner perceives his or her own language abilities, the results are not consistent with how a language educator or trained rater would assess the learner’s proficiency. (Brown et al., 2014; Cox et al., 2018; Summers et al., 2019).

Explicit training has also been discussed as a potential way to improve calibration in language self-assessments (Boud, & Falchikov, 1989; Blue, 1988; Brown et al., 2014; Butler & Lee, 2010; Dlaska & Krekeler, 2008; Engelhardt & Pfingsthorn, 2013; Jamrus & Razali, 2019; Lappin-Fortin, & Rye, 2014; Ma, & Winke, 2019; Mehami, & Razmjoo, 2016; Shleykina, 2020; Summers et al., 2019; Wolochuk, 2009). The research that has already been conducted regarding training has somewhat inconsistent results with some researchers showing improvement in the calibration with self-assessment (Chen, 2008; Wolochuk. 2009), while others only showed improvement with peer-assessment (Khonbi & Sadeghi, 2013; Taras, 2003). However, the literature is overall optimistic regarding the potential effectiveness of training, with many researchers calling for further research regarding the type of training and effect of training in self-assessment (Adams & King, 1995; Chen, 2008; Khonbi & Sadeghi, 2013; Li, & Zhang,
The purpose of this research is to report on the outcomes of an explicit language proficiency training module created to improve the accuracy of learner perception of personal language proficiency. The module was part of a self-assessment placement test given to third year language learners. The placement test used Can-Do statements to assess speaking and writing proficiency of third year world language students, at an American university.

**Literature Review**

Researchers have identified several factors that influence how accurately a learner can rate his or her own language abilities when using self-assessments. Other than learner inability to perceive his or her own proficiency, these factors include the likelihood that students will purposely rate themselves inaccurately to achieve a specific purpose (such as receiving a high grade), as well as students having an unclear understanding of what is being assessed (Adams & King, 1995). However, self-assessment research also provides insight into ways to narrow the gap between rater and learner perceptions (Boud & Falchikov, 1989; Li & Zhang, 2020; Ross, 2006). The most consistent recommendation, found in half the articles, was to provide the learner with training.

Although many articles request more research on the effectiveness of training, few offer recommendations as to how the training should be conducted. The recommendations that are stated mostly fall into the 3 categories: 1) providing and clarifying criteria and purpose for the assessment; 2) teaching students to apply criteria by instructing, modeling, and practicing; 3) giving feedback as to how well a learner assessment aligns with a teacher or raters assessment.
(Adams & King 1995, Li & Zhang, 2020; Ross, 2006). These categories will be further described below.

**Criteria and Purpose of the Self-assessment**

Although a learner’s ability to understand and identify the assessment criteria has been discussed as an important part of self-assessment training (Adams & King, 1995; Li & Zhang, 2020; Ross, 2006), this ability alone has been shown to not always improve self-assessment calibration. For example, Patri (2002) provided a 2-hour training for college age students in a remedial second-language English class. The purpose of the training was to familiarize the students with assessment criteria to prepare them to do self- and peer-assessments. However, while the student ability to peer-assess improved to the point that it correlated well with the teacher’s assessment, the individual self-assessments did not. The author cautioned that the number of participants was small and theorized that the level of language proficiency might have factored into the results as well as the fact that the course was remedial and the task of being autonomous and self-assessing was very new to the students.

Khonbi and Sadeghi (2013) had similar results using training to help learners to accurately self- and peer-assess. After receiving training on how the criteria of the assessment, the students gave short presentations on topics while peers filled out assessment forms. After the presentations, peers gave feedback to the presenter orally. Then the presenter filled out a self-assessment and all assessment forms were given to the teacher. The teacher found that peers were able to calibrate their assessment of their classmates more in line with the teacher’s assessment than the participants were able to calibrate their self-assessment. The authors concluded that the peer assessors were able to be stricter than the self-assessors. In addition, the authors theorize that learner ability to be stricter with one’s personal assessment would likely
improve through practice. The ability to be stricter with a peer is another possible reason for the results Patri (2002) observed.

Both studies included some instruction on the criteria of the assessment, and both fell short of improving learner ability to self-assess. However, there was improvement in learner ability to accurately identify how well the criteria were met by other students. The ability to better judge another learner’s work indicates there was an improvement in calibration, but only within a less personal context. That is, participants were not better calibrated when considering their own work. This limited improvement indicates that even a basic opportunity in training can improve calibration, depending on the type of assessment and perhaps on continued practice.

In several studies (Adams, C., & King, K., 1995; Dlaska, A., & Krekeler, C., 2008; Taras, 2003), learners were given the opportunity to set criteria themselves, allowing the participants to understand the criteria and purpose of the assessment more fully. Although there was improvement in calibration, it wasn’t with exception. Dlaska & Krekeler (2008) found that the advanced students especially were stricter with themselves, even to the point of underestimating their abilities while Tara (2003) found that the opportunity to create was not enough to improve self-assessment scores.

**Instruction, Modeling, and Practice**

Khonbi and Sedaghi (2013) suggested that more practice would likely improve the self-assessor’s ability to self-assess. Wolochuk (2009), provided a similar idea noting that those learners who not only understood what was required of them to complete a language task, but who also had helpful meta-cognitive strategies (acquired often through training and practice) had more accurate calibration.
Chen (2008) found that including practice as part of a 12-week cycle (2 weeks of initial training and 10 weeks of assessment) was instrumental in improving the accuracy of learner self-assessment. However, as Piñana (2009) noted, it is important to recognize that repetition alone (implicit training) has does not guarantee improvement in this area.

Schmidt and Wehmeyer (2016) sought to understand more fully how learners assess themselves and focused their training activities specifically on developing appropriate attitudes toward self-assessment in their explicit training. The 4 training sessions spread over the 8-week course dealt with reflecting, transferring, monitoring, and evaluating and took students through similar levels as Adams and King (1995). The students were taught the goals of self-assessment, to identify criteria, to separate judgement from observation, and to discuss and choose effective self-assessment strategies. Although the results were mixed, there was statistical significance in the improved learner ability to honestly assess themselves. These results led the authors to conclude that, among other things, training can help students reflect more honestly on their own abilities.

The ability to show statistically significant improvement through methodical training and practice is hopeful for those looking to better calibrate student perceptions. Given the comments from Khonbi and Sadeghi (2013) above, what seems a key change from a single focus on criteria identification was the focus on separating judgement from observation.

**Feedback**

Highlighting the role of feedback in self-assessment training, Taras (2003), who had intentionally moved away from explicit training having found that it can cause unwanted anxiety to students, incorporated the practice of helping the students consistently question the quality of their own work and seek out peer and tutor feedback. In this translation-based class, the students
were first trained on the criteria of a good translation. Taras found that the students were unable to accurately assess without feedback from their peers or a tutor, but with feedback were able to become more accurate. However, the author noted that students considered time spent on an assignment to be more important in determining their grade than actual quality of the translation.

Feedback was also part of Chen’s (2008) study involving 28 university level English learners in Taiwan. The students received two weeks of training followed by ten weeks of assessing in class. The cycle was then repeated. The training included practice assessing recordings with teacher feedback afterward, and self- and peer-assessment of oral presentations. Chen concluded that the training and practice did improve student accuracy making it more similar to teacher ratings in the second cycle.

**Time on Training**

Another aspect of self-assessment training that varied within the research was the amount of training provided. Whereas Taras (2003) and Butler and Lee (2010) only provided learners with an explanation, other training required considerably more time by both teacher and learner. Patri (2002) used a two-hour pre-training to prepare learners. Khonbi and Sadeghi (2013) had a similar pre-training while Chen (2008) had a much more extended training, spanning two weeks before the course followed by practice during the ten-week class itself. This cycle was even repeated a second time. Similarly, Schmidt and Wehmeyer (2016) and McDonald and Boud (2003) spread their training over 8 weeks (Schmidt & Wehmeyer) or 12 modules (McDonald & Boud).

Considering that the studies with the least training, only a short explanation before the assessment (Taras, 2003 and Butler and Lee 2010), saw the least improvement, while those with more substantial training (Chen, 2008; Khonbi & Sadeghi, 2013; McDonald & Boud, 2003;
Patri, 2002; Schmidt & Wehmeyer, 2016) saw larger improvement, one can infer that the more extensive the training the more accurate the calibration (no doubt with decreasing returns at some point). Nevertheless, training of any length does seem to improve calibration, and shorter trainings may be more practical for some purposes.

Motivation for the Current Study

This study draws on data collected using the Language Ability Self Evaluation Resource (LASER\(^1\)). The LASER is a low-stakes language self-assessment developed by language specialists at a large private university in the western United States. It was designed to better understand the speaking and writing proficiency of in-coming third year language students in their target language in an effort to make sure students were registering for the correct language class. The LASER uses Qualtrics experience management software and includes both a survey to better understand the learners’ language backgrounds as well as a self-assessment section containing 24 Can-Do statements. The Can-Do statements were created to represent defining aspects of language functions, accuracy, contexts, and text types at progressively complex proficiency levels as described by the ACTFL proficiency guidelines. Results from the LASER previous to 2019, highlighted the tendency among learners to over-assess their own language abilities. (T. Cox, personal communication, January 12, 2022).

Because of previous research on the benefits of training to improve calibration between the learner’s self-assessed (perceived) proficiency and their rater-assessed (actual) proficiency, the current study was designed to determine whether training could improve participant

\(^1\) The LASER was formerly called the Language Proficiency Diagnostic Assessment (LPDA). The name change was made for the Fall 2021 version.
calibration in the LASER. A short training module was added to the LASER in Fall 2021 that included explanation regarding proficiency as well as practice and immediate feedback. In an effort to keep the LASER accessible for both teachers wanting to use it and students needing to take it, the placement test has purposely been kept short. As such the training module was purposefully designed to add minimal time to the LASER with the goal being to increase the total length of the assessment by no more than 10 minutes. With that in mind, the study was guided by the following research questions:

1. What effect does a minimal (approximately 10 minute), training module (involving basic information about proficiency, examples and practice, and immediate feedback) have on the calibration of a language learner’s *perceived* and *actual* proficiency?
2. What effect does this minimal training module have on the calibration of learners of different proficiency levels?

**Methodology**

**Participants**

The participants in the present study were university level students enrolled in third-year world language courses Fall of 2021. Figure 1 displays the participants according to their target languages: French (*n* = 79, 19.3%), Italian (*n* = 21, 5.1%), Japanese (*n* = 11, 2.7%), Portuguese (*n* = 134, 32.8%), Russian (*n* = 21, 5.1%), and Spanish (*n* = 143, 35%) for a total of 409 students. There were 264 male and 145 female participants.
Nearly all the participants were native English speakers with the exception of 2 native Portuguese speakers, 1 native Spanish speaker and 3 bilingual native speakers of English and either Italian or Spanish.

Figure 2 displays an approximation of the time participants spent in the target language culture. All but 42 participants reported having spent time in the target culture. This self-reported time as well as how much access to native speakers the participant reported having during the that time were used to approximate the total number of immersion hours for each participant. These hours ranged from 180 to just under 4500 hours.
Figure 2

Approximation of Immersion Hours

Additionally, participants were asked to report the age that they began learning their target language. Figure 3 displays the distribution of participants across the age ranges they reported. Although the majority ($n = 337, 82.4\%$) started between the ages of 14 and 20, 8.1\% reported beginning their target language before the age of 10 ($n = 33$), and 7.4\% began between the ages of 11 and 13 ($n = 32$). Only 2.5\% reported starting to learn the target language after 20 years of age ($n = 7$).
Over half the participants reported not having received formal language training in school \((n = 216, 53.3\%)\) including many of the participants who started learning their target language between the ages of 17-20 \((n = 203, 56.2\%\) of total participants; 84.9\% of 17-20 group). It is likely that the large percentage of students in this category can be accounted for by acknowledging that most upper division language students at this private university spend one to two years in target languages around the world usually between the ages of 18-22. Often the learners have had no prior language training in the language of the target culture other than a 4-to 6-week intensive course before for they go abroad.

**Materials**

The materials for this study included the Language Self Evaluation Resource (LASER), the Oral Proficiency Interview – Computer (OPIc), and an explicit training module added to the LASER. The LASER contributed the *perceived* proficiency score using 24 Can-Do statements.
(12 speaking and 12 writing) of progressing complexity to allow learners to identify their own proficiency level. The OPIc, designed and administered by ACTFL, provided the actual proficiency score using trained raters to rate speech samples obtained through its internet-based proficiency test. The explicit training module was designed for this study and included two parts. The first part was designed to train learners on the ACTFL proficiency guidelines. The second part was designed to give learners an opportunity to practice the principles they just learned and receive immediate feedback. The additional material was designed to require approximately 10-15 minutes of additional time. Table 1 displays the differences between average duration spent taking the LASER for both groups. These data indicate that the experimental group did indeed spend more time on the LASER by an average of 11 minutes².

Table 1

Descriptive Statistics of Duration for All Groups.

<table>
<thead>
<tr>
<th>Duration (min)</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experimental</td>
<td>211</td>
<td>33.7</td>
<td>31.7</td>
<td>9.81</td>
<td>0.676</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>136</td>
<td>22.2</td>
<td>20.0</td>
<td>8.90</td>
<td>0.763</td>
</tr>
</tbody>
</table>

² There was a noticeable jump in the time participants took to complete the LASER at around one hour. It was found that often students were interrupted mid-test and returned to finish it much later. As such, tests that took longer than one hour were not considered in finding the average time it took for participants to complete the LASER. There were 136 participants who did not have any additional training and 211 participants who had additional training that finished in under one hour. Considering only those that finished within an hour, the groups differed in assessment duration significantly, by an average of 11 minutes, with a p value <0.001.
Figure 4 displays the confidence interval between the time on task between the two groups. Notice that there is no overlap between the groups and the average differs by approximately 11 minutes. Considering that the experimental group was the group that spent the additional time, this 11-minute difference aligns well with the intended 10-minute duration of the training module.

**Figure 4**

*Confidence Interval Between the Control and Experimental Groups Regarding Duration*

Explicit training on the ACTFL proficiency guidelines. The first part of the training module included explanations of the functions, text types, and accuracy associated with each ACTFL proficiency level. It also included a written sample of each level. These explanations were presented as video clips using Prezi video and communications software. There was a video clip for each level, and each video lasted just longer than one minute. The information about the ACTFL levels was presented visually in short written statements that described the function, text type and accuracy of the level. (See Appendix A for screen shots and audio scripts from the Training module). The module was presented to the participants directly prior to the start of the Can-Do statements.

**Practice and Feedback.** After each video clip the students were asked to consider how often they used that level complexity in their native language. They reported this using a 5-point
Likert scale which ranged from “never” (1) to “always” (5). After this short exercise, students were given the opportunity to test their understanding of the characteristics of each proficiency level and receive feedback. Using video and audio prompts and responses available on the ACTFL website, participants were asked to identify the appropriate level, function, accuracy, and/or text type for each sample. Choices were provided for the learner to choose from. As the samples were embedded into an H5P framework, learners were able to receive immediate feedback to their responses.

To further synthesize the information, each sample was paired with a sample of a different level. This allowed learners to compare speech samples of different levels with Superior and Advanced grouped together and Novice and Intermediate grouped together. After identifying the aspects of the ACTFL scale relevant to the video, the learners could then check their answers and receive immediate feedback including some additional explanation. The feedback reiterated key principles from the instructional portion of the training. (See Appendix B for screen shots of the Practice and Feedback module).

**Procedure**

All the participants received a link to the LASER from their language teacher and were asked to complete it as homework within the first several weeks of class. For the first week the test was open, all the students were assigned to the experimental group. After that, the students were randomly assigned into two groups: the control group – those who did not receive any additional training on proficiency; and the experimental group – those who received the additional 10-minute module on proficiency right before the self-assessment portion of the test. As shown in Figure 5, 147 students were randomly assigned to the control group (female $n = 63,$
male \( n = 84 \) while 262 were assigned to the experimental group (female \( n = 82 \), male \( n = 180 \)) for a total of 409\(^3\) participants.

**Figure 5**

*Gender Split Between the Control and Experimental Groups.*

As is displayed in Figure 6, both the control group and the experimental group started the LASER by watching a video explanation describing the purpose of the test, then completing a survey to gather personal information about the student: their age, gender, language learning background and experience with the language. At this point the experience of the two groups

\(^3\) The total number of participants in this study was 409 students. However, due to an oversight, random assignment was not used during the first 12 days of LASER administration. As such students were ONLY assigned to the experimental group for those days. After the oversight was identified and corrected, random assignment began and participants were assigned to either the experimental or control group. In order to maintain the integrity of the experiment, those students not randomly assigned were placed in their own group (preliminary). Although the preliminary group was removed from experimental analysis it is still included for informational purposes and the calibration of the participants in the preliminary group will be discussed as appropriate. After removing the preliminary group, the remaining participants were N=278.
diverged with the experimental group here receiving additional training in proficiency. The rest of the LASER was the same for both groups: first completing the 24 items to self-assess language proficiency and then responding to speech and writing prompts that were sent to their teachers. Within a month following their LASER assessment, all participants completed the ACTFL Oral Proficiency Interview – Computerized (OPIc).

Figure 6
Visualization of LASER Participants and Test Procedure
Data Analysis

The first research question asked to what extent an explicit training module based on introducing learners to the ACTFL proficiency guidelines and providing practice and immediate feedback improves learner calibration between perceived and actual proficiency.

The learners’ calibration score was calculated by subtracting the perceived proficiency from the actual proficiency. The LASER was used to assess the learners’ perceived proficiency. The responses for the LASER Can-Do statements were reported using a five-point Likert scale with 1 being “Not Confident” and 5 being “Very Confident.” Each answer was weighted according to the complexity of the item corresponding to what functions, accuracy, content and text type were required to complete it. For example, if a participant selected 1 – Not confident, for any item, they would receive zero points for that item. Whereas, if they selected 2, they received one point if the item was rated as simple, beginning level complexity, four points if the item was more complex, and six points if the item required the highest level of complexity. At the opposite end, if the participant selected 5 - Very Confident, for an item, they would receive two points for the simplest items, four points or seven points for more complex items, and nine points for the most complex items. These scores were then averaged to get the perceived proficiency score.

The learners’ actual proficiency was determined by converting their OPIc score to a 10-point scale with 1-3 representing Novice Low, Mid, and High; 4-6 representing Intermediate Low, Mid, and High; 7-9 representing Advanced Low, Mid, and High; and 10 representing Superior. For example, if the learner rated as an Advanced Low on the OPIc, their numerical score would be a 7. The calibration was then determined by subtracting the participants OPIc score from the LASER score providing either a positive score (indicating by how many sub-
levels the learner over-assessed), zero (indicating the learner was calibrated not over- or under-
assessing), or a negative number (indicating by how many sublevels the learner under-assessed
their own proficiency.). So, if the learner above had received an average perceived score of 8 and
then received a Advanced Low (7) on the OPIc, the OPIc score (7) would be subtracted from the
perceived score (8) giving the calibration score of +1. This positive number would indicate that
they overestimated their ability by one sublevel.

An independent t-test was used to measure the difference in calibration between the
control and experimental groups.

The second research question asked if the training module’s effect was related to the
participants’ actual proficiency level. To answer this question, the participants were analyzed by
proficiency level. An ANOVA was used to measure the difference in calibration between the
experimental and control group at each proficiency level.

Results

Removing Outliers

To stay consistent with the intended course populations, we found it necessary to remove
four outliers: two French speakers, one Italian, and one Portuguese. The data from one French
and one Portuguese participant were removed because they did not have ratable samples on the
OPIc making it impossible to compare their actual and perceived proficiency. The remaining
French speaker had an OPIc rating of Intermediate Low and the Italian speaker was rated at
Novice High. The objectives for second-year language students in French and Italian at this
university indicate that learners will be able to talk about the present in discreet sentences as well
as begin to narrate in all time frames and use some paragraph level speech. These last descriptors
are characteristics of the Advanced Level and are functions and text types that learners begin to
gain some mastery of in the Intermediate High sublevel. This means that learners entering third-year courses are expected to at least be moving into the Intermediate High from Intermediate Mid. As such, we determined Intermediate Mid to be a generous cutoff for the actual proficiency expected of third-year students at this university and removed any participant who was not able to sustain Intermediate level communication (i.e. anyone below Intermediate Mid).

Nearly half of the remaining participants were in the Advanced Low (AL) proficiency level (44.7%, \( n = 181 \)). As displayed in Table 2, Advanced High was the highest actual proficiency level found among the participants. Intermediate Mid, the lowest actual proficiency level (after removing outliers), had an equal population to Advanced High with both accounting for 4.2% of the entire population \( (n = 17) \). Note that no participants had an actual proficiency level of Superior.

<table>
<thead>
<tr>
<th>Levels</th>
<th>Counts</th>
<th>% of Total</th>
<th>Data Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate Mid</td>
<td>17</td>
<td>4.2%</td>
<td></td>
</tr>
<tr>
<td>Intermediate High</td>
<td>85</td>
<td>21.0%</td>
<td></td>
</tr>
<tr>
<td>Advanced Low</td>
<td>181</td>
<td>44.7%</td>
<td></td>
</tr>
<tr>
<td>Advanced Mid</td>
<td>105</td>
<td>25.9%</td>
<td></td>
</tr>
<tr>
<td>Advanced High</td>
<td>17</td>
<td>4.2%</td>
<td></td>
</tr>
</tbody>
</table>

### Group Populations

**Creation of the Preliminary Group.** As was noted previously, for the first 12 days of this experiment, due to an oversight in the survey settings, all participants were assigned to the experimental group (meaning they received training on proficiency). To maintain the integrity of
the experiment itself, these initial participants were split off into their own group, the preliminary group (Pre). The preliminary group population will be included in the group comparisons that follow but the results of the preliminary group calibration testing will be analyzed separately from the experimental group (Exp) and the control group (Control). This grouping is to maintain the validity of the experiment itself by keeping as many factors (time when they could take the test being important as is shown below) equal. The results of the preliminary group calibration will be reported at the end of the results section.

Actual Proficiency Levels Across Groups

The preliminary group had the highest average actual proficiency follow by the experimental group and then the control group. The actual proficiency of the preliminary group ($n = 134$), the experimental group ($n = 128$) and the control group ($n = 143$) was determined using OPIc scores. As displayed in Table 3, the preliminary group had more participants than either the experimental group or the control group that were rated in the highest levels of proficiency (Advanced Mid, AM, and Advanced High, AH). Conversely, the control group had more participants than either the preliminary group or the experimental group in the lowest levels or proficiency (Intermediate Mid, IM, and Intermediate High, IH). The number of participants in the experimental group was consistently between or equal to the numbers of the other groups.
Table 3

*Breakdown of OPIc Level by Group*

<table>
<thead>
<tr>
<th>OPIc ACTFL Levels</th>
<th>Groups_All</th>
<th>Data Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Exp</td>
</tr>
<tr>
<td>IM (5)</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>IH (6)</td>
<td>19</td>
<td>25</td>
</tr>
<tr>
<td>AL (7)</td>
<td>52</td>
<td>62</td>
</tr>
<tr>
<td>AM (8)</td>
<td>48</td>
<td>33</td>
</tr>
<tr>
<td>AH (9)</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>134</td>
<td>128</td>
</tr>
</tbody>
</table>

Unfortunately for the experiment, this creates an unequal balance of proficiency between groups with the preliminary group having 26 more participants in the highest two proficiency levels and the control group having 19 more participants in the lowest two levels of proficiency. All groups had the most participants in the Advanced Low (AL) proficiency level.

The lack of overlap in proficiency scores is made more clear in Table 4 which shows that with at least 95% confidence we can conclude that the preliminary group (with the lower parameter of 7.23) does not overlap with the experimental group (which has its higher parameter at 7.15) or the control group (which has its higher parameter at 6.93). We also see that there is an overlap in the confidence intervals of the experimental group and the control group (6.85 being the lower parameter for the experimental and 6.93 being the higher parameter of the control group).
Table 4

Estimated Marginal Means of Each Group

<table>
<thead>
<tr>
<th>Groups_All</th>
<th>Mean</th>
<th>SE</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td>7.38</td>
<td>0.0746</td>
<td>7.23</td>
<td>7.53</td>
</tr>
<tr>
<td>Exp</td>
<td>7.00</td>
<td>0.0763</td>
<td>6.85</td>
<td>7.15</td>
</tr>
<tr>
<td>Control</td>
<td>6.78</td>
<td>0.0722</td>
<td>6.64</td>
<td>6.93</td>
</tr>
</tbody>
</table>

Data Visualization

In addition, an ANOVA used to compare the three groups found that the populations were significantly different regarding actual proficiency. The difference between the average proficiency of the preliminary group and the experimental group \([t = 3.57, p = .001]\) as well as difference between the preliminary group and the control group \([t = 5.75, p = <.001]\) are shown in Table 5. While these differences were statistically significant, note that the overlap between in proficiency between the experimental group and the control group was not significant \([t = 2.06, p = 0.099]\).

Table 5

Post Hoc Comparison Between the OPIc Levels of All Groups

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Groups_All</th>
<th>Groups_All</th>
<th>Mean Difference</th>
<th>SE</th>
<th>df</th>
<th>t</th>
<th>p_hkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td>-</td>
<td>Exp</td>
<td>0.381</td>
<td>0.107</td>
<td>402</td>
<td>3.57</td>
<td>0.001</td>
</tr>
<tr>
<td>Pre</td>
<td>-</td>
<td>Control</td>
<td>0.597</td>
<td>0.104</td>
<td>402</td>
<td>5.75</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Exp</td>
<td>-</td>
<td>Control</td>
<td>0.217</td>
<td>0.105</td>
<td>402</td>
<td>2.06</td>
<td>0.099</td>
</tr>
</tbody>
</table>

Note. Comparisons are based on estimated marginal means
Overall, these results indicate that the preliminary group was very different from the other two in actual proficiency while the experimental group and the control group, though not as similar as we would like, were more similar to one another in actual proficiency than either is to the preliminary group.

Aside from the need to create an additional group, this imbalance in actual proficiency also contributes to how we will analyze further data, which will include comparing results by OPIc level as well as by group.

**A Noted Systematic Difference.** In analyzing these results, it became clear that there was a difference between participant proficiency level correlated to when they took the LASER. The participants who took the LASER near the beginning of the 35-day window in which it was available were more likely to have higher actual proficiency scores. Conversely, those participants who took the LASER near the end of the 35-day window in which it was available were more likely to have lower actual proficiency scores. This trend was confirmed statistically, showing there was significant negative relationship (*Pearson's r* = -.24, *p* < .001) with a weak effect size. In other words, *on average*, the further from the LASER availability date a participant took the LASER, the lower their actual proficiency level. Specifically, the average actual proficiency in the beginning of the testing period was approximately Advanced Mid while the average actual proficiency in the end of testing period was closer to Intermediate High. While how quickly participants choose to start the LASER was not a part of this study, it is interesting to note and some possible reasons for this finding will be discussed later.

All languages followed this trend, as shown in Figure 7. Japanese had the steepest decline.
This correlation between the time a participant took the LASER and their proficiency level was particularly problematic in that, as noted previously, for almost the first 12 days of testing, participants were only assigned to the experimental group. As such, many high proficiency speakers were removed from the calibration comparisons of the experiment (being placed into the preliminary group), leaving the experimental group and the control group to have significantly lower proficiency (statistically).

**Perceived Proficiency across groups**

The largest number of participants *perceived* their proficiency to be at the Advanced Mid proficiency level \( n = 105, 25.9\% \) with the numbers tapering off the further away from Advanced Mid the participants rated themselves. This fairly regular distribution is displayed in Table 6. However, as noted above in Table 3, Advanced Low (AL) was the *actual* proficiency level with the largest population \( n = 181, 44.7\% \). This is an immediate indication that participants, in general, are over-assessing their proficiency.
Table 6
*Frequencies of Self-Assessment by ACTFL Sublevel of Entire Population*

<table>
<thead>
<tr>
<th>Levels</th>
<th>Counts</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL (4)</td>
<td>2</td>
<td>0.5%</td>
</tr>
<tr>
<td>IM (5)</td>
<td>32</td>
<td>7.9%</td>
</tr>
<tr>
<td>IH (6)</td>
<td>56</td>
<td>13.8%</td>
</tr>
<tr>
<td>AL (7)</td>
<td>96</td>
<td>23.7%</td>
</tr>
<tr>
<td>AM (8)</td>
<td>105</td>
<td>25.9%</td>
</tr>
<tr>
<td>AH (9)</td>
<td>73</td>
<td>18.0%</td>
</tr>
<tr>
<td>Sup (10)</td>
<td>41</td>
<td>10.1%</td>
</tr>
</tbody>
</table>

When looking within the groups, the highest self-assessed proficiency level on average was found in the experimental group, followed by the preliminary group, leaving the control group rating themselves lowest as shown in Table 7. Note that the preliminary group peaks at Advanced Mid (AM, n = 43), while the experimental group (n = 34) and the control group (n = 30) both peak at Advanced Low (AL). However, despite the fact that the preliminary group had the highest *actual* proficiency, the experimental group and the control group both had more participants rating themselves as Advanced High (n = 25, 27 respectively) than the preliminary group. Noting these differences, it is clear that the three groups are not calibrating their *perceived* and *actual* proficiency the same way. Also note that each group had several participants rating themselves as Superior (Sup) (n = 15, 17, 9 respectively) even though no student received that rating on the OPIc.
Table 7

*Frequencies of Self-Assessment by ACTFL Level across the Population*

<table>
<thead>
<tr>
<th>ACTFL Levels</th>
<th>Groups_All</th>
<th>Data Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Exp</td>
</tr>
<tr>
<td>IL (4)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>IM (5)</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>IH (6)</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>AL (7)</td>
<td>32</td>
<td>34</td>
</tr>
<tr>
<td>AM (8)</td>
<td>43</td>
<td>33</td>
</tr>
<tr>
<td>AH (9)</td>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td>Sup (10)</td>
<td>15</td>
<td>17</td>
</tr>
</tbody>
</table>

Additionally, the estimated marginal means shown in Table 8 indicate that with at least 95% confidence we can conclude that there is a large overlap in the self-assessment ratings between the preliminary group and the experimental group, but the control group does not overlap with the experimental group at all (7.50 upper margin to 7.62 lower margin) or the preliminary group (7.50 upper to 7.51 lower margin).
An ANOVA verified these results indicating that the groups were significantly different \( F = 6.98, p = 0.001 \) when it came to self-assessment scores. The post hoc comparisons further showed that the significant difference is between the control group and either of the other groups as shown in Table 9. Note that the comparisons show the experimental group and the control group to be more different \( t = 3.483, p = 0.002 \) than the preliminary group and the control group \( t = 2.853, p = 0.013 \).
Table 9

*Post-Hoc Comparisons of Self-assessment Ratings*

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Groups_All</th>
<th>Groups_All</th>
<th>Mean Difference</th>
<th>SE</th>
<th>df</th>
<th>t</th>
<th>p&lt;sub&gt;tukey&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre - Exp</td>
<td>-</td>
<td>Exp</td>
<td>-0.113</td>
<td>0.173</td>
<td>402</td>
<td>-0.653</td>
<td>0.791</td>
</tr>
<tr>
<td>Pre - Control</td>
<td>-</td>
<td>Control</td>
<td>0.481</td>
<td>0.168</td>
<td>402</td>
<td>2.853</td>
<td>0.013</td>
</tr>
<tr>
<td>Exp - Control</td>
<td>-</td>
<td>Control</td>
<td>0.594</td>
<td>0.170</td>
<td>402</td>
<td>3.483</td>
<td>0.002</td>
</tr>
</tbody>
</table>

*Note.* Comparisons are based on estimated marginal means

**Calibration**

The real consideration of this experiment is how proficiency training affects the calibration of a learner’s proficiency self-assessment. At this point we will remove the preliminary group from comparison with the randomly assigned experimental and control groups.

In comparing the self-assessment calibration of the experimental group and the control group, it is evident that the control group was better calibrated than the experimental group. Table 10 displays the difference between learner calibration in the experimental group and the control group. A zero (0) calibration means that the participant matched their *perceived* proficiency level (LASER) with their *actual* proficiency level (OPIc score). A positive integer indicates how many sublevels above their *actual* proficiency the learner rated their proficiency. A negative integer indicates how many sublevels below their *actual* proficiency the learner rated their own proficiency. Notice that the control group has more participants either accurately self-assessing (n = 42, 29.4%) or under-assessing (n = 33, 23%) than the experimental group (n = 25, 19.5% and 22, 17.2% respectively). Whereas, the experimental group had more participants overrating their proficiency (n = 81, 63.3%) than the control group (n = 68, 47.6%). The data
visualization clarifies that whereas the control group peaks at zero (perceived and actual proficiency are aligned), the experimental group peaks at overrating by one sublevel.

Table 10

*Frequencies of Speaker Calibration by Sublevel Between Experimental and Control Groups*

<table>
<thead>
<tr>
<th>Calibration</th>
<th>Groups</th>
<th>Exp</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-2</td>
<td></td>
<td>2</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td>20</td>
<td>21</td>
<td>41</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>25</td>
<td>42</td>
<td>67</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>40</td>
<td>32</td>
<td>72</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>29</td>
<td>25</td>
<td>54</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>12</td>
<td>11</td>
<td>23</td>
</tr>
</tbody>
</table>

The observation that the control group is calibrating differently was confirmed using a t-test that showed slight significance \( t = 2.00, p = 0.046 \). However, it should be noted that the effect size is small at .243. The small effect size is a reminder that most learners are only mis-calibrated by one or two sub levels. Even trained raters have been known to disagree at the sub-level, so this minimal mis-calibration is actually quite good. However, the fact that participants with training were, on average, less calibrated than those without training is both unexpected and worth exploring further.

The Effect of OPIc Level and Training on Calibration

As we have shown already, proficiency level plays an interesting role in this experiment especially as the groups were not evenly matched in proficiency. An ANOVA used to compare the multiple variables of actual proficiency level with group (and by extension training) showed
that proficiency level is a very significant factor in how calibrated a participant was \(F = 3.823, p = 0.005\). Figure 8 displays two aspects of how proficiency interacted with calibration. First, it shows that lower and higher proficiency levels are more calibrated. This is seen as the circles representing the means for Intermediate Mid, Advanced Mid and Advanced High of both groups are closer to alignment (0), than Advanced Low. In addition, this graph also shows that at every proficiency sublevel, the control group has more accurate calibration than the experimental group (the circle indicating the control group mean is closer to zero). This is particularly noticeable at the High sublevels where the experimental group jumps further from alignment (0) than the control. Indicating that learners in the High sublevels are more likely to overestimate their abilities even more than might be found at other sublevels.

**Figure 8**

*Plot of Calibration by OPIc Rating Compared Between Experimental and Control groups*

It should be noted that there were low or very low numbers of participants in the Intermediate Mid \(n = 15\) and Advanced High \(n = 4\) proficiency levels compared to the number of participants at other levels. This likely accounts for the larger confidence intervals and requires careful consideration as to the generalizability of the results for those specific levels.
As evident in Figure 8 above, each sublevel had a slightly different calibration between groups. However, as shown in Table 11, only the difference between participant calibration at the Intermediate High level was statistically significant \( t = 2.62, p = .011 \).

**Table 11**

*Results of Independent T-tests Comparing Calibration by Proficiency Level*

<table>
<thead>
<tr>
<th>Calibration Level</th>
<th>Statistic</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration_IM</td>
<td>Student's t</td>
<td>0.624</td>
<td>13.0</td>
</tr>
<tr>
<td>Calibration_IH</td>
<td>Student's t</td>
<td>2.62</td>
<td>64.0</td>
</tr>
<tr>
<td>Calibration_AL</td>
<td>Student's t</td>
<td>1.24</td>
<td>127</td>
</tr>
<tr>
<td>Calibration_AM</td>
<td>Student's t</td>
<td>0.666</td>
<td>55.0</td>
</tr>
<tr>
<td>Calibration_AH</td>
<td>Student's t</td>
<td>1.00</td>
<td>( ^* )</td>
</tr>
</tbody>
</table>

**Calibration of the Preliminary Group**

Overall, the preliminary group was different in calibration from what we found in the control group and the experimental group. As displayed in Figure 9, the preliminary group had a calibration plotline that had similarities to both the experimental group and the control group while also being very different from both as well. Some of these differences are likely explained by the varied number of participants at different proficiency levels within each group. For example, the preliminary group had less than a third as many participants than the experimental group or the control group in Intermediate High \( (n = 2) \) and more than six times the number of either other group in the Advanced High group. Because the preliminary group had so few participants at the Intermediate High level, we cannot be certain we are seeing generalizable results. This is displayed as a very wide confidence bar in the graph. This may account for why the average of Intermediate High in the preliminary group is much higher than the average of the
others. In addition, since the preliminary group has a much higher number of participants in the Advanced High level (correlated to the much smaller confidence interval bar when compared to the experimental and controls groups) we can likely generalize the results from the Advanced High preliminary group better than that of the other Advanced High plots. As for the other three sublevels, the preliminary group, with similar numbers of participants was closer to the control group on Intermediate High and Advanced Mid and closer to the experimental group on Advanced Low.

Although higher proficiency correlates with better calibration (the learner aligns their \textit{actual} and \textit{perceived} proficiency better), it is interesting to note that, despite the learners in the preliminary group having higher \textit{actual} proficiency, their calibration was a less accurate than those with the lowest \textit{actual} proficiency (the control group). This is another indication that the proficiency training did more immediate harm than good.

\textbf{Figure 9}

\textit{Plots of Calibration by OPIc Score between Groups}
Conclusion and Discussion

Although unexpected, the results from this experiment are very insightful. They are not conclusive as to how or how not to train learners to self-assess, or even on whether training can truly benefit learner calibration. However, they do provide motivation for reflection on potential weaknesses of self-assessment training and emphasize the importance of pilot testing and data analysis. In particular, it is of interest to discuss the effect the training had on participants of different proficiency levels.

Proficiency and Calibration

As was seen in previous research, proficiency did play a role in calibration, showing that learners at the higher end of the ACTFL proficiency scale and to some degree those at the lower end were better calibrated than those with proficiency in the middle range- displayed above in Figure 9 (Boud & Falchikov, 1989; Brown et al., 2014; Ma & Winke, 2019; Piñana, 2009). Given this connection between actual proficiency (found using the OPIc) and calibration, the groups with higher proficiency, the preliminary group and the experimental group, would have been expected to be better calibrated. However, the results proved otherwise. In addition, as was shown in Table 4, despite both the preliminary group and the experimental group having higher overall proficiency ratings than the control group, both experimental groups were less calibrated than the control group. The preliminary group, with the highest proficiency of all, was only able to assess almost as accurately as the control group. In effect, the training undermined the ability of the experimental group to accurately self-assess. Observing such, careful planning when providing training becomes crucial so as to not unknowingly make the ability to correctly calibrate language self-assessment worse for the learners.
Moreover, although there were differences in the group composition other than proficiency (such as the number of participants from each language) none of those factors were found to be significant unless connect with the OPIc (proficiency) ratings (See Appendix D for more on groups comparisons). As such, training seems to be the reasonable cause for the difference in calibration. As Table 7 shows, this was especially relevant for the experimental group at the High end of the major levels—Intermediate High and Advanced High. Consistently, more participants in the experimental group self-assessed in the Advanced Low sublevel when they should have been at the lower, Intermediate High. Whereas the control group was the opposite with more participants likely to self-rate into the Intermediate High level instead of the higher Advanced Low. This is a problematic split in proficiency as learners will be doing some of the functions and text types from both groups. Being able to clearly determine whether you are peaking into the next level or dipping down into the lower level is no easy task, even for a trained rater. It seems a place in particular where limited training would fall short.

One possible explanation is that the minimal training mixed with limited practice and feedback allowed the learners to justify their desired perception of their own ability. These findings and considerations seem to indicate statistically that broad and limited training on ACTFL proficiency should be avoided as a way to find quick gains in learner self-assessment calibration. However, the long-term washback effect, the practical effect, of such a training is still unknown. And certainly, we must allow for future gains as the learner continues working in the language and note that what the learner was made aware of in this short training may later become more concrete in the learner’s mind and lead to greater calibration.
It is also not certain what effect this limited training module would have on peer-assessments but, given previous findings, the results are likely to be different and potentially more beneficial (Khonbi & Sadeghi, 2013; Patri, 2002).

A note here to remember that we are talking about sublevels, so the effect size of being off one sublevel is very small. Also, it is interesting to note that learners were not trained on the nuances of sub levels and the LASER itself did not require them to specifically assign themselves to a level at all, but rather to answer how well they might complete a task that happened to be at that level. As was described above, these Can-Do statements were built based on the FACTS of the four ACTFL proficiency levels. The learners’ weighted scores (determined by personally indicating how well they thought they could do a task) were added up and averaged to figure the overall proficiency score. How certain a learner rated themselves as capable of completing the FACTS of a major level led to what sublevel they were figured into for the self-assessment portion. It might have been interesting to have them assess themselves from the training alone as well.

**Limitations and Future Research**

The results of the experiment were limited in several ways. Two limitations have already been mentioned. The first was the lack of randomization for the first 12 days of testing and the second was the dissimilarity between groups with regard to *actual* proficiency level. In addition to these limitations there are several limitations regarding the training module that should be discussed.

**The Training Module**

First, the actual participation or quality of participation of the experimental group is not really known. On average, we know that the participants in the experimental group spent longer
on the LASER as a whole than those in the control group, but we have no way to confirm how much of that time was spent on the initial demographic information, on the module, or on the Can-Do statements. Even averaging the additional time at just over 10 minutes for the modules, we still wouldn’t know whether they watched all or even some of the videos or completed any of the H5P activities. The data displayed in Figure 10 suggests that some participants, in particular, likely did not effectively complete the training module as their total test time was less than the minimal time it would take for participants to watch every video in the training. However, the videos could be watched at a faster speed.

**Figure 10**

*Test Duration in Minutes for the Preliminary, Experimental and Control Groups*

There is some indication of interaction with the training module for many of the participants in the experimental group through post video follow-up questions. These questions asked participants to rate how often they used the functions and text types of that proficiency level in their native language. There was an additional question at the end of the module asking for general comments. The participant responses were insightful, but also raised questions, that lead to a second limitation: the uptake from the design of the module itself.
The module design limitation relates to both the accessibility of the information in the module as well as the ability of the participants to perform the tasks in the module – such as to practice and receive feedback. As for the former, many participants did not seem to pick up the overarching ideas behind the proficiency levels. This was manifested in many participants indicating that they “never” use Novice level functions and text types or “seldom” use Intermediate level functions and text types in their native language as shown in Figure 11. Similarly, the majority of learners indicated that they often or very often used Advanced or even Superior level language functions and text types in their native language. As Intermediate level language functions are the interactions most typically used by natives and non-natives, it seems clear that the participants were conflating the improved ability to speak a language with the moving up in proficiency levels without particular attention to the actual functions and text types that proficiency level required. In other words, instead of being a descriptor of language production, the proficiency levels were a descriptor of intelligence or time on task. (A problem that, I would argue, is common among language learners and is not helped by the possibly negatively nuanced names of the first two ACTFL proficiency levels–Novice and Intermediate).

**Figure 11**

*Participants Self-reporting Use of Specific Proficiency Levels in Their Native Language*

It should be noted that not everyone in the experimental group answered these questions. In addition, we do not know how carefully they were answering them. This again calls into
question the first limitation of how carefully and thoroughly the participants went through the training module.

In addition, many comments were made indicating that there was a malfunction in one of the H5P activities, making it impossible to drag and drop into one of the fields. Despite being checked numerous times before students began taking the LASER, there was indeed a problem with that activity. The participants were still able to drag into the Novice area, but not into the Intermediate area, so in effect they could still check their answers as half were in one and half in the other. However, the consequences of the malfunction (either positive or negative) we cannot know.

Yet another possible limitation in the training module was due to the choice to use the ACTFL website videos that show non-native English speakers completing tasks at different proficiency levels. In watching the Advanced and Superior level videos it has been commented that the accent of the speakers distracts learners from paying attention to the FACTS mastered by the speaker. Instead, the learner could use their own difficulty to understand the foreign accent as justification to raise their own abilities to those higher levels. Accent has been found to affect comprehensibility especially in relation to prosodic proficiency (similarity in the changes of pitch and flow with a native speaker) (Derwing & Munro, 1997). Native perception of accented speech has even been studied and in the effort to determine whether perception can be improved (Kasparek, 2008). With this in mind, future training may be improved by avoiding heavily accented speakers. Alternately, helping the participants understand the role that accent plays in language proficiency might have led to more accurate understanding of the proficiency levels.

A final limitation in the training module is due to the information contained in the module and the wording. One phrase, in particular, indicated that often language majors are
Advanced level speakers. This could have been taken to mean that if the participant is a language major, they are likely at the advanced level. Any student who is a language major could have been swayed to answer other questions inaccurately in order to align them with this claim. In addition, for a population so likely to have had immersion experiences that included the high-level functions of description, hypothesis and providing support for opinions on limited topics, there likely needed to be more emphasis not on the ability to attempt the task, but on completing it fully, that is with the appropriate control of the language, including cultural understanding, appropriate vocabulary, accuracy and text types.

The LASER

Since the LASER is a low stakes test, used mainly to either check that students are in the correct course or inform teachers on student needs, it follows that there is not a large draw to students over-assessing their own abilities. However, there are also several limitations regarding the LASER that should be discussed as they may encourage learners to over- or under-estimate their proficiency. The first limitation in this category stems from the collection of information about the participant before the assessment. Researchers have found (R. Erickson, personal communication, January 25, 2022) that learners respond to questions differently if they have already provided personal information. For example, if a participant stated they are female, they may answer questions in a way that they think is expected (or not expected) for a female as opposed to what they really think. (Note, genders were compared with no significant difference. See Appendix D for more information.) These findings could apply to the LASER in that students provided a large amount of information about their language background. Besides gender, this included how old they were when they started learning the target language, how many years they studied the target language in school and how long they have been in immersive
target culture settings. It is possible that they then felt the need to answer the self-assessment questions in a way that they felt would be right for that amount of time spent in the language, not wanting to admit what they couldn’t do. The training module may have increased their likelihood to over-assess in this case as it could have encouraged them to think of themselves further along on the proficiency scale due to time on task alone.

Additionally, as a possible second limitation in this category, it should be noted that the Can-Do statements were grouped by proficiency level. In consideration of the previous point, if learners wanted to think of themselves at a particular level, providing them with the knowledge of what questions belong to each level could have made it more difficult for them to answer truthfully. If they think they should be a certain level, whether due to the time on task or some other reason, telling them which tasks they should be able to complete in order to be at that level is likely problematic. This is of particular interest considering the limitation above regarding the wording used to describe the Advanced level in relation to language majors.

Also, the review of the ACTFL guidelines before that level’s questions may have reminded the participants of what they want to say about their language ability instead of just reminding them what the FACTS were. In addition, as mentioned above, the use of the names of the ACTFL proficiency levels could be problematic, as stated above, since the names could be considered negative by a learner. For example, a learner who has spent considerable time in classes and/or immersive settings might consider it negative to be referred to as Novice or Intermediate. They may think of themselves as advanced (not by ACTFL standards, but by the regular definition of the term as one who has spent a considerable amount of time on something).

4 Grouping the Can-Do statements by proficiency level was the choice of the designers of the Fall 2019 LASER.
Can-Do Statements

The wording of the Can-Do statements makes up a final limitation. As was indicated earlier (Adams & King, 1995), if a learner is unsure of what is being assessed, self-assessment is not very accurate. Tigchelaar et. al (2017) found that learners do not accurately self-assess when a Can-Do statement is vague, dependent on experience, or uses multiple skills. Ten of the 24 Can-Do statements used in the LASER arguably fit into one or more of these categories. Nine of the 10 are in the Advanced and Superior levels. (See Appendix C for a table of the Can-Do statements and possible weaknesses).

Less effective Can-Do statements could account for at least some of the inaccurate calibration as the participants could have been unclear as to their ability to fulfill the task in question.

For example, the second Can-Do statement in the Intermediate section reads: “I can ask and answer questions of a new friend to explore what our family members do for their job.” In this statement there are actually two skills being measured. First, whether the participant can ask questions and secondly whether the participant can answer questions. It is also unclear whether the learner can understand the answers to the questions that he or she asks the new friend, or if uttering the question is the only requirement. This Can-Do statement could be seen to require not only the grammar for both asking and answer questions, but also the skills required to adequately speak and listen/understand. If a participant feels confident asking questions, but not confident in understanding the response from the friend, that participant might be unsure how to respond to the statement and inaccurately represent his or her language abilities.
Tigchelaar et al (2017) argue that items focused on multiple skills have different degrees of difficulty and likely different degrees of mastery for the participant and should therefore be avoided.

Another example comes from the second Superior level Can-Do statement: “I can critically evaluate and expound upon the impact of government-sponsored daycare programs on family life and its economic implications in a public forum such as a Parent-Teacher Association or school board meeting.”

This statement is experience-dependent in that a participant who has never thought about government-sponsored daycare or been to a public forum might struggle to know how to answer. The participant would need to consider whether they know the specific vocabulary, whether they know anything about the topic, or even whether they know the protocol for presenting in a public forum. Most students at this American university, around the age of 20, would not have had experience with government-sponsored daycare, especially since in America government-sponsored daycare is uncommon if not unheard of. If the participant has never had experience with the experience side of the statement, the participant might struggle to know how to answer. Tigchelaar et al (2017) argue that there is a need to provide descriptors that are appropriate for the age level of the test takers. Out of 24, there are arguably seven experience-dependent Can-Do statements that could be modified to become more inclusive for students of this population.

It is likely that if the Can-Do statements were assessed and re-worked to avoid the pitfalls mentioned, participant self-assessment calibration would be more accurate.

As mentioned above, although general training on proficiency might not be helpful, perhaps focusing instead on Can-Do statements, might be better, allowing the participants to
better understand the functions, accuracy, content, text types and contexts described in the Can-Do statements themselves to better recognize their abilities and limitations.

Having expressed this limitation it is important to note that the internal consistency of the statements is .94. This indicates that the higher-level statements are indeed requiring different and more difficult language from the participants. In fact, a Rasch analysis showed that only three items misaligned (T. Cox, personal communication, Jan 12, 2022). The reliability of the statements may in fact mean that changing them would make no difference. It may also, considering learner over-estimation, be possible that although a statement has been found to certainly be one of the most difficult, if it is vague, experience dependent, or requires multiple skills, learners who cannot truly complete the task might sometimes convince themselves they can.

**Implications for Practitioners**

One implication of this study is to remind language instructors that learners struggle to accurately rate their own language abilities. However, enabling autonomous learners has been shown to be extremely beneficial (Engelhardt & Pfingsthorn, 2013; LeBlanc & Painchaud, 1985). As part of this, teaching language learners about proficiency has become much more common in recent years. In presenting proficiency, language instructors should be careful to present it accurately and then continue to train and provide feedback to learners about how calibrated they are. Using objective-based Can-Do statements consistently would likely also help to accustom learners to the assessment process. Instructors should also be aware that self-assessment for learners who are at the boundaries of a major level is more problematic.
Future Research

There are several areas where further research would be beneficial. First, with regard to the LASER, it would be helpful to understand more about why participants who took the LASER near the beginning of the availability window had higher average proficiency scores on the OPIc (on average). One possible reason would be learner motivation level. Other scores in the LASER, such as the GRIT and willingness to communicate scores could be informative in this. Comparing the Fall 2022 results with those of other semesters would also provide greater understanding as to whether this is a consistent trend in the LASER and perhaps be generalizable to other self-assessments that have flexible start dates.

Adjusting the LASER Can-Do statements would also provide insight into the creation of self-assessment items. Whether there was or wasn’t improvement in calibration after clarification and other alterations would inform others trying to create strong items. Another project could include training to help participants understand what it would look like to complete the specific can-do statements. Or students could be asked to actually try the can-do statement first (maybe even recording themselves) after which they are asked to tell how difficult or easy it was.

In response to the possible issues with the ACTFL videos, it could be useful to find videos of speakers at each proficiency level for the actual target language of the participant. Testing whether they would be more likely to calibrate their proficiency having seen someone do what they are being asked would provide insight into whether experiencing the accents and other obvious qualities of the native language interferes with the learner’s ability to recognize the characteristics that effect their own proficiency.

More general to self-assessment training, finding an optimal balance between time spent on training and gains in learner calibration would be beneficial. Exploring how much time is
needed and the effect of repeated trainings would be informative. Also, identifying the washback from the LASER and training module would be insightful. Were learners able to apply the principles from the module in their language learning later on?

In addition, determining the factors that are most advantageous to train on to improve calibration is improving to the usefulness of self-assessment. It would be insightful to train in-depth on one particular FACT to identify which were particularly salient to learner calibrating their proficiency. For example, if focusing on text type, it would be interesting to measure how greater focus on what a particular text type really includes, and what learner ability to maintain that text type really looks like would affect calibration.

In addition, most of the studies cited above, whether they had training or not, were done in classes of English as a second or foreign language (Blue, 1988; Butler & Lee, 2010; Chen, 2008; Engelhardt & Pfingsthorn, 2013; Khonbi & Sadeghi, 2013; Patri, 2002; Schmidt & Wehmeyer, 2016; Shleykina, 2020; Taras, 2003). Most of those were college level classes with the exception of Butler and Lee (2010) who worked with sixth grade students. Several other articles focused on non-language learning classes completely (Logan, 2015; McDonald & Boud, 2003). Only four (Brown, Dewey, & Cox, 2014; Dlaska & Krekler, 2008; Lappin-Fortin & Rye, 2014; Ma & Winke, 2019) deal with language learners studying languages other than English. The languages represented include Chinese, French, Spanish, Russian and German. Although ESL and world language classrooms are similar, it would be beneficial to have more research on self-assessment and explicit training for it within the larger world language context. More research on self-assessment in the world language classroom would likely offer new insight into the usefulness of self-assessment in language classrooms.
References


The training module contained the following slides and audio divided into 4 videos presentations. The first presentation started with a general overview of the ACTFL proficiency scale.

**Figure A1**

*ACTFL Proficiency levels slide*

This explanation was accompanied by the following narration: “Let’s look at the 4 stages of language development! They are Novice, Intermediate, Advanced and Superior. As each level is described notice: 1. What tasks can a learner do? 2. What does the task require? and 3. What or who does a learner rely on to complete those tasks?”
The Novice level slide was accompanied by the following narration: “Novice level: the tourist who learned phrases from a book. Novice level language learners communicate by recalling rehearsed material. Their responses are usually limited to words and phrases that they have learned in a class, from a book, or on an app. Novice speakers are very hard to understand and often rely on visuals, like gestures and pictures, as well as helpful people who understand learner errors, like language teachers.”
**Figure A3**

*Novice Level Writing Sample*

The Novice level writing sample was accompanied by the following: “A tourist who has minimal knowledge of the language might respond to an SMS text asking what to buy at the store by writing the following:”

**Figure A4**

*Intermediate Level Slide.*

The Intermediate level slide was accompanied by the following: “Intermediate level: the foreigner who speaks their native language at work or school. Intermediate level language
learners use their increasing understanding of grammar and vocabulary to express personal
meaning about daily life that goes beyond what they have been taught to say or write. In this way
they create more personalized meaning by adapting, combining, and recombining known
elements. Many daily tasks are at the intermediate level, even for native speakers. The
intermediate level learner uses statements and questions that are simple most often in the form of
sentences or strings of sentence that do not use transition words to connect them. Despite relying
on familiar vocabulary and word forms, an intermediate level learner often needs to repeat or
reformulate ideas to be understood. Speaking with or writing to people familiar with learner
errors is also helpful.”

**Figure A5**

*Intermediate Level Writing Sample.*

The Intermediate level writing sample was accompanied by the following: “A foreigner
who uses the target language to complete daily tasks but not at work or school might write a
review of tortillas that looks like this.”
The Advanced level slide was accompanied by the following: “Advanced level: the language major who has spent a substantial amount of time abroad. Advanced level language learners are characterized by their ability to use interconnected details as they describe and narrate what is happening, has happened, and will happen in the world around them. They are usually limited to topics of personal or general interest and are able to easily work around complications that arise within those realms. Advanced learners connect their ideas and questions using transition words and logical sequencing. As a result, their speech and writing take the form of written or oral paragraphs. Despite some errors, they have control of the basic structures of the language. This control, in addition to practiced communicative strategies and generic vocabulary, allow Advanced level learners to be understood without undue difficulty or effort on the part of their conversation partner.”
The Advanced level writing sample was accompanied by the following: “The language major who has spent a lot of time in a country where they have been required to speak the target language might write this entry on their food blog:”

The Superior level slide was accompanied by the following: “Superior level: The CEO of a large company based in a foreign country who transacts business in the foreign language.”
Superior level language learners use both cognitively and linguistically complex language to discuss, hypothesize and support both real world AND abstract ideas fully and effectively. Doing this requires that a learner move beyond single paragraphs and use extended discourse, connecting and separating their ideas with more precise vocabulary and rhetorical structures to ensure a logical progression. The Superior level learner relies on their own fluency and accuracy, projecting ease and showing no pattern of error despite the use of complex language forms and nuanced vocabulary. Any non-native patterns that emerge, do not distract from their message.

**Figure A9**

*Superior Level Writing Sample*

The Superior level writing sample was accompanied by the following: “The CEO might write a report that reads in part like this:”
APPENDIX B

Practice and Feedback module

Figure B1

Novice and Intermediate Practice and Feedback. Text Type, Accuracy, and Level
Figure B2

Novice and Intermediate Practice and Feedback and Answers

Figure B3

Novice and Intermediate Practice and Feedback with Explanations
Figure B4

Advanced and Superior Level Practice and Feedback. Text type, Accuracy, and Level.

Figure B5

Advanced and Superior Level Practice and Feedback. Immediate Feedback.
Figure B6

Advanced and Superior level practice and feedback exercises. With Explanations.

Understanding the ratings:

The Advanced-Level speaker is able to connect his sentences using transitions words and phrases (first, in, then). He uses generic vocabulary as well as some specialized words, but all are very common for his interests (electrical engineering, image processing). Although he does make some mistakes (engineering instead of engineer), he has control over the basic structures of English grammar and speaks about the past without much concern.

The Superior-Level speaker is able to communicate with ease and accuracy in order to participate fully and effectively in a conversation covering several aspects of the topic from both an abstract and a concrete perspective. The speaker provides structured arguments to support his opinions (education should be holistic) and construct his hypotheses. The speaker uses extended discourse to make his points. There is no pattern of errors in basic structures and the occasional structural and phonetic errors do not interfere with the communication or distract the listener.

Figure B7

All Levels by Prompt Exercise. Identification of Functions and Levels
Figure B8

All Levels by Prompt Exercise with Immediate Feedback

2. Drag the correct Level:

1. Drag the correct Description:
## APPENDIX C

### Can-Do Statements Considerations

#### Table B1

*LASER Speaking and Writing Can-Do Statements with Categorization*

<table>
<thead>
<tr>
<th>N1: I can name the members of my family to a new friend.</th>
<th>Vague</th>
<th>Experience dependent</th>
<th>Multiple skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>N2: I can tell my friend the types of activities I do in class/at work.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N3: I can write a shopping list of supplies needed for school or work.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N4: I can fill out a form with biographical data (age, gender, etc.) about my family.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i1: I can describe my family's hobbies to a new friend.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i2: I can ask and answer questions of a new friend to explore what our family members do for their job.</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>i3: I can tell my friend what a typical school/work day is like from beginning to end.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i4: I can ask a teacher/boss a series of questions to clarify how to complete an assignment.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i5: I can write a short birthday message to a family friend.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i6: I can send an email asking a series of questions about an upcoming activity my family will participate in.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i7: I can send text messages to arrange a meeting with a classmate/co-worker.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i8: I can email a classmate/co-worker and ask a series of questions to find out what I missed when I was sick.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1: I can share a long, detailed story with a new friend about a memorable childhood experience such as a favorite vacation or holiday.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2: I can tactfully bring a problematic issue to the attention of a customer service representative (hotel clerk, travel agent, etc.) to resolve a complication related to a family vacation/holiday.</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>A3: I can tell a classmate/co-worker an experience in detail so they understand my point of view on a current situation.</td>
<td></td>
<td>X?</td>
<td></td>
</tr>
<tr>
<td>A4: I can negotiate a solution that is mutually agreed with my teacher/boss when I am unable to keep a deadline.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A5: I can write a memorable childhood event in vivid detail across multiple paragraphs in an email/letter to a friend.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A6: I can write a multiple paragraph review of a place of business that caters to families that includes both positive and negative aspects in a respectful tone.</td>
<td></td>
<td>X?</td>
<td></td>
</tr>
<tr>
<td>A7: I can write a detailed report for a teacher/boss that summarizes what I read as part of a school/work assignment.</td>
<td></td>
<td>X</td>
<td>X?</td>
</tr>
<tr>
<td>A8: I can write a formal Petition for Exception to school/vendor that describes why the current policy/procedure is problematic and propose different alternatives that could provide a mutually beneficial resolution.</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>S1: I can critically evaluate and expound upon the impact of government-sponsored daycare programs on family life and its economic implications in a public forum such as a Parent-Teacher Association or school board meeting.</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>S2: I can discuss implications of employee dress and grooming standards on workplace morale, productivity and brand representation in a panel discussion at a professional conference.</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>S3: I can extensively review different arguments of bringing children into the workplace and write a position paper on the relative strengths and weaknesses of various interventions.</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
S4: I can write a critique of current practices or policies in the workplace and highlight the resulting unintended consequences, give supported argumentation on a revised policy and hypothesize how conditions could be ameliorated through adopting the revisions.

Note: Shaded boxes are Can-Do statements referencing writing ability whereas unshaded boxes reference spoken ability.

Example change to S3. Think of a current issue that is being reported on in the news or in your community: 1. I can read page long arguments for and against this issue and understand the details and logic that support them. 2. I can write a paper/OpEd/response that is a page or more in length outlining the strengths and weaknesses of these arguments and present reasons for my own preference on this issue.
APPENDIX D

Additional Comparisons and Discussion

Gender Across Groups

As noted in the Methods section, there were more male than female participants in our test population. However, one would hope the populations would be close to evenly distributed across the three groups. As shown in Figure D1, the male population was had the largest split between the preliminary group \(n = 99\) and the experimental group \(n = 81\) with the control group \(n = 83\) only slightly above the experimental group. The female dispersion was more problematic, however, with the control group having nearly twice as many \(n = 60\) than the preliminary group \(n = 35\), and the experimental group coming more in the middle with \(n = 47\). A chi-square test\(^5\) showed that there was a significant difference between gender populations across these three groups \(x^2 = 7.84, p = 0.020\).

Figure D1

*OPIc Levels of all Participants Divided by Gender*

---

\(^5\) It is recognized that a chi-squared test is usually for random data only. As such, including the preliminary group does not technically meet the requirements for the chi-square. It is offered just for consideration.
Considering that in this particular population the female participants are likely to have less immersion hours, this unequal assignment into groups may play a role in the unequal proficiency levels across groups. Specifically, the female population had more participants who reported no immersion experience \((n = 27\) compared to male \(n = 13\)) and no participants with estimated immersion hours above 3000. The male population, on the other hand, had \(n = 20\) participants with estimated immersion hours above 3000. Whether due to less immersion hours or other factors, as Figure D1 shows, female actual proficiency (determined by the OPIc) was lower with only 18.3% \((n = 26)\) rating above Advance Low compared to 38.5% \((N=96)\) for the male population. These figures were almost flipped for the lowest levels (IM and IH) where 38% of women were rated \((N=54)\) compared to only 18.2% of men \((n = 48)\).

**Survey Language Between Groups**

Languages with smaller sample sizes (Japanese and Russian especially) were less likely to be randomly assigned across groups. If the languages had an imbalance as to how early on students took the LASER is also problematic considering that the first 12 days were grouped into the preliminary group. As Table D1 shows, these factors are especially problematic for French, Italian, and Russian who had 63%, 70% and 5% respectively of their population end up in the preliminary group. Spanish and Portuguese, with larger populations, were more forgiving with only a quarter of their population in the preliminary group still leaving a large number of participants to be split fairly evenly among the experimental group and the control group. A chi-square test on all three groups indicating that there is a significant difference between these groups in regard survey language \([x^2 = 65.8, p = <.001]\). However, a chi-square for only the experimental group and the control group, being much more similar groups, did not show significant difference \([x^2 = 3.13, p = 0.680]\). It should be noted that Japanese \((n = 9)\) and Italian
(n = 6) only have two participants each in the experimental group making the distribution there problematic for finding meaningful results related to those two languages.

Table D1

*Frequencies of Survey Language Across Groups*

<table>
<thead>
<tr>
<th>Survey Language</th>
<th>Groups_All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
</tr>
<tr>
<td>Fre</td>
<td>49</td>
</tr>
<tr>
<td>Ital</td>
<td>14</td>
</tr>
<tr>
<td>Jap</td>
<td>2</td>
</tr>
<tr>
<td>Port</td>
<td>33</td>
</tr>
<tr>
<td>Rus</td>
<td>1</td>
</tr>
<tr>
<td>Span</td>
<td>35</td>
</tr>
</tbody>
</table>

**Immersion Hours between Groups**

As shown in Table D2, immersion hours were roughly similar across all three groups. A chi-square test showed that each group is representative of the entire population \( \chi^2 = 16.9, p = 0.532 \). Notice that the participants (n = 40) who reported having no immersion experience are roughly even across groups as well with \( n = 15 \) (Pre), \( n = 13 \) (Exp), and \( n = 12 \) (Control), while the total participants (n = 365) who reported having an immersion (regardless of how long) were also comparable with \( n = 119 \) (Pre), \( n = 115 \) (Exp), and \( n = 131 \) (Control).

Table D2

*Immersion Hours across Groups.*
Contingency Tables

<table>
<thead>
<tr>
<th>Immersion hours</th>
<th>Groups_All</th>
<th>0</th>
<th>500</th>
<th>1500</th>
<th>2000</th>
<th>2500</th>
<th>1000</th>
<th>3000</th>
<th>3500</th>
<th>4000</th>
<th>4500</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td></td>
<td>15</td>
<td>4</td>
<td>24</td>
<td>25</td>
<td>31</td>
<td>13</td>
<td>13</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>134</td>
</tr>
<tr>
<td>Exp</td>
<td></td>
<td>13</td>
<td>4</td>
<td>15</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>18</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>128</td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td>12</td>
<td>9</td>
<td>28</td>
<td>21</td>
<td>23</td>
<td>29</td>
<td>16</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>143</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>40</td>
<td>17</td>
<td>67</td>
<td>69</td>
<td>78</td>
<td>47</td>
<td>11</td>
<td>6</td>
<td>3</td>
<td></td>
<td>405</td>
</tr>
</tbody>
</table>

Age at the Start of Language Acquisition Between Groups

Table D3 displays the distribution of participants by the age that they started learning their survey language. A chi-squared test indicated that the groups were satisfactorily from the same population \(x^2 = 11.8, p = .463\).

Table D3

Age at the Start of Language Acquisition Displayed by Age and Group

<table>
<thead>
<tr>
<th>Age at start of Language Acquisition</th>
<th>Groups_All</th>
<th>Pre</th>
<th>Exp</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 10</td>
<td></td>
<td>12</td>
<td>12</td>
<td>9</td>
<td>33</td>
</tr>
<tr>
<td>11-13</td>
<td></td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>30</td>
</tr>
<tr>
<td>14-16</td>
<td></td>
<td>34</td>
<td>21</td>
<td>28</td>
<td>83</td>
</tr>
<tr>
<td>17-18</td>
<td></td>
<td>46</td>
<td>45</td>
<td>43</td>
<td>134</td>
</tr>
<tr>
<td>19-20</td>
<td></td>
<td>33</td>
<td>36</td>
<td>49</td>
<td>118</td>
</tr>
<tr>
<td>21-22</td>
<td></td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>23-25</td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>134</td>
<td>128</td>
<td>143</td>
<td>405</td>
</tr>
</tbody>
</table>

The Effect of Survey Language and Training on Calibration

As was noted earlier, some of the languages had either a small \(n\) size within a group or a small \(n\) size in general. With this as a possible limitation, survey language was found to be
significant with regard to calibration \( F = 6.125, p < .001 \). However, an ANOVA comparing survey language and group was highly insignificant \( F = .660, p = .654 \) indicating that survey language and group together were not significant. The closest language to be significant between groups was Portuguese \( F = 2.81, p = .097 \).

As Figure D2 displays, both the experimental group and the control group follow the same basic pattern with member of the control group rating themselves more accurately in all languages except Italian and Russian (which was pretty even distance from zero). It should be noted that Italian in general had very low numbers in the experimental group \( n = 2 \) and the control group \( n = 4 \), while Japanese had especially low number in the experimental group \( n = 2 \). As such the data from those languages are probably not very generalizable. As the experimental group calibrates slightly higher than the control group in every language, it would appear that there is no different effect from training dependent on language.

**Figure D2**

*Plot of Calibration by Language Between the Experimental Group and the Control Group*
The Effect of Immersion and Training on Calibration

Although immersion together with training showed no significant effect on calibration \([F = 1.07, p = .386]\) Figure D3 displays an interesting trend difference between the two groups. Those with training (Exp) who have reported no immersion experience rate themselves higher with training than those in the control group without immersion experience. However, with immersion experience, those with training (Exp) start to move closer to exact calibration whereas those without immersion become less calibrated then originally, on average overrating their proficiency slightly more. Note that those with estimate hours of 3500 and above have less than 11 participants total, likely contributing to larger differences between groups.

**Figure D3**

*Plot of Calibration by Reported Immersion Experience Between Experimental and Control Groups*
This is interesting considering the different experiences of the experimental and control groups for those who did not have any immersion experience. As noted, participants who had no immersive experience rated themselves differently depending on whether they had training or not. Those with the training we likely to be less calibrated, assessing themselves too high, than those in the control group. With minimal proficiency, participants in the control group were likely to even under-assess their proficiency. What is striking is that those who had never been a part of and immersive experience seemed to be more affected by the training, and more likely to consider their abilities better than they were. The training arguably allowed participants to justify their perception more than those who had been in an immersive setting. This may be that the Can-Do statements were less applicable to them because the participants could not relate to the FACTS as well having never been in the situations described.

It is also interesting to point out that 17 of the participants with zero immersion hours (no study abroad or service mission work) were still in the Advanced proficiency level with 2 reaching advanced high, the highest any participant was rated. Both of these participants began language learning before the age of 10, possible in dual language immersion programs which are widely available in the area. (Dual language immersion programs were not counted toward immersion hours in this study.)

The Effect of Initial Language Acquisition Age and Training on Calibration

Age at the start of language acquisition was also found to not be a significant contributing factor to how a participant responded to training \((p = 0.366)\). As Figure D4 displays, the largest difference was between the experimental group and the control group for participants that started learning their survey language between 14-16 years of age with the control group being much
more calibrated than the experimental group. The two oldest age groups had few enough participants \((n = 6, n = 1)\) to not be generalizable trends.

Figure D4

*Plots of Calibration by Reported Initial Age at the Start of Language Training between Groups*

![Graph showing calibration by reported initial age at the start of language training between groups.]

The Effects of Years Studying Language in School and Training on Calibration

The number of years a participant spent studying their survey language in school had no significant effect on how the training affect their calibration \([F = 1.053, p = .370]\). As displayed in Figure D5 there is an interesting jump for participants with only one year of formal language training are more calibrated for the experimental group and less calibrated for the control group. Otherwise, as has been reported elsewhere, there the control group is more calibrated than the experimental group.
Figure D5

Plots of Calibration by Reported Years of Formal Language Training Between Groups