Studies in Second Language Acquisition (2023), **45**, 416–441
doi:10.1017/S0272263122000079

RESEARCH ARTICLE

A closer look at a marginalized test method: Self-assessment as a measure of speaking proficiency

Paula Winke[1]*, Xiaowan Zhang[2] and Steven J. Pierce[1]

[1]Michigan State University, East Lansing, MI, USA; [2]MetaMetrics, Durham, NC, USA
*Corresponding author: E-mail: winke@msu.edu

(Received 06 September 2021; Revised 02 February 2022; Accepted 07 February 2022)

Abstract

Second language (L2) teachers may shy away from self-assessments because of warnings that students are not accurate self-assessors. This information stems from meta-analyses in which self-assessment scores on average did not correlate highly with proficiency test results. However, researchers mostly used Pearson correlations, when polyserial could be used. Furthermore, self-assessments today can be computer adaptive. With them, nonlinear statistics are needed to investigate their relationship with other measurements. We wondered, if we explored the relationship between self-assessment and proficiency test scores using more robust measurements (polyserial correlation, continuation-ratio modeling), would we find different results? We had 807 L2-Spanish learners take a computer-adaptive, L2-speaking self-assessment and the ACTFL Oral Proficiency Interview – computer (OPIc). The scores correlated at .61 (polyserial). Using continuation-ratio modeling, we found each unit of increase on the OPIc scale was associated with a 131% increase in the odds of passing the self-assessment thresholds. In other words, a student was more likely to move on to higher self-assessment subsections if they had a higher OPIc rating. We found computer-adaptive self-assessments appropriate for low-stakes L2-proficiency measurements, especially because they are cost-effective, make intuitive sense to learners, and promote learner agency.

Introduction

We are interested in self-assessment of second language (L2) oral proficiency at the college level because we suspect that self-assessment can be valid as an external measure of oral proficiency within second language acquisition (SLA) research studies and as a measure of oral proficiency achievement or growth for students at different curricular levels within college-level language programs. As we have found as educators, self-assessments are inexpensive when compared to standardized tests, and have test-taking processes that appear to make great intuitive sense to students, especially when compared to other assessments that have been suggested as valid for measuring proficiency within SLA research, such as cloze tests (Tremblay, 2011), C-tests (Eckes &

Grotjahn, 2006; Norris, 2018), or elicited imitation tests (e.g., Deygers, 2020; Erlam, 2006; Gaillard & Tremblay, 2016; Yan et al., 2016). With this article, we seek to validate a particular L2 speaking self-assessment for measuring college-language learners' proficiency, with the understanding that test validation means to "seek evidence for the construct meaning of the test score" (Chapelle, 2021, p. 12). In particular, we aim to "present theoretical rationales and evidence for interpretations and uses of a test" (ibid.), with the test under scrutiny being a computer-adaptive, self-assessment of oral skills, with items based on communicative-oriented proficiency standards that span 10 levels of proficiency. We did this work because even though theoretical rationales for self-assessment as a formative part of the L2 learning process are evident (Andrade, 2019; Brantmeier, 2006; Kissling & O'Donnell, 2015), less apparent within applied linguistics is evidence for the summative interpretation and uses of L2 self-assessment scores.

Early promise with self-assessments, and then doubts

Self-assessments are valuable educational tools (Dolosic et al., 2016) that have been used in diverse educational settings for more than 40 years (see a review by Panadero et al., 2017) because they have processes and results that are meaningful for students (Joo, 2016). As described by Babaii et al. (2016, p. 414), "self-assessment, as a formative assessment tool, promotes learning, establishes a goal-oriented activity, alleviates the assessment burden on teachers, and finally continues as a long-lasting experience (Kirby & Downs, 2007; Mok, Lung, Cheng, Cheung, & Ng, 2006; Ross, 2006)." In the late 1970s, the groundwork for the use of self-assessments in adult foreign, second, or additional language learning (henceforth called L2) was laid by Oskarsson (1978). He noted that standardized, self-assessment forms would help learners set individual goals, define threshold levels of proficiency, and be available for both summative and formative assessment. In the 1980s and 1990s, researchers in applied linguistics began investigating the design and validity of self-assessment instruments to see if they could reliably measure proficiency (Bachman & Palmer, 1989; LeBlanc & Painchaud, 1985; Peirce et al., 1993) and be used for lower-stakes assessment purposes, such as placement into a program's sequence of coursework (LeBlanc & Painchaud, 1985; see Figure 1). Those early studies found that self-assessments placed students at least as well as standardized tests did (ibid., p. 684), and that self-assessment items that asked about specific tasks related to students' personal learning situations worked best (ibid.). Abstract, decontextualized self-assessment items without reference to specific language-use situations, LeBlanc and Painchaud warned, were less reliable.

Since the publication of those early, promising studies (Bachman & Palmer, 1989; LeBlanc & Painchaud, 1985; Peirce et al., 1993) that suggested that self-assessment can be used as a low-stakes measure of L2 proficiency, self-assessments in L2 educational programming, especially at the college level, have not become as widespread as perhaps they should have become. This may be because L2 researchers have promoted a sense of skepticism concerning the usefulness of self-assessment scores, even when considered for low-stakes purposes. In particular, two meta-analytic studies (Li & Zhang, 2021; Ross, 1998) investigated how well self-assessment scores would be able to stand in for objective measurements of L2 proficiency. Both demonstrated a large variability in students' ability to accurately gauge their L2 proficiency through self-assessment. We describe these two meta-analytic studies in more detail next because of their great influence on the field of L2 self-assessment.

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

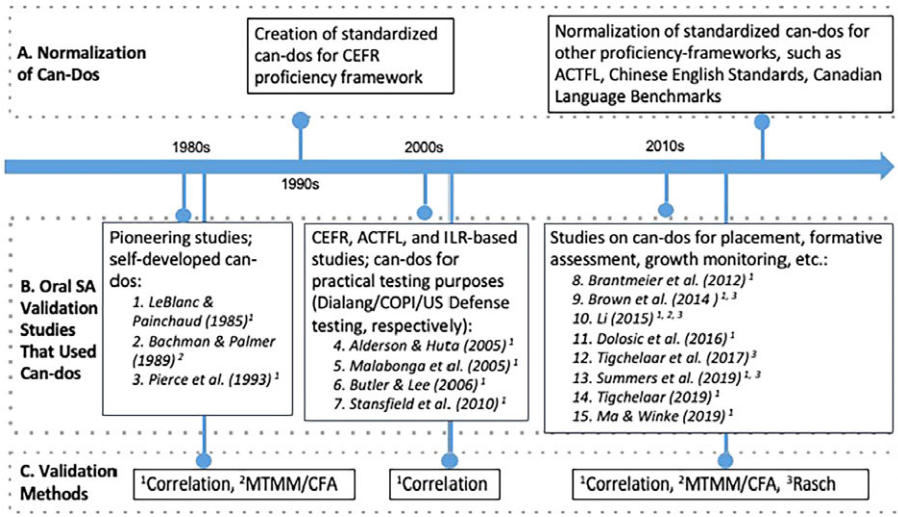

Figure 1. Timeline of 15 studies on L2 self-assessment of speaking 1985–2019

First, in 1998, Ross published a meta-analytic review of 10 studies published in applied linguistics over a 25-year period that contained 60 correlations between L2 self-assessment outcomes and other L2 proficiency scores. He found an average correlation of .63 between self-assessment and other measures, with speaking measures having an average correlation of 0.55. Ross reported that there is "considerable variation in the ability learners show in accurately estimating their own second language skills" (p. 5). Second, Li and Zhang (2021) meta-analyzed 67 studies published over a 42-year period with 214 correlations between L2 self-assessments and external L2 proficiency measures. They found self-assessments correlated with proficiency measures at about .45, with speaking measures correlating on average at .44. Li and Zhang noted overall that the correlations were relatively weak, but that the average correlation (across skills) was statistically significantly improved when the self-assessment described tasks that were specific to one's real life (.49), or when the self-assessment included a rubric or was criterion-referenced (.49). Training the students to perform the self-assessment also improved the relationship significantly (.48). When the self-assessments were computer adaptive, the correlation was also significantly stronger (.52), yet still not sufficiently high enough to swap out standardized test scores with self-assessment scores. Li and Zhang concluded that future researchers should attend to variables that can "improve the correlation between SA [self-assessment] and language performance" (p. 210), and they wrote that they hoped that their results would spur more interest in self-assessments within the field of applied linguistics.

We believe the focus on self-assessments' meta-analyzed correlation with external proficiency measures may be misplaced for at least two reasons. First, self-assessments in applied linguistics have improved drastically over the last few decades. They force a more critical process of self-evaluation through a larger depth and breadth of questioning, with learners asked to envision their proficiency in terms of standardized, positively worded L2-learning descriptors (e.g., Brantmeier et al., 2012;

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

Summers et al., 2019), rather than in comparison to native-speaker norms (e.g., Bachman & Palmer, 1989; Peirce et al., 1993). However, Ross (1998) and Li and Zhang (2021) did not consider these time-bound changes in L2 self-assessment design. The research outcomes meta-analyzed by Ross were published between 1978 and 1992, and those by Li and Zhang were between 1978 and 2019. Meta-analyzing self-assessments' correlations with external measures over multiple decades, we believe, may be like averaging cars' fuel efficiency over multiple decades to inform the public on how fuel-efficient cars are, all the while ignoring recent automotive-engineering development: The outcome would be rather meaningless in regard to contemporary design and use. To our knowledge, no systematic evaluation of the impact of older data on the outcomes of meta-analyses within applied linguistics exists, although there have been calls to measure effect-size changes over time within meta-analyses within the field of SLA to understand if, for example, instrument or statistical modernizations are responsible for effect-size changes over time, which would cloud meta-analytic, effect-size-average results. One could do that by adding time as a moderator variable in the field's meta-analytic studies (see Plonsky & Oswald, 2014). In other fields, such as medicine, however, it has been shown empirically that older data can disproportionately impact meta-analytic outcomes, and meta-analytic studies in medicine involving more than a 10-year window are seldom conducted (less than 33%), and most or all have less than a 20-year window (Patsopoulos & Ioannidis, 2009). If self-assessments have only recently become more robust and critical in their processes, as researchers have pointed out (Andrade, 2019; Summers et al., 2019), shouldn't self-assessments be viewed and evaluated more contemporarily? The answer may be yes, especially because earlier L2 self-assessments were less statistically reliable, had item types that produced less valuable score-meaning interpretations, and focused less on promoting self-regulated learning, one of the goals of self-assessment today (Andrade, 2019).

The second reason we believe a focus on self-assessments' meta-analyzed correlation with external proficiency measures may need revisiting is because correlation is only one type of evidence entailed in test validation. While correlation is informative, it does not present a complete picture of the usefulness of test scores. As described by Chapelle (1999, p. 265; see also 2021), "for applied linguists who think that 'the validity of a language test' is its correlation with another language test, now is a good time to reconsider." Test validation involves collecting evidence to support test-score *uses*, including the consequences involved in those uses (Messick, 1994). Test validation includes demonstrating, with evidence, that the test method and the item characteristics are good, both psychometrically and psychologically (Chapelle, 1998). Most recently, self-assessments of L2 speaking proficiency have been made to be computer adaptive (Summers et al., 2019; Tigchelaar, 2019; Tigchelaar et al., 2017) to improve self-assessment's psychometrics (as evidenced by Li & Zhang, 2021) *and* psychological aspects: Computer-adaptive tests ensure that students will not receive items that are too far away from the students' ability level, which means that all test takers are "challenged but not discouraged" (Wainer, 2000, p. 11; see also Malabonga et al., 2005). Importantly, computer-adaptive self-assessments are not linearly administered. They involve sequential selection processes. Correlations based on sum or final scores cannot reveal if the internal algorithms within a computer-adaptive test are functioning well, which would be an important piece of validity evidence for a computer-adaptive self-assessment. Thus, on several fronts, more than correlation is needed to understand if a self-assessment is working well as a summative measure, especially if that self-assessment is computer adaptive.

Different forms, different purposes: Self-assessments then and now

In this study, we are interested in investigating one single type of self-assessment: self-assessment of L2 speaking. Thus, we investigated what other validation types researchers have used to demonstrate the usefulness of oral self-assessment scores. We itemized 15 studies on L2-speaking self-assessment whose authors investigated how valid those assessments' scores were for summative purposes. We reviewed these papers to better understand *what* the authors validated in terms of the L2 speaking self-assessments, and *how* they did this work. We plotted a timeline of the 15 studies in Figure 1: A complete table of the 15 studies appears in a supplemental file available alongside the online version of this paper (see Table A in the supplement). We divided the studies into three basic phases: the pioneering studies (n = 3) (Bachman & Palmer, 1989; LeBlanc & Painchaud, 1985; Pierce et al., 1993), whose authors designed their own can-do statements; the studies (n = 4) from the first 10 years of the 2000s that first used can-do statements from national or international testing scales; and the studies (n = 8) published since 2010 that have normalized using can-do statements from the standardized scales.

We would like to point out that after 2010, an important transition occurred in speaking self-assessments: Some of them became computer-adaptive (Ma & Winke, 2019; Summers et al., 2019; Tigchelaar, 2019; Tigchelaar et al., 2017) with two or more prearranged testlets (i.e., item sets) of increasing difficulty. As test takers move through the self-assessment, they must pass implicit, predetermined *thresholds* between testlets at adjacent difficulty levels. *Thresholds* are a computer-adaptive test's *adaptive* points, that is, points within the test that calculate (run an algorithm to see) whether a test taker will either stop being tested, or continue on to a new set of items, based on the test taker's performance on the item set that appeared between the current and last threshold (e.g., in Tigchelaar, 2019, participants had to indicate they could do well on at least 8 out of 10 self-assessment can-do statements on a given testlet to be able to take the next testlet; otherwise, the self-assessment was terminated for the participant). Like non-computer-adaptive tests (in which all test takers take all items), computer-adaptive tests need "defensible measurement models, validity, and reliability, and fairness" (Bunderson, 2000, p. xi). But they have some clear differences over their linearly administered counterparts: They have thresholds that are set based on hypotheses. The test designers choose the number of thresholds and set the algorithms that run the thresholds for many reasons, including to ensure a test taker will not be faced with "too many inappropriately chosen items" and to assure "that the examinee understands the task" (Wainer, 2000, p. 9). The appropriateness of the thresholds can be investigated, yet none of the authors of the four studies in Table A with computer-adaptive tests did that. (To sum forthwith, Summers et al. [2019] investigated the psychometrics of a computer-adaptive speaking test with one threshold, but did not investigate the validity of the threshold, and Tigchelaar et al. [2017], Tigchelaar [2019], and Ma and Winke [2019] investigated the psychometrics of different versions of this paper's computer-adaptive speaking test with four thresholds, but did not investigate the validity of the thresholds.)

Approaches to providing evidence for the validity of oral self-assessments

As outlined in Table A in the online supplement and overviewed in Figure 1, the applied linguistics researchers of the 15 L2 speaking self-assessment studies collected four different types of statistical evidence to support the utilization (see Chapelle, 2021,

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

Table 2.1, p. 15) of their tests' self-assessment scores for particular real-world uses, like placement (e.g., LeBlanc & Painchaud, 1985; Li, 2015; Summers, 2019), to measure gains (e.g., Brantmeier et al., 2012; Brown et al., 2014), or as an alternative measure of proficiency (e.g., Butler & Lee, 2006; Pierce et al., 1993). They used these statistical approaches to test the claims regarding what self-assessment scores mean, or for what the scores can be used. We summarize the 15 studies' statistical methods here in relation to the inferences they support.

1. **Correlation.** When researchers claim that self-assessment scores are useful for specific purposes like placement or estimating L2 speaking proficiency levels, the researchers are suggesting that the self-assessment can replace or stand-in for an assessment that was previously used for the same purpose. The main method to test for this claim, as seen the studies in Table A in the supplement, has been through correlation of the self-assessment scores with the same test takers' scores on a trusted measurement used for the same purpose. Correlation was used in 13 of the 15 studies on oral self-assessment in Table A. Ma and Winke (2019) used approaches similar to correlation, that is, by calculating exact and adjacent agreement between test takers' self-assessment and standardized proficiency test scores. In most cases, Pearson correlation was used.
2. **Rasch modeling.** When researchers claim that L2 speaking self-assessment scores reflect the construct of L2 speaking well, the researchers are suggesting that the test takers' L2 speaking ability explains or gives rise to their L2 self-assessment scores. One way to provide evidence for this construct-explanation of scores is to test the underlying assumption that L2 speaking is a unidimensional construct with a single, underlying ability scale. Testing that unidimensional theory with Rasch modeling using the L2 assessment scores provides evidence in support of the notion that the construct (L2 speaking) is being measured, and that L2 speaking as a trait thus explains the self-assessment scores. Rasch modeling also helps determine if the test scores and items that construct that score accurately summarize the relevant performance, that is, if the items discriminate, are of appropriate difficulty, and fit the model well. Rasch modeling was used in 4 of the 14 studies for these purposes.
3. **Multitrait multimethod (MTMM).** MTMM is another way to show evidence for a test's construct validity. Like Rasch modeling, it can provide evidence that L2 speaking ability gives rise to the self-assessment score. The MTMM design is "a classic approach to designing correlational studies for construct validation" (Bachman, 1990, p. 263). In this approach, a test is considered to be a combination of trait (construct) and test method (e.g., self-rating or oral proficiency interview): MTMM reveals how much of the score is trait-based and how much is method-based. Using MTMM, a test is considered construct valid only if it has high correlations with external tests that measure the same trait using different methods (it has convergent validity) and has low correlations with external tests that measure different traits using the same method (it has discriminant validity). MTMM was used in 2 of the 14 studies.
4. **Factor analysis.** Factor analysis, exploratory or confirmatory, is another tool that can be used for assessing an instrument's construct validity, that is, whether the meaning of the test's scores is based on the defined construct (see Chapelle, 2021, for more on the surmising of score meaning). Bachman and Palmer (1989) used confirmatory factor analysis (CFA) to confirm the hypothesized structure of a

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

self-assessment's subscales, including speaking, providing evidence that the self-assessment scores reflect the assumed construct.

To summarize, while correlation coefficients are easily obtainable and offer a straightforward interpretation, they take insufficient account of important scale-based assumptions underlying newer, more specially designed self-assessments of a computer-adaptive nature: Correlation coefficients do not provide information about the goodness of the thresholds predetermined by designers for a computer-adaptive self-assessment. More than correlation needs to be used to provide evidence for appropriate uses of scores from a computer-adaptive self-assessment, but other commonly used analyses (Rasch, MTMM, and factor analysis), are not right for the job. In particular, robust, computer-adaptive self-assessments must provide evidence that the hypotheses underlying the thresholds can withstand tests of those hypotheses. None of the studies that employed and validated a computer-adaptive speaking self-assessment have investigated the goodness of the assessments' thresholds. This article serves as an example of how to do this.

The present study

For the present study, we collected evidence for the uses of scores from a computer-adaptive, self-assessment of L2 speaking for measuring L2 proficiency within a college program. Specifically, we addressed a main claim (that self-assessment scores reflect the learners' oral proficiency as measured by a standardized oral proficiency test), for which we have a specific hypothesis or, as test validation researchers would call it, a *warrant,* which has *backings* that can be supported with empirical evidence (Chapelle, 2021; Kane, 2006). If the warrant in support of the claim does not prove true, we have suspicions on why that might be, and this is called the *rebuttal*, which in turn may have backings from prior studies. We diagramed the connections among these validation argumentation points in Figure 2.

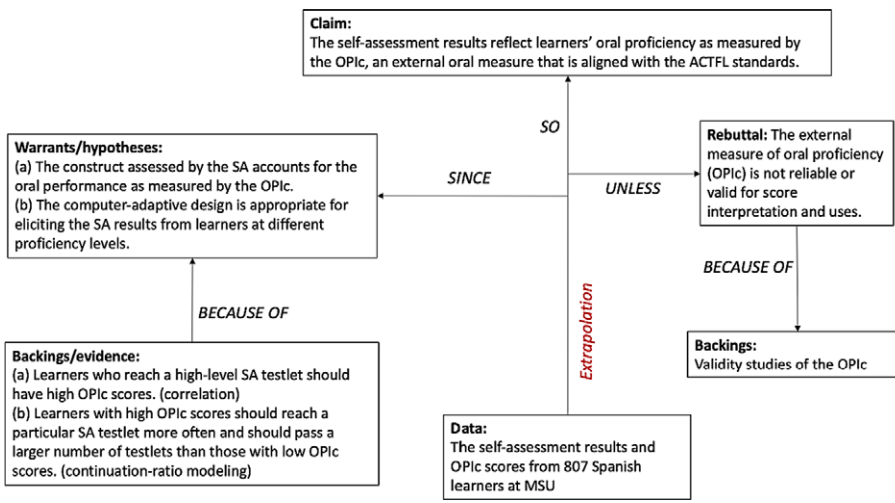

Figure 2. This study's validity claim, warrants, backings, and rebuttal that will be tested through the analysis of the data.

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

As a cross-walk to more traditional empirical research methodology, the study can also be seen as having two main research questions:

1. Does the construct assessed by the self-assessment account for students' oral performance as measured by the OPIc?
2. Is the computer-adaptive design of this oral self-assessment appropriate for eliciting the self-assessment results from learners at different proficiency levels?

Method

Participants

The data in this study are a subset of the data collected for the Language Proficiency Flagship project at Michigan State University. For the project, a sample of intact Chinese, French, Russian, Spanish classes at Michigan State University were pseudo-randomly selected to have their proficiency measured on five occasions over the course of three academic years (fall 2014 through spring 2017). At each time of testing, the sampled classes were brought by their language instructors to a computer lab to take a background survey, a self-assessment of oral skills, and a computerized oral proficiency interview test from Language Testing International (LTI; https://www.languagetesting.com/, a test officially known as ACTFL's OPIc). For the current study, we focus on the Spanish students who were tested in spring 2017, and we use their oral proficiency interview test scores and their self-assessment outcomes. Note that we only use data from Spanish-learning students who received interpretable OPIc test scores. This means that we excluded the data from 64 students who received either "above range" (AR = 4), "below range" (BR = 59), or "unratable" (UR = 1) on the OPIc test. AR was given when a student selected a test form that was too easy; BR was given when a student selected a test form that was too difficult; and UR was given when a student submitted no response or a response that was not ratable due to various reasons (e.g., technology failure).

The students who received interpretable OPIc scores were enrolled in first-year (100-level; $N = 131$), second-year (200-level; $N = 251$), third-year (300-level; $N = 346$), and fourth-year (400-level; $N = 79$) Spanish courses within the four-year program, for a total of 807 students in this study. The sample size was not determined using *a priori* power analysis. It was determined by the number of students who enrolled in the Spanish courses during the study period. The data and a codebook explaining the study's variables are publicly available (see Winke & Zhang, 2022).

Materials

As mentioned already, we used two sets of test data (oral self-assessment, and ACTFL's OPIc scores) for this project, and we additionally recorded the students' year in the 4-year Spanish program as a gross indicator of Spanish ability (see Winke & Zhang, 2022). We describe these in more detail next.

Self-assessment of oral proficiency

This self-assessment of oral proficiency was originally developed by the research team at Michigan State University to assist individual students in identifying their approximate level of oral proficiency on the ACTFL (2012) proficiency scale. The self-assessment is semi-computer adaptive and comprises five testlets of 10 National

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

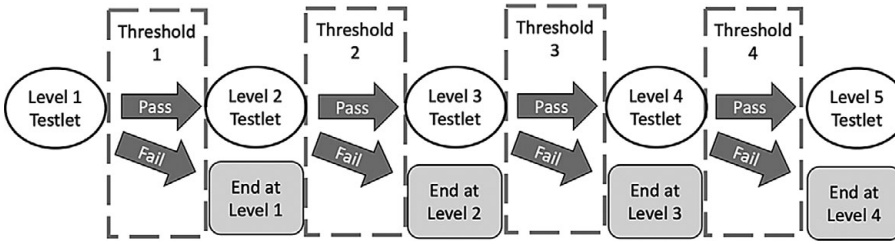

Figure 3. The sequential selection process of the self-assessment. Level 1 through level 5 represent the five testlets of 10 can-do statements in the self-assessment. Threshold 1 through threshold 4 represent the four thresholds that implicitly exist between every two levels of self-assessment testlets.

Council of State Supervisors for Languages-ACTFL (NCSSFL-ACTFL 2015) can-do statements, with each testlet labeled as level 1 through level 5. The 50 statements were selected from NCSSFL-ACTFL's larger 2015 list of can-do statements to roughly represent the five ranges of proficiency as measured by the five different ACTFL OPIc test forms: (1) novice-low to intermediate-mid, (2) novice-high to intermediate-mid, (3) intermediate-mid to advanced-low, (4) intermediate-high to advanced-mid, and (5) advanced-mid to superior. Students responded to individual statements by rating their ability to perform the task described on a 4-point Likert scale: 1 ("Not yet"), 2 ("With much help"), 3 ("With a little help"), and 4 ("Yes, I can do this well"). All students started the self-assessment from the testlet labeled as level 1 and were only allowed to proceed to the next higher testlet level when they indicated mastery ("Yes, I can do this well") on 8 out of 10 statements. Thus, when a student takes this self-assessment, they are partaking in a *sequential selection process*, as illustrated in Figure 3. Conceptually, one can perceive the self-assessment at hand as having four sequentially presented thresholds, with one threshold between every two testlets (or levels) of 10 statements. Passing over a threshold to the next set of 10 can-do statements requires a demonstration of mastery (a score of 4 out of 4) on 8 out of 10 statements. This qualifies the student to be able to try to pass over the next threshold. However, failing to pass over a threshold (by *not* indicating mastery on 8 out of 10 statements) terminates the self-assessment. For example, if a student does not pass over threshold 1, they are given their raw score (out of 200 points possible) and information that they are most likely between novice low and novice high in speaking on the ACTFL (2012) proficiency scale. They are additionally told, if they will take the ACTFL OPIc, to take the Level 1 ACTFL OPIc test form. Through such a sequential selection process, the transition testing sequentially filtered students out of the self-assessment process. In other words, the test with five testlets has an "adaptive stopping rule" (Wainer, 2000, p. 249) that is applied at each threshold.

Consequently, those who self-assessed themselves lower were presented with fewer can-do statements (as few as 10), and those who self-assessed themselves higher were presented with more (as many as 50). More details about the development of the self-assessment can be found in the paper by Tigchelaar et al. (2017). The self-assessment instrument is available at https://tinyurl.com/MSUselfassess.

Standardized, oral proficiency assessment

As mentioned above, the ACTFL OPIc is an internet-delivered oral proficiency test. (For an open-access description and review of the test, see Isbell & Winke, 2019.)

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

Table 1. The five ACTFL OPIc test forms offered to the students in spring 2017 (adapted from Isbell & Winke, p. 469).

ACTFL OPIc form	Target proficiency range	Score range	No. of prompts
1	NL–NH	NL–IL	12
2	IL–IM	NL–IH	15
3	IH–AL	NL–AL	15
4	AL–AM	IH–AH	17
5	AH–S	AM–S	13

At the time of testing, LTI offered five OPIc test forms targeting five ranges of proficiency, as shown in Table 1. Each student was asked to self-select the test form to take based on their outcome on the self-assessment administered immediately before they were to take the OPIc.

Procedure

The data were collected in spring 2017 at Michigan State University as part of the Language Proficiency Flagship Project. Each Spanish language learner came into a computer lab with their class and their instructor toward the end of the spring 2017 semester to have their language proficiency evaluated as part of their regular, programmatic coursework. The students first took the computer-adaptive self-assessment. They immediately received their self-assessment results upon completion of the self-assessment. They were encouraged to use the outcome of the self-assessment to help them choose their level of the LTI ACTFL OPIc (level 1, 2, 3, 4, or 5, as in Table 1). They then proceeded with taking the ACTFL OPIc. All students completed these two tasks within the normally allocated class time. The students received their OPIc results from LTI by email approximately 2 weeks later.

Analyses

To provide a robust understanding of the self-assessment data and their relationship with the students' OPIc ratings, we analyzed the data using multiple methods, each providing a different perspective on the validity of the self-assessment-score uses. To prepare the data for analysis, we numerically converted OPIc ratings to a scale of 1 to 10 (novice-low = 1, superior = 10; see Tigchelaar, 2019 for a discussion of other scaling methods researchers using ACTFL proficiency ratings have employed). For all analyses, we centered the numeric scores at 5 (i.e., intermediate-mid) prior to use. In our analyses, we treated the centered (–4 to 5) OPIc scores as a continuous variable and treated self-assessment levels (levels 1 to 5) as an ordinal variable. Year in the program was also treated as ordinal.

Correlations

To investigate the first warrant (warrant a) in Figure 2, we examined the correlation between self-assessment levels and centered OPIc scores. We estimated a polyserial correlation (r_{ps}) between the continuous OPIc scores and the ordinal self-assessment levels. Polyserial correlation measures the strength of the linear association between a continuous and a discretized ordinal variable (Drasgow, 2006; Hasegawa, 2013). In our

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

current dataset, the ordinal variable, self-assessment level, arose from discretization, which is when a spectrum of continuous scores are divided into a finite number of discrete elements. Although we measured self-assessed oral proficiency on an ordinal scale with five discrete levels (similar to a Likert-scale item) in this study, it does not mean that the underlying construct of self-assessed oral proficiency is distributed on an ordinal scale. Like other proficiency-based variables (e.g., reading proficiency), self-assessed oral proficiency can be reasonably assumed to have a normal distribution on a continuous scale. In other words, we artificially discretized the underlying continuous variable of self-assessed oral proficiency by using our ordinally scored self-assessment measure.

The polyserial correlation coefficient is more appropriate for computing the correlation between a discretized ordinal variable and a continuous variable, as compared with other correlation coefficients that are more familiar to applied linguists, such as Pearson's *r*, Spearman's rho, and Kendall's tau. Pearson's *r* is the coefficient most widely used by researchers in applied linguistics to correlate self-assessment data with other (external) measures of proficiency. It should be noted that Pearson's *r* is appropriate only if both measures are continuous; otherwise, for example, when one of the measures is ordinal, Pearson's *r* likely underestimates the relationship. Some self-assessment researchers who had ordinal self-assessment scores used Spearman's rho (Li, 2015) or Kendall's tau (Peirce et al., 1993) to investigate the predictive validity of the self-assessment. These researchers, however, neglected the fact that their ordinal self-assessment scores, similar to the self-assessment data in this present study, arose from discretization. Spearman's rho and Kendall's tau are most appropriate with *ranked* ordinal data rather than *discretized* ordinal data; when they are used to correlate *discretized* ordinal data, Spearman's rho and Kendall's tau likely underestimate the relationship (Ekström, 2011). Polyserial and polychoric correlation coefficients are specifically designed for discretized ordinal data: Polyserial correlation is appropriate if, for example, self-assessment is a discretized ordinal variable, and proficiency is a continuous variable; polychoric correlation is appropriate if both self-assessment and proficiency are discretized ordinal variables. An easy way, perhaps, for applied linguists to conceptualize these differences is to envision Spearman's rho and Kendall's tau as appropriate for norm-referenced ordinal data (when students' abilities are comparatively ranked), and polyserial and polychoric correlations as appropriate for criterion-referenced ordinal data (when students' abilities are mapped to an external scale or set[s] of criteria). For the purposes of comparison with correlational outcomes from former studies only, we also calculated the Pearson's *r* and Spearman's rho between self-assessment levels and OPIc scores, despite their inappropriateness for handling discretized ordinal data.

Continuation-ratio modeling

A correlation coefficient is a convenient way to rapidly assess the strength of a relationship between two variables. However, distilling the relationship down to a single number then risks oversimplifying the phenomenon.

The self-assessment test in this study was not linearly administered; rather, it employed a "hierarchical branching scheme" (Wainer & Kiely, 1987, p. 190) at each threshold, which can also be called, as we referred to it previously, as a sequential selection process. Thus, any correlation coefficient based on the self-assessment data will inaccurately estimate the relationship because the correlation coefficient

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

will provide the degree of any linear relationship present (see Klugh, 1986, p. 89). To best account for the sequential selection process underlying the self-assessment data (which is nonlinear and "hierarchically structured;" see O'Connell, 2006, p. 60), we used more appropriate continuation-ratio modeling to further examine the relationship of self-assessment levels with OPIc scores. In other words, we employed continuation-ratio modeling to investigate the second warrant (warrant b) in Figure 2. Supportive validity evidence should show that the probability that a student would pass a given threshold test is positively related to the student's oral proficiency as externally measured by the OPIc. Continuation-ratio modeling is a test of whether a computer-adaptive test's thresholds were set appropriately. We explain, as plainly as possible, what readers need to know about continuation-ratio modeling next.

The continuation-ratio model is a regression model specifically designed for ordinal outcome data generated from a sequential selection process (O'Connell, 2006, ch. 5). The model parameters can be translated into additional estimates and graphs that are easy to interpret and meaningful for assessing the validity and characteristics of the self-assessment.

In the current study, we define *continuation ratio* as the proportion of students who were presented with a particular threshold (when taking level 1, 2, 3, or 4 testlet of the semi-computer adaptive self-assessment; see Figure 1) and passed over the threshold to the next testlet (exhibited mastery on eight or more of the testlet's can-do statements). Each of the four thresholds is associated with a continuation ratio, which, as defined already, is identical to the *conditional pass rate* (i.e., conditional probability) of that threshold. We used continuation-ratio modeling to investigate the relationship between the conditional pass rates (dependent variable) and the OPIc scores (predictor). Specifically, we performed logistic regression using data in a person-threshold format, wherein thresholds were nested within persons. Passing a threshold was coded as 1, whereas failing to pass was coded as 0. We expected to observe a positive relationship between the OPIc scores and conditional pass rates; that is, we expected students who obtained high OPIc scores to (a) pass a given threshold more often and (b) to pass a larger number of thresholds in general than those students with low OPIc scores.

Analysis software

We used R 4.1.2 and several R packages (e.g., polycor, car, multcomp) to perform the correlational and continuation-ratio analyses. We evaluated a set of continuation-ratio models (which we will define in the results section) with respect to goodness of fit, calibration, and discrimination because we adopted a predictive modeling approach rather than an explanatory one (Fenlon et al., 2018; Sainani, 2014; Schmueli, 2010). We assessed goodness of fit with (a) an R^2 based on deviance residuals (Cameron & Windmeijer, 1997; Fox, 1997, p. 451), (b) the Akaike information criterion (AIC), and (c) the Bayesian information criterion (BIC) (Sainani, 2014). Our calibration measures included (a) the Hosmer-Lemeshow test statistic, (b) calibration plots, and (c) scaled Brier scores (Fenlon et al., 2018; Steyerberg et al., 2010). Our discrimination measures were (a) classification accuracy, (b) specificity, (c) sensitivity, and (d) area under the curve (AUC). We compared the alternative models using (a) log-likelihood ratio tests, (b) AIC, and (c) BIC, along with the aforementioned goodness-of-fit measures.

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

Reproducibility

We published the data in a public archive (Winke & Zhang, 2022). We also published a research compendium (Marwick et al., 2018) consisting of a custom R package (Pierce & Zhang, 2022) containing our analysis scripts and our raw statistical output.

Results

Correlations

Before we present the correlation, we first describe the data that formed the bases of the correlation. We first present frequency counts of the students' highest self-assessment levels by OPIc score (Table 2). (See our research compendium, Pierce and Zhang [2022], for a distribution of the self-assessment levels of the students who received either AR, BR, or UR on the test.) Overall, students who obtained higher OPIc scores tended to reach higher self-assessment levels, indicating a positive relationship between self-assessment levels and OPIc scores as expected.

The trends in the descriptive data seen in Table 2 were corroborated by the significant and positive correlation of self-assessment levels with OPIc scores, as shown in Table 3. In this article, we present Pearson's *r* and Spearman's rho, but we focus on the polyserial correlation coefficient because, as described in the analysis section, Pearson's *r* and Spearman's rho substantially underestimate a linear association with a discretized ordinal variable (Drasgow, 2006; Ekström, 2011; Hasegawa, 2013), and our self-assessment levels were a discretized ordinal scale. To recapitulate, the self-assessment levels 1 through 5 represent the observed values discretized from an underlying, continuous variable of self-assessed oral proficiency.

Table 2. Frequency counts of students' highest self-assessment level by OPIc score

	Student's highest self-assessment level					
Student's OPIc score	Level 1	Level 2	Level 3	Level 4	Level 5	Total
NL	10	0	0	0	0	10
NM	59	1	1	1	1	63
NH	118	15	1	0	0	134
IL	152	62	10	4	2	230
IM	110	102	23	5	9	249
IH	25	30	11	9	17	92
AL	2	2	5	4	8	21
AM	1	1	0	1	4	7
AH	0	0	0	0	1	1
Total	477	213	51	24	42	807

Table 3. Polyserial, Pearson, and Spearman correlations between self-assessment levels and OPIc scores

			95% Confidence Interval		
Correlational design	Self-assessment levels and OPIc scores	SE	Lower	Upper	p value
Polyserial correlation	0.61	0.03	0.56	0.66	<.001
Pearson's r	0.50	0.03	0.45	0.55	<.001
Spearman's rho	0.50	0.03	0.45	0.55	<.001

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

Continuation-ratio modeling analyses

As described earlier in the analysis section, we used continuation-ratio modeling to examine whether and to what extent OPIc scores predicted the conditional rates at which students passed the four thresholds in the self-assessment. First, we illustrate the derivation of conditional pass rates using descriptive statistics. Table 4 lists the number of students who completed, demonstrated mastery on, and failed to demonstrate mastery on the can-do statements in each testlet level of the self-assessment. By design, the number of students completing a given self-assessment testlet level shrank as the self-assessment level increased. Based on the count data in Table 4, we calculated the conditional pass rate for each threshold by dividing the number of students who passed that threshold by the number of students who took the testlet.

In Table 5, we present the number of students who took and who passed each threshold, as well as each threshold's conditional pass rate. A high conditional pass rate indicates it was relatively easy for students to pass that threshold conditional on that they had passed all previous thresholds. The conditional pass rate, as a type of conditional probability, ranges between 0 and 1.

Here we summarize our steps in model testing. Specifically, we used continuation-ratio modeling to examine whether and to what extent self-assessment thresholds' conditional pass rates, as a dependent variable, were predicted by our independent predictor variable, OPIc scores. We tested two logistic regression models to determine if the predictor had a parallel effect (Model 1) or a nonparallel effect (Model 2) on the conditional pass rates. (Models 1 and 2 in this article are referred to as Models 2a and 2b, respectively, in the Pierce and Zhang, 2022, compendium.) Model 1 included the main effects for OPIc scores and the four thresholds, assuming that OPIc scores had an identical effect across all thresholds. We relaxed the assumption of a parallel OPIc effect in Model 2 by including the main effects for, as well as interaction terms between, OPIc

Table 4. The number of students who completed, demonstrated mastery on, and failed to demonstrate mastery on the statements in each testlet level of the self-assessment

Self-assessment testlet level	N of students who completed the level	N of students who terminated the self-assessment at the level	N of students who demonstrated mastery at the level
1	807	477	330
2	330	213	117
3	117	51	66
4	66	24	42
5	42	42	0

Table 5. The number of students who took and passed each threshold (see Figure 3) and each threshold's conditional pass rate

Self-assessment threshold	N of students who took the threshold	N of students who passed the threshold	Conditional pass rate (continuation ratio)
1	807	330	0.409
2	330	117	0.355
3	117	66	0.564
4	66	42	0.636

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

scores and each of the four thresholds. Model 2 therefore allowed the effect of OPIc scores to vary across thresholds.

Readers can envision the differences in the two models of the predictor in this way: If a predictor has a parallel effect on the conditional pass rate, the predictor will affect each of the four thresholds' conditional pass rate in the same way or by the same amount: the predictor will have a *fixed* effect. If a predictor has a nonparallel effect, the predictor will have threshold-specific effects, meaning that the predictor's effect on the conditional pass rate will *vary*, and will depend on the level of the threshold.

In both models, we omitted the normal intercept term (meaning we dropped the constant term from the models, which is a statistical procedure or tool called "regression through the origin," or RTO), as RTO is appropriate with categorical or ordinal predictors (Casella, 1983; Eisenhauer, 2003). This means that rather than having a shared intercept term associated with a reference value for the transition threshold plus additional parameters testing whether the baseline pass rates for other transition thresholds differed from it, the model instead estimated a separate intercept parameter for each threshold. This makes sense in the language proficiency testing context because those intercepts represent the baseline difficulty of the task at each transition when all other predictors are set to zero. That difficulty should vary because the tasks involved are different at each transition. This parameterization simplified the task of defining contrasts that estimate specific quantities of interest. Furthermore, the results still permit us to examine whether those intercepts (and the pass rates they allow us to compute) are similar or different across transitions.

Model results (see Table 6) were transformed into odds ratios and conditional and unconditional pass rates to facilitate interpretation. In contrast to the *conditional* pass rate, a threshold's *unconditional* pass rate is its absolute pass rate, which can be understood as the proportion of the whole sample that passed the threshold.

Table 7 displays the goodness-of-fit statistics for Models 1 and 2. Both models demonstrated acceptable fit to the data (see the Hosmer-Lemeshow test results in Table 7). We then compared the two models using AIC, BIC, and a likelihood-ratio test (LRT), all shown in Table 7. We were given conflicting answers in terms of model selection. The LRT suggested that Model 2 was preferred over Model 1, revealing a significant effect for the interaction between OPIc scores and thresholds ($p = .017$). Difference in the AIC values also favored Model 2 by a small margin of about four points. However, the difference in the BICs favored Model 1 by a larger margin of about 11 points, suggesting that the increase in model complexity from Model 1 to

Table 6. Model estimates, standard errors, p-values, and odds ratio for Models 1 and 2

	Model 1		Model 2	
	Estimate (SE/*p*)	Odds Ratio [95% CI]	Estimate (SE/*p*)	Odds Ratio [95% CI]
Transition 1	0.13 (0.09/.130)		0.19 (0.09/.035)	
Transition 2	−0.70 (0.12/<.001)		−0.69 (0.13/<.001)	
Transition 3	−0.15 (0.21/.482)		−0.01 (0.21/.952)	
Transition 4	−0.06 (0.29/.824)		0.31 (0.30/.309)	
OPIc	0.84 (0.06/<.001)	2.31 [2.05, 2.61]		
OPIc @ Transition 1			0.96 (0.08/<.001)	2.61 [2.10, 3.25]
OPIc @ Transition 2			0.79 (0.13/<.001)	2.21 [1.55, 3.14]
OPIc @ Transition 3			0.54 (0.18/.008)	1.72 [1.07, 2.77]
OPIc @ Transition 4			0.31 (0.21/.252)	1.37 [0.78, 2.38]

Note: OPIc @ Transition j represents the simple effect of OPIc scores on the j transition.

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

Table 7. Goodness-of-fit indices for Models 1 and 2 and likelihood-ratio test result

Fit statistic	Model 1	Model 2
Log-likelihood	−761.34	−756.22
Deviance	1522.68	1512.43
AIC	1532.68	1528.43
BIC	1558.61	1569.91
Hosmer-Lemeshow's chi-square statistic (DF/*p*)	7.05 (8/.531)	2.41 (4/.660)
Brier score	0.20	.20
R-squares based on deviance	.17	.17
Specificity	.731	.731
Sensitivity	.676	.676
Accuracy	.708	.708
AUC of ROC [95% CI]	.752 [.726, .778]	.756 [.729, .781]
Likelihood-ratio test		
Delta deviance	**DF**	***p***
10.25	3	0.017

Model 2, which was due to the addition of the interaction terms, could not be justified by the concomitant improvement in model fit. Because other goodness-of-fit indices (Brier score, AUC, and classification accuracy, sensitivity, and specificity, in Table 7) all suggested little difference in model fit between Model 1 and Model 2, we decided to select Model 1, the parallel-effect model, as the final model because of its parsimony.

Results of Model 1 (see Table 6) indicated that each unit of increase on the OPIc scale (i.e., one sublevel) was associated with a 131% increase in the odds of passing every threshold in the self-assessment. In other words, the odds of passing a given threshold for students who scored advanced-high on the OPIc was expected to be 2.31 times higher than for students who scored advanced-mid, 5.34 (2.31^2) times higher than students who scored advanced-low, 12.33 (2.31^3) times higher than students who scored intermediate-high, and so on, given that the students reached the opportunity to take the threshold.

In Figure 4 and Figure 5, we visualize the model-predicted effects of OPIc scores on conditional and unconditional pass rates, respectively (see Table B in the online supplement for the results' values). A visible trend in both figures is that students were increasingly more likely to pass each threshold, as well as to pass a larger number of thresholds, as their OPIc ratings increased from novice-low to advanced-high. Unique in Figure 5, which illustrates the unconditional (or absolute) pass rates, one can see that among all advanced-level Spanish learners (7 = advanced-low, 8 = advanced-mid, or 9 = advanced-high), over 50% were expected to pass thresholds 1 through 3 in the self-assessment, compared to less than 1% of those who scored in the novice range on the OPIc (1 = novice-low, 2 = novice-mid, or 3 = novice-high).

Discussion

In this article, we analyzed the validity of an L2 speaking self-assessment for determining college-level Spanish language learners' proficiency on the ACTFL oral proficiency scale (2012). We did this work because the field of applied linguistics has been hesitant to fully endorse the use of self-assessments, despite the strong, theoretical rationales for self-assessments as part of a positive learning process (e.g., Brantmeier, 2006; Falchikov & Boud, 1989; Kaderavek et al., 2004; Kissling & O'Donnell, 2015).

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

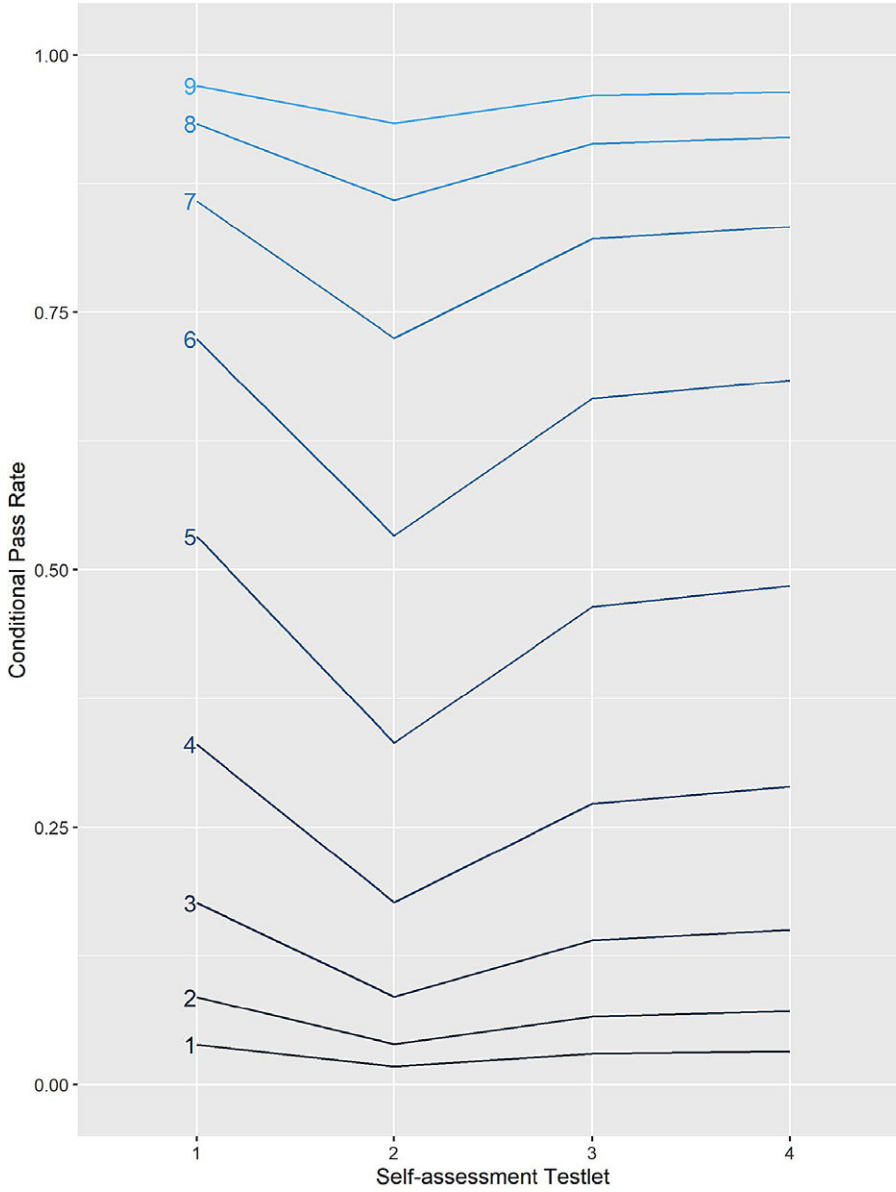

Figure 4. Predicted conditional pass rates by OPIc rating for each self-assessment threshold. The nine lines represent the nine observed OPIc ratings (1 = novice-low, 2 = novice-mid, 3 = novice-high, 4 = intermediate-low, 5 = intermediate-mid, 6 = intermediate-high, 7 = advanced-low, 8 = advanced-mid, 9 = advanced-high). Due to how OPIc ratings were centered for modeling purposes, the line for OPIc = 5 visualizes the interpretation of the set of transition-specific intercepts.

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

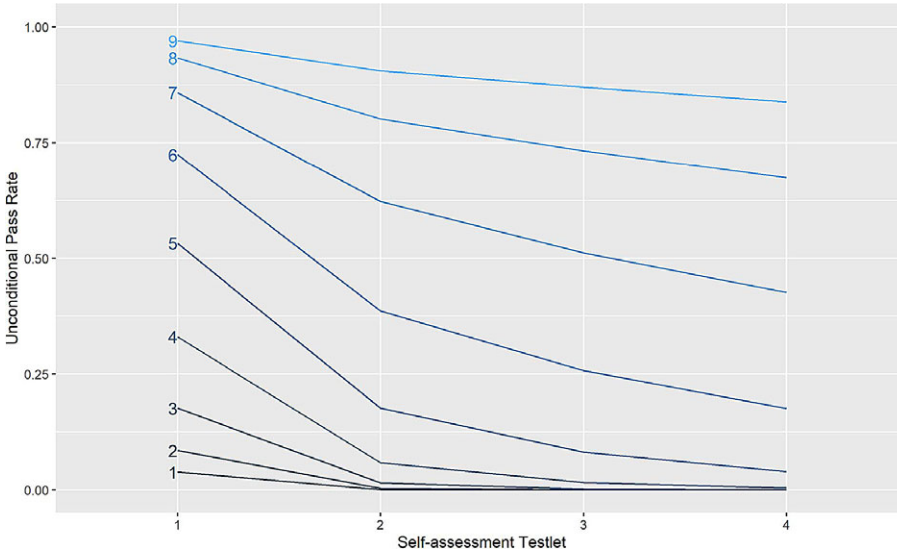

Figure 5. Predicted unconditional (or absolute) pass rates by OPIc rating for each self-assessment threshold. The nine lines represent the nine observed OPIc ratings (1 = novice-low, 2 = novice-mid, 3 = novice-high, 4 = intermediate-low, 5 = intermediate-mid, 6 = intermediate-high, 7 = advanced-low, 8 = advanced-mid, 9 = advanced-high).

Empirical evidence of the high value of inferences from computer-adaptive self-assessment

First, we investigated our claim (that the self-assessment results reflect learners' oral proficiency as measured by the OPIc) through the following warrant and backing, as listed in Figure 2:

- **Warrant a:** The construct assessed by the self-assessment accounts for students' oral performance as measured by the OPIc.
- **Backing a:** Learners who reach a high-level self-assessment testlet should have high OPIc scores.

We found that the self-assessment levels aligned moderately with the OPIc scores (polyserial = .61), which provides evidence that self-assessment scores would work well to differentiate learners according to their proficiency, although perhaps not as well as OPIc scores would. However, we would like to point out, compared to the OPIc or any other performance-based assessment, the self-assessment would do the job of separating individuals within classes or on a larger scale efficiently, cost-effectively, and with little potential for student stress and anxiety, as is common in high-stakes testing (Huang & Hung, 2013; Shi, 2012). Overall, learners who reached high self-assessment levels tended to have high OPIc scores. This relationship was revealed as stronger than when estimated with Pearson's *r* (.50) or Spearman's rho (.50), as expected. The results suggest that previous studies that calculated Pearson's *r* or Spearman's rho for discretized ordinal self-assessment data (e.g., Li, 2015) may have underestimated the predictive validity of the self-assessment.

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

We needed to test the functionality of the thresholds within the testing process through something other than correlation, factor analysis, or MMTM, as those statistical methods fall short in providing information on whether the assessment's thresholds work as well as intended. We therefore applied more robust statistical modeling—continuation-ratio modeling—to investigate the functionality of the self-assessment and to understand the goodness of the test's thresholds. We did this in relation to our claim's second warrant and backing, which were as follows:

- **Warrant b:** The computer-adaptive design is appropriate for eliciting the self-assessment results from learners at different proficiency levels.
- **Backing b:** Learners with high OPIc scores should reach a particular self-assessment testlet more often and should pass a larger number of testlets than those with low OPIc scores.

Our best-fitting model (Model 1) indicated that each unit of increase on the OPIc scale corresponded with a 131% increase in the odds of passing over every threshold. In other words, a student was more likely to pass each self-assessment threshold, as well as to pass a larger number of thresholds, if they correspondingly had a higher OPIc rating. The overall outcomes of the continuation-ratio modeling, which investigated the effect of OPIc scores (Models 1 and 2) on learners' progression though the self-assessment, demonstrated clearly that students' OPIc scores predicted their self-assessments of oral skills rather well.

Arguments for self-assessment score uses by language programs and classroom teachers

We believe that this study provides clear evidence that language teachers and programs should use more self-assessments and use them in conjunction with performance-based testing, not in lieu of it. As we reviewed at the start of this article, decades of research have found positive effects from self-assessment in general education, educational psychology, and other diverse education sub-fields such as medical education (see Panadero et al., 2017). Applied linguists know that learners' vision, agency, autonomy, and conceptualizations of their *possible selves* (i.e., realistic and goal-oriented visions of what they will be able do in the language in the future if their learning continues positively) are part of an enjoyable and engaging language-learning experience (Dörnyei, 2009), and self-assessment can be part of forming, solidifying, or bringing to the fore (for open discussion) those conceptualizations. Self-assessments with positively worded, can-do-based statements can promote agency and motivation (see Brantmeier, 2006; Kissling & O'Donnell, 2015). Meta-analyses from higher education have demonstrated that self-assessment can be reliable (Falchikov & Boud, 1989). Moreover, self-assessment can positively influence students' strategy implementation and self-efficacy (Panadero et al., 2017). Even with younger children (ages 5 to 12), research in general education has found that self-evaluation of learning increases self-efficacy, achievement, and school-based motivation (Kaderavek et al., 2004).

Applied linguistics researchers, however, have tended to take a different stance on self-assessment. Focusing on self-assessments' correlation with other measures of oral proficiency, researchers have concluded that self-assessment provides "too cloudy of a picture of proficiency" (Ross, 1998, p. 12), a notion we countered with this research. With this research, we pondered, too cloudy of a picture of proficiency *for what*

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

purposes? We take up the claim by Chapelle (2021) that each test must be validated according to the *uses* of its scores. She wrote (p. 12), that "beyond construct validation, validation needs to take into account issues of relevance and utility, value implications, and the social consequences of testing." Because self-assessment and OPIc scores are generally not used for the same purposes, and because their social consequences are very different, we argue that one should perhaps not overinterpret their correlational estimates. It could be that a medium correlation should not be unexpected. We found correlational results similar to Ross's findings when we inspected the 15 studies (Table A in the online supplement) that investigated the predictive validity or extrapolation inferencing of L2-speaking self-assessments. We also found a similar result with this present study when we correlated self-assessment levels and OPIc scores (polyserial = .61). We would like to stress here that expectations of anything higher than what we found in terms of the correlations are problematic in that such expectations would ignore the different practical and formative uses of self-assessments over standardized testing. The modern question is not whether self-assessments can be used to *replace* standardized tests (which could be determined through correlational studies), but rather the question is, are self-assessment outcomes reliable and meaningful enough to warrant their perhaps very different uses? Based on our research, we believe the answer is yes, when the uses are for low-stakes assessment, diagnostic purposes, spurring discussions on goals, learning objectives, and building learning motivation, vision, and agency. We also believe computer-adaptive self-assessments, such as this study's test, lend themselves well for repeated use over time because computer-adaptive tests offer different items to students as they gain in ability (Wainer, 2000). Thus, the self-assessment as showcased in this study can be used by teachers and program directors to measure proficiency growth reliably and cost-effectively on the ACTFL scale.

The complementary uses of self-assessments and proficiency-based tests

Unlike the language of meta-analyses that cautioned the field on using self-assessments (Li & Zhang, 2021; Ross, 1998), with this study we have similar results, but interpret them differently. The empirical correlational evidence in this study aligns with those from other studies (Alderson & Huhta, 2005; Butler & Lee, 2006; Stansfield et al., 2010; Summers et al., 2019), but we see the evidence as suggesting that L2 self-assessments of oral skills measure something slightly different from standardized performance tests, and that the two test types have complementary uses. For example, self-assessments may measure the language learner's conceptualization of their *possible self*, that is, their "real-to-themselves" and goal-oriented vision of what they can do in the language (see Dörnyei, 2009, for further definitions of the *possible self*). One can view the self-assessment as an internal, cognitive reflection on what the learner thinks they can do with the language, task-by-task. Meanwhile, standardized performance-based assessments are a direct, external measure of observable performance that may align or not align with the learner's actual vision or estimation of what they can do. And this makes sense. Classroom language teachers and educators need to use both self-assessment and performance-based assessment types over the course of an academic program to foster positive and realistic conceptualizations of the possible self, and to spur practice (which drives SLA; see DeKeyser, 2015) in using the target language, respectively. If one comes to think closely about the two assessment types used in this study (self-assessment and performance-based assessment), one will not expect them to correlate too highly, for

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

even though they both measure aspects of the same larger construct (speaking ability), they tap into it from extremely different, nonoverlapping angles. This need not be seen as problematic, and in fact can be seen as a better way to represent and measure the development of the overall construct: Using both self-assessment and performance-based tests may represent a whole-learner approach to measuring speaking, one that considers the personal, psychological, and cognitive aspects of speaking development, and the more behavioral, procedural, and external manifestation of the learner's underlying knowledge, as judged by someone else.

Arguments for self-assessment score uses by researchers

Calls for robust measures of proficiency for SLA research purposes have been around for a long time (Thomas, 1994). As outlined by Gaillard and Tremblay (2016) and Révész and Brunfaut (2021), SLA researchers must use measures of proficiency and language growth that are reliable and for which the researchers have score-use validity evidence. The test must be discriminating, and often needs to be quick to administer and easy to score given the researchers' typical limited time and resources. For proficiency assessment, the assessment should, Gaillard and Tremblay noted, also be "sufficiently global" so it does not assess the same construct as the L2-learning construct under investigation (p. 420). Three proficiency test formats that have been put forward as meeting such requirements for SLA research have been cloze testing (Tremblay, 2011), C-testing (Eckes & Grotjahn, 2006; Norris, 2018), and elicited imitation testing (Deygers, 2020; Erlam, 2006; Gaillard & Tremblay, 2016; Yan et al., 2016). But what these functionally measure, and the appropriateness of the assessments in measurement, have been long debated due to the abstract and nonauthentic nature of the assessments (Grothjahn et al., 2002; Kim et al., 2016; Spada et al., 2015; Winke et al., in press). We believe our computer-adaptive self-assessment can serve appropriately and functionally as proficiency measurement for SLA researchers, especially because its administration and use may be seen as more intuitive and meaningful than cloze, C-testing, or elicited imitation. The self-assessment also has the advantage of being directly interpretable on a national (ACTFL) language proficiency scale and is less reliant on abstract, scale-score mapping procedures. Furthermore, the self-assessment's items can be adapted for local use. Brantmeier et al. (2012) demonstrated clearly how standardized, scale-based can-do items can be adapted and localized for higher specificity and discriminatory power. Butler and Lee (2006) demonstrated how can-do statements can be tailored for young learners. Tailoring can-do statements to align with a specific language programs' objectives can help with the accuracy in self-assessment when the self-assessment outcomes are used to investigate specific programmatic questions about attainment and growth, while using nonadapted, general can-do statements culled directly from proficiency framework materials can be used to gather proficiency estimations from learners across multiple schools or learning contexts.

Limitations

We would be remiss if we did not review this study's limitations. We based this study in a Spanish language program at one university in the United States, thus the generalizability of this study is limited in that we don't know how well this particular self-assessment would work with learners of other languages or in other learning contexts,

https://doi.org/10.1017/S0272263122000079 Published online by Cambridge University Press

such as in other countries. Additionally, this particular assessment was designed for and used with college-level language learners. Whether computer-adaptive self-assessment functions in the same manner with learners of other age groups and proficiency levels or proficiency-level ranges is also an area that needs further empirical investigation. We also centered this research on the notion that the ACTFL OPIc accurately measures college-level language learners' proficiency with a high degree of accuracy, a notion that has come under question in recent years (see Isbell & Winke, 2019). The higher the errors in measurement (from either random error or systematic error) from the instruments on either side of the correlation would weaken any directional hypothesis because by nature error terms are uncorrelated or correlated, without directional assumptions being possible, as the terms are unknown (see Raykov et al., 2015). Nonetheless, our study seems to point out exactly what Oskarsson found in 1978: that self-assessment estimates are generally good, or in some cases very good, at estimating learners' proficiency.

Few learners in this study had reached the advanced levels on the OPIc test. That necessarily limits the available data for drawing conclusions about advanced learners. We recommend some caution with respect to applying our findings to advanced learners. Replication with new samples containing more advanced learners would be advisable. Because advanced learners are a smaller fraction of the population than those of more modest proficiency, addressing that issue may therefore require starting with larger sample sizes, adopting a narrower focus and recruiting only advanced learners, or pooling results across multiple studies using meta-analysis.

Only 66 of our learners reached the fourth self-assessment transition testlet. Two of them had novice OPIc scores and 18 had advanced scores, leaving most (46) with intermediate scores. Such range restriction tends to reduce regression coefficients (Aguinis et al., 2017), suggesting that our study may underestimate the true OPIc effect at that transition. Meanwhile, the smaller sample size available for examining that final transition also means our estimate of the OPIc effect on that conditional pass rate is less certain than our estimates for the earlier transitions. That uncertainty is already reflected in the larger standard error (and thus wider confidence intervals) associated with those estimates. Solving the sample size problem is straightforward in an ideal world: recruit larger samples for future studies to increase the probability of having sufficient numbers of learners reach the final transition. Practically this is difficult simply because there are fewer language learners who reach the highest levels of proficiency in a given language program. Solving the range restriction issue may be a thornier problem requiring creative research designs or more sophisticated statistical methods.

Data availability statement. The experiment in this article earned an Open Data badge for transparent practices. The materials are available at https://doi.org/10.3886/E164981V1

References

ACTFL. (2012). *ACTFL Proficiency Guidelines 2012.* ACTFL. http://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

Aguinis, H., Edwards, J. R., & Bradley, K. J. (2017). Improving our understanding of moderation and mediation in strategic management research. *Organizational Research Methods*, *20*, 665–685. https://doi.org/10.1177/1094428115627498

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*, *22*, 301–320. https://doi.org/10.1191/0265532205lt310oa

Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education*, *4*, 1–13. https://doi.org/10.3389/feduc.2019.00087

Babaii, E., Taghaddomi, S., & Pashmforoosh, R. (2016). Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing*, *33*, 411–437. https://doi.org/10.1177/0265532215590847

Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing*, *6*, 14–29. https://doi.org/10.1177/026553228900600104

Bachman, Lyle F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

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

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

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

Bunderson, C. V. (2000). Foreword to the first edition. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, & R. J. Mislevy (Eds.,) *Computerized adaptive testing: A primer* (pp. ix-xii). Routledge.

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

Cameron, A. C., & Windmeijer, F. A. G. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, *77*, 329–342. https://doi.org/10.1016/s0304-4076(96)01818-0

Casella, G. (1983). Leverage and regression through the origin. *American Statistician*, *37*, 147–152. https://doi.org/10.1080/00031305.1983.10482728

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In Lyle F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge Applied Linguistics.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272. https://doi.org/10.1017/S0267190599190135

Chapelle, C. A. (2021). Validity in language assessment. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 11–20). Routledge.

DeKeyser, R. M. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (2nd ed., pp. 94–112). Routledge.

Deygers, B. (2020). Elicited imitation: A test for all learners? Examining the EI performance of learners with diverging educational backgrounds. *Studies in Second Language Acquisition*, *42*, 933–957. https://doi.org/10.1017/S027226312000008X

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

Dörnyei, Z. (2009). *The psychology of the second language learner*. Oxford University Press.

Drasgow, F. (2006). Polychoric and polyserial correlations. In S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences* (pp. 1–6). John Wiley & Sons. https://doi.org/10.1002/0471667196.ess2014.pub2

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23*, 290–325. https://doi.org/10.1191/0265532206lt330oa

Eisenhauer, J. G. (2003). Regression through the origin. *Teaching Statistics*, *25*, 76–80. https://doi.org/10.1111/1467-9639.00136

Ekström, J. (2011). A generalized definition of the polychoric correlation coefficient. *UCLA Department of Statistics Papers*, *36*. https://escholarship.org/uc/item/583610fv

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical investigation. *Applied Linguistics*, *27*, 464–491. https://doi.org/10.1093/applin/aml001

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, *59*, 395–430. https://www.jstor.org/stable/1170205

Fenlon, C., O'Grady, L., Doherty, M. L., & Dunnion, J. (2018). A discussion of calibration techniques for evaluating binary and categorical predictive models. *Preventive Veterinary Medicine*, *149*, 107–114. https://doi.org/10.1016/j.prevetmed.2017.11.018

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. SAGE.
Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning*, *66*, 419–447. https://doi.org/10.1111/lang.12157
Grotjahn, R., Klein-Braley, C., & Raatz, U. (2002). C-tests: An overview. In J. A. Coleman, R. Grotjahn, & U. Raatz (Eds.), *University language testing and the C-test* (pp. 93–114). AKS-Verlag.
Hasegawa, H. (2013). On polychoric and polyserial partial correlation coefficients: A Bayesian approach. *METRON*, *71*, 139–156. https://doi.org/10.1007/s40300-013-0012-1
Huang, H.-T. D., & Hung, S.-T. A. (2013). Comparing the effects of test anxiety on independent and integrated speaking test performance. *TESOL Quarterly*, *47*, 244–269. https://doi.org/10.1002/tesq.69
Isbell, D., & Winke, P. (2019). Test review: ACTFL Oral Proficiency Interview—Computer (OPIc). *Language Testing*, *36*, 467–477. https://doi.org/10.1177/0265532219828253
Joo, S. H. (2016). Self- and peer-assessment of speaking. *Working Papers in TESOL and Applied Linguistics*, *16*, 68–83. https://doi.org/10.7916/salt.v16i2.1257
Kaderavek, J. N., Gillam, R. B., Ukrainetz, T. A., Justice, L. M., & Eisenberg, S. N. (2004). School-age children's self-assessment of oral narrative production. *Communication Disorders Quarterly*, *26*, 37–48. https://doi.org/10.1177/15257401040260010401
Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.
Kim, Y., Tracy-Ventura, N., & Jung, Y. (2016). A measure of proficiency or short-term memory? Validation of an elicited imitation test for SLA research. *Modern Language Journal*, *100*, 655–673. https://doi.org/10.1111/modl.12346
Kissling, E. M., & O'Donnell, M. E. (2015). Increasing language awareness and self-efficacy of FL students using self-assessment and the ACTFL proficiency guidelines. *Language Awareness*, *24*, 283–302. https://doi.org/10.1080/09658416.2015.1099659
Kirby, N. F., & Downs, C. T. (2007). Self-assessment and the disadvantaged student: Potential for encouraging self-regulated learning? *Assessment & Evaluation in Higher Education*, *32*, 475–494. https://doi.org/10.1080/02602930600896464
Klugh, H. E. (1986). *Statistics: The essentials for research*. Psychology Press.
LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly*, *19*, 673–687. https://doi.org/10.2307/3586670
Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing*, *38*, 189–218. https://doi.org/10.1177/0265532220932481
Li, Z. (2015). Using an English self-assessment tool to validate an English placement test. *Papers in Language Testing and Assessment*, *4*, 59–96. https://arts.unimelb.edu.au/__data/assets/pdf_file/0003/1770672/Li.pdf
Ma, W., & Winke, P. (2019). Self-assessment: How reliable is it in assessing oral proficiency over time? *F*oreign Language Annals, *52*, 66–86. https://doi.org/10.1111/flan.12379
Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing*, *22*, 59–92. https://doi.org/10.1191/0265532205lt297oa
Marwick, B., Boettiger, C., & Mullen, L. (2018). Packaging data analytical work reproducibly using R (and friends). *The American Statistician*, *72*, 80–88. https://doi.org/10.1080/00031305.2017.1375986
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, 13–*23*.
Mok, M. M. C., Lung, C. L., Cheng, D. P. W., Cheung, R. H. P., & Ng, M. L. (2006). Self-assessment in higher education: Experience in using a metacognitive approach in five case studies. *Assessment & Evaluation in Higher Education*, *31*, 415–433. https://doi.org/10.1080/02602930600679100
NCSSFL-ACTFL. (2015). *NCSSFL-ACTFL Can-do statements*. ACTFL.
Norris, J. M. (Ed.). (2018). *Developing C-tests for estimating proficiency in foreign language research*. Peter Lang.
O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables. Quantitative applications in the social sciences*. SAGE.
Oskarsson, M. (1978). *Approaches to self-assessment in foreign language learning*. Pergamon Press.
Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review*, *22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004

Patsopoulos, N. A., & Ioannidis, J. P. (2009). The use of older studies in meta-analyses of medical interventions: a survey. *Open Medicine*, *3*, 62–68. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2765773/pdf/OpenMed-03-e62.pdf

Peirce, B. N., Swain, M., & Hart, D. (1993). Self-assessment, French immersion, and locus of control. *Applied Linguistics*, *14*, 25–42. https://doi.org/10.1093/applin/14.1.25

Pierce, S. J., & Zhang, X. (2022). *SAWpaper: Self-assessment works paper research compendium* (Version 1.0.0) [Reproducible research materials and computer program, R package]. GitHub and Zenodo. https://github.com/sjpierce/SAWpaper, https://doi.org/10.5281/zenodo.6388011

Plonsky, L., & Oswald, F. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. https://doi.org/10.1111/lang.12079

Raykov, T., Marcoulides, G. A., & Patelis, T. (2015). The importance of the assumption of uncorrelated errors in psychometric theory. *Educational and Psychological Measurement*, *75*, 634–647. https://doi.org/10.1177/0013164414548217

Révész, A., & Brunfaut, T. (2021). Validating assessments for research purposes. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 21–32). Routledge.

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research, and Evaluation*, *11*, 1–13. https://doi.org/10.7275/9wph-vv65

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experimental factors. *Language Testing*, *15*, 1–20. https://doi.org/10.1177/026553229801500101

Sainani, K. L. (2014). Explanatory versus predictive modeling. *American Academy of Physical Medicine and Rehabilitation*, *6*, 841–844. https://doi.org/10.1016/j.pmrj.2014.08.941

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310. https://doi.org/10.1214/10-STS330

Shi, F. (2012). Exploring students' anxiety in computer-based oral English test. *Journal of Language Teaching and Research 3*, 446–451. https://doi.org/10.4304/jltr.3.3.446-451

Spada, N., Shiu, J. L.-J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning*, *65*, 723–751. https://doi.org/10.1111/lang.12129

Stansfield, C. W., Gao, J., & Rivers, W. P. (2010). A concurrent validity study of self-assessments and the federal Interagency Language Roundtable Oral Proficiency Interview. *Russian Language Journal*, *60*, 299–315.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, *21*, 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

Summers, M. M., Cox, T. L., McMurry, B. L., & Dewey, D. P. (2019). Investigating the use of the ACTFL can-do statements in a self-assessment for student placement in an Intensive English Program. *System*, *80*, 269–287. https://doi.org/10.1016/j.system.2018.12.012

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*, 307–336. https://doi.org/10.1111/j.1467-1770.1994.tb01104.x

Tigchelaar, M. (2019). Exploring the relationship between self-assessments and OPIc ratings of oral proficiency in French. In P. Winke & S. M. Gass (Eds.), *Foreign language proficiency in higher education* (pp. 153–173). Springer. https://doi.org/10.1007/978-3-030-01006-5_9

Tigchelaar, M., Bowles, R., Winke, P., & Gass, S. (2017). Assessing the validity of ACTFL can-do statements for spoken proficiency. *Foreign Language Annals*, *50*, 584–600. https://doi.org/10.1111/flan.12286

Tremblay, A. (2011). Proficiency assessment in second language acquisition research: "Clozing" the gap. *Studies in Second Language Acquisition*, *33*, 339–372. https://doi.org/10.1017/S0272263111000015

Wainer, H. (2000). Introduction and history. In N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computer adaptive testing: A primer* (2nd ed., pp. 1–21). Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computer adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–201. https://www.jstor.org/stable/1434630

Winke, P., Yan, X., & Lee, S. (in press). What does the Cloze test really test? A cognitive validation of a French Cloze test with eye-tracking and interview data. In G. Yu & J. Xu (Eds.), *Language test validation in a digital age*. UCLES/Cambridge University Press.

Winke, P., & Zhang, X. (2022). *Data and codebook for SSLA article: "A closer look at a marginalized test method: Self-assessment as a measure of speaking proficiency."* [Data set]. Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E164981V1

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*, 497–528. https://doi.org/10.1177/0265532215594643

Cite this article: Winke, P., Zhang, X. and Pierce, S. J. (2023). A closer look at a marginalized test method: Self-assessment as a measure of speaking proficiency. *Studies in Second Language Acquisition,* **45,** 416–441. https://doi.org/10.1017/S0272263122000079