Language Teaching (2024), **57**, 42–56
doi:10.1017/S0261444822000489

CAMBRIDGE UNIVERSITY PRESS

RESEARCH TIMELINE

Self-assessment in second language learning

Yuko Goto Butler

University of Pennsylvania, Philadelphia, USA
Email: ybutler@gse.upenn.edu

(Received 11 November 2022; accepted 19 November 2022)

Introduction

Self-assessment (SA), as an activity for reflecting on one's own performance and abilities (Black & Wiliam, 1998), has been a topic of interest to educators over the years. Among second language (L2) educators, SA began growing in popularity in the 1970s and 1980s, when L2 educators' focus shifted from analyzing linguistic systems to examining how learners learn a language. Many can-do statements and SA descriptors have been developed for L2 language learning, including SA grids aligned with the Common European Framework of Reference (CEFR, Council of Europe, 2022) and can-do statements prepared by the American Council on the Teaching of Foreign Language (ACTFL) in collaboration with the National Council of State Supervisors for Languages (NCSSFL) (ACTFL, n.d.). Textbooks and other L2 learning materials, including online apps, often contain SA items. SA can be used in conjunction with other assessments, such as traditional objective assessments, peer assessments, and portfolios. Teachers are often encouraged to incorporate SA into their curricula as part of the promotion of constructivist approaches to education, which have been particularly popular since the late 1980s (e.g., Nunan, 1988; Tarone & Yule, 1989); SA resonates well with modern learning theories such as learner-centered education, self-regulated learning, and AUTONOMOUS LEARNING (Butler, in press).

Despite widespread promotion of SA through policy and curricular initiatives, the actual implementation of SA in language classrooms varies considerably, and SA is not often used as effectively as expected in practice (Bullock, 2011; Nikolov & Timpe-Laughlin, 2020). Reluctance to use SA in classrooms may be owing, in part, to users' perception of SA; for example, teachers may be skeptical about the accuracy of their students' SA, and students may not see SA as helpful for their learning (e.g., Mäkipää, 2021*).

Mixed views of SA among L2 educators and students may partially stem from the fact that SA entails multiple functions and purposes. Broadly speaking, varied definitions of SA reflect two major functions. One common focus is on the measurement functions of SA, namely, ASSESSMENT OF LEARNING. As exemplified in Bailey's (1998) definition of SA, "procedures by which learners themselves evaluate their language skills and knowledge" (p. 227), some researchers emphasize its measurement functions. The other major focus is on the aspects of SA that support learning, or ASSESSMENT FOR LEARNING. An example of the latter can be seen in Andrade and Valtcheva's (2009) definition: "a process of formative assessment during which students reflect on the quality of their work, judge the degree to which it reflects explicitly stated goals or criteria, and revise accordingly" (p. 13). SA can be used for both summative purposes (i.e., attributing values or scores to one's learning outcome, primarily for grading) or formative purposes (i.e., monitoring, or self-reflecting on, the ongoing process of learning), and the aforementioned definitions of ASSESSMENT OF LEARNING and ASSESSMENT FOR LEARNING roughly correspond to the summative and formative purposes of its use, respectively.

*Indicates full reference is described in the subsequent timeline.

The purpose of this timeline is to review major research on SA in L2 learning (including foreign language [FL] learning) conducted in the last 30 years and to illustrate how the field came to better understand the use of SA both as a measurement tool and a learning/teaching tool for L2 learning. Because of space constraints, the timeline is limited to select studies that were published in English in major academic journals and book chapters; I selected major studies that were highly cited and/or provided important new insights that influenced successive research. While many studies also examined in-service or pre-service teachers' SA of their teaching performance or language proficiencies as well as their attitudes towards SA, these studies are excluded in the timeline below. The selected studies are categorized according to the following themes:

A. Assessment of learning orientation
  A1. Theoretical frameworks
  A2. Learners' perception
  A3. Reliability and validity
  A4. Variables influencing students' SA
  A5. SA development and implementation
  A6. Meta-analyses, qualitative reviews, etc.
B. Assessment for learning orientation
  B1. Theoretical frameworks
  B2. Effectiveness of SA on learning and self-regulation
  B3. Innovative use of SA (e.g., SA as a social activity, via technology)
  B4. Meta-analyses, qualitative reviews, etc.
C. Targeted age groups
  C1. Young learners (up to primary school)
  C2. Secondary school students
  C3. Adults
  C4. General or unspecified

Reflecting the strong psychometric tradition of language assessment since the introduction of modern assessment theories in language education in the 1960s (e.g., Carroll, 1968; Lado, 1961), research on SA in L2 learning has largely examined the efficacy of SA from a measurement point of view until relatively recently (Category A in the timeline). More specifically, researchers were interested in examining the reliability and validity of SA. A few studies examined the reliability and validity of can-do statements or descriptors, including ACTFL can-do statements (Brown et al., 2014*; Ma & Winke, 2019*; Malabonga et al., 2005*; Summers et al., 2019*; Tigchelaar, 2019*; Tigchelaar et al., 2017*); CEFR descriptors (Little, 2005*); and the Diagnostic Language Assessment System (DIALANG), which was developed based on the CEFR (Brantmeier et al., 2012*; Luoma & Tarnanen, 2003*; Ünaldi, 2016*; see https://dialangweb.lancaster.ac.uk for the items in DIALANG). While studies with measurement-oriented approaches are still very popular, in the last decade or so a growing number of studies have examined SA as a learning/teaching tool (Category B in the timeline). In these studies, researchers were interested in understanding how best to implement SA to maximize its effect on students' L2 learning.

As a measurement tool, SA generally has moderate correlations with external assessments, according to meta-analyses (Li & Zhang, 2021*; Ross, 1998*). Thus, depending on the purpose of its use and the importance placed on it (i.e., whether it is a high-stakes context), SA can replace or complement other external assessments (e.g., teachers' assessments and objective language measures) (Malabonga et al., 2005*), although it may not be as reliable and valid as peer assessment (PA) (Matsuno, 2009*; Patri, 2002*). SA can also be used as a reasonably reliable measure of one's learning progress over time (Brown et al., 2014*).

It is important to note, however, that there are substantial variabilities in the accuracy of learners' SA across studies. Three types of variables can influence the accuracy of SA by L2-learning students:

https://doi.org/10.1017/S0261444822000489 Published online by Cambridge University Press

(a) variables related to item construction and administration, (b) learner-related variables, and (c) external or environmental variables. The variables related to item construction and administration include item wording (Ross, 1998*; Tigchelaar et al., 2017*) and response formats (e.g., can-do or dichotomous formats vs. Likert-scale formats) (Butler, 2018a*). Compared with general and holistic descriptions, specific descriptors, particularly descriptors consistent with learners' experiences, tend to increase accuracy (Butler, 2018b*; Butler & Lee, 2006*; Edele et al., 2015*; Ross, 1998*; Suzuki, 2015*). Other factors that matter include the point of reference that learners relied on when self-evaluating (Butler, 2018a*, 2018b*; Moritz, 1996*; Swain & Hart, 1993*) and the tasks or skill domains being assessed (Bachman & Palmer, 1989*; Brantmeier et al., 2012*; Ross, 1998*). Influential learner-related variables include learners' L2 proficiency (AlFallay, 2004*; Brantmeier et al., 2012*; Dolosic et al., 2016*; Ma & Winke, 2019*; Matsuno, 2009*; Ross, 1998*, Ünaldi, 2016*), age (Butler, 2018a*, 2018b*; Butler & Lee, 2006*), attitudes and personality factors such as self-esteem (AlFallay, 2004*), and learning experience (Suzuki, 2015*). Finally, external or environmental factors—including cultural environments (Blanche & Merino, 1989*; Matsuno, 2009*) and heritage or nonheritage learning contexts (Ashton, 2014*)—seem to play significant roles as well. Researchers have documented response biases associated with various learner characteristics. For example, lower proficiency students or students with less experience with language learning tend to overestimate their abilities—a phenomenon often referred to in psychology as the Dunning-Kruger effect (Dunning et al., 2003) (Heilenman, 1990*; Lappin-Fortin & Rye, 2014*; Suzuki, 2015*; Trofimovich et al., 2016*; Ünaldi, 2016*). Similarly, younger children tend to overestimate their abilities (Butler & Lee, 2006*). Learners tend to be more strict when evaluating their own performance compared with assessing their peers' performance (Matsuno, 2009*; Tigchelaar, 2016*). Finally, in certain cultures, people might be expected to be humble when self-assessing their abilities and performance (Edele et al., 2015*; Matsuno, 2009*).

As the following timeline illustrates, over time researchers have shown increasing interest in understanding SA's role not only as a measurement tool but also as a learning and teaching tool; namely, how it affects students' L2 learning, self-regulation, and self-efficacy. SELF-REGULATION refers to one's ability to control one's cognition, affect, and behaviors to achieve a goal, and SELF-EFFICACY means one's confidence in ability to perform relevant actions to accomplish a goal. SA is thought to promote learners' self-regulation because it can help them set goals and criteria, monitor their performance, reflect on their performance, and internalize the whole learning experience. SA can improve learners' self-efficacy by helping them understand the requirements of targeted tasks, which can in turn improve the likelihood that they will successfully complete the task (Butler, in press).

Qualitative or mixed methods have been employed to uncover the process of learning and/or learners' and teachers' perceptions and experiences of using SA as a learning/instructional tool. As predicted, studies have found that SA improves learners' self-reflection on their abilities and performance and leads to greater self-efficacy (Blanche & Merino, 1989*; Brantmeier et al., 2012*; Butler & Lee, 2010*; Glover, 2011*; Jang et al., 2015*; Kissling & O'Donnell, 2015*). As with the accuracy of SA in measurement-oriented studies, in learning-oriented studies, the effects of SA on learning were also influenced by several variables, including the duration of the SA intervention and the wording and structures of the rubrics used (Wang, 2017*). Students' perception of SA as a L2 learning/instructional tool is generally positive if the criterion is clearly provided and/or some form of training (including repeated use of SA) is offered (Babaii et al., 2016*; De Saint Léger, 2009*; Glover, 2011*; Hung, 2019*; Sullivan & Lindgren, 2002*). With guidance, repeated use of SA can not only improve students' perceptions of their L2 learning but also lead to actual learning gains, as measured by external assessments (Butler & Lee, 2010*). While the importance of feedback—including self-feedback—through SA on one's learning is acknowledged, its effect is a complicated combination of factors that includes one's previous experiences and future goal setting, aspirations, and self-confidence (Butler, 2018a*, 2019b*; Huang, 2016*; Tigchelaar, 2016*).

Although college students in classroom settings have historically been the primary target of studies on SA, recent studies have considered more diverse populations such as young learners (Ashton,

https://doi.org/10.1017/S0261444822000489 Published online by Cambridge University Press

2014*; Butler, 2018a*, 2018b*; Butler & Lee, 2006*, 2010*; Dolosic et al., 2016*; Jang et al., 2015*; Liu & Brantmeier, 2019*) and immigrants (Edele et al., 2015*). SA is increasingly administered through computers, and computer-administrated SA appears to increase accuracy (Li & Zhang, 2021). Moreover, while the general conceptualization of SA as an "internal or self-directed" cognitive activity (Oscarson, 1989*, p. 1) has been dominant, researchers have started paying greater attention to the social and emotional aspects of SA rather than viewing it as a purely individual cognitive activity (Andrade & Brown, 2016; Butler, in-press). Most recently, SA is used to evaluate L2 learners' intercultural communication proficiency as part of communicative competence (e.g., Lenkaitis, 2021*).

In sum, SA has gained the attention of L2 researchers and educators in the last couple of decades, both as a potential measurement and learning/teaching tool. Several variables that are influential for accuracy (as a measurement tool) and learning (as a learning/teaching tool) have been identified. Most recently, research on SA is more diversified in terms of its target population and means of administration (e.g., computer-administered SA), and it takes a more ecological perspective, viewing SA as a social activity as well as an individual cognitive activity.

References

American Council on the Teaching of Foreign Language. (n.d.). *NCSSFL-ACTFL can-do statements*. https://www.actfl.org/resources/ncssfl-actfl-can-do-statements

Andrade, H. L., & Brown, G. T. L. (2016). Student self-assessment in the classroom. In G. T. L. Brown, & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 319–334). Routledge.

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice*, *48*(1), 12–19. doi:10.1080/00405840802577544

Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions, and directions*. Heinle & Heinle.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, *5*(1), 7–74. doi:10.1080/0969595980050102

Bullock, D. (2011). Learner self-assessment: An investigation into teachers' beliefs. *ELT Journal*, *65*(2), 114–127. doi:10.1093/elt/ccq041

Butler, Y. G. (in-press). Expanding the role of self-assessment: From assessing to learning English. In D. Valente, & D. Xerri (Eds.), *Innovative practices in early English language education*. Palgrave Macmillan.

Carroll, J. B. (1968). The psychology of language testing. In A. Davies (Ed.), *Language testing symposium: A psycholinguistic approach* (pp. 46–69). Oxford University Press.

Council of Europe. (2022). *Self-assessment grid-Table 2 (CEFR 3.3): Common reference levels*. https://www.coe.int/en/web/common-european-framework-reference-languages/table-2-cefr-3.3-common-reference-levels-self-assessment-grid

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*(3), 83–87. doi:10.1111/1467-8721.01235

Lado, R. (1961). *Language testing: The construction and use of foreign language test. A teachers' book*. McGraw-Hill Book Company.

Nikolov, M., & Timpe-Laughlin, V. (2020). Assessing young learners' foreign language abilities. *Language Teaching*, *54*(1), 1–37. doi:10.1017/S0261444820000294

Nunan, D. (1988). *The learner-centered curriculum: A study in second language teaching*. Cambridge University Press.

Tarone, E., & Yule, G. (1989). *Focus on the language learner*. Oxford University Press.

Yuko Goto Butler is Professor of Educational Linguistics at the Graduate School of Education at the University of Pennsylvania. She is also the director of Teaching English to Speakers of Other Languages (TESOL) program at Penn. Her research interests are primarily focused on the improvement of second/foreign language education among young learners in the U.S. and Asia in response to the diverse needs of an increasingly globalizing world. Her work has also focused on identifying effective English-as-a-second language/English-as-a-foreign-language (ESL/EFL) teaching and learning strategies and assessment methods that take into account the relevant linguistic and cultural contexts in which instruction takes place.

Cite this article: Butler, Y. G. (2024). Self-assessment in second language learning. *Language Teaching*, *57*(1), 42–56. https://doi.org/10.1017/S0261444822000489

Year	References	Annotations	Theme
1989	Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing, 6*(1), 14–29. https://doi.org/10.1177/026553228900600104	**Bachman & Palmer's (1989)** classic psychometric investigation found that SA is a valid and reliable measure. The study, conducted among 116 English-learning adults, also found that students' self-ratings (on a four-point scale) of their grammatical competence were a better indicator of the trait compared with pragmatic and sociolinguistic competencies. Additionally, items that asked participants to rate their perceived difficulty in production were more effective than other types of items that appeared in can-do statements (composed of affirmative statements).	**A3, C3**
1989	Blanche, P., & Merino, B. (1989). Self-assessment of foreign language skills: Implications for teachers and researchers. *Language Learning, 39*(3), 313–340. https://doi.org/10.1111/j.1467-1770.1989.tb00595.x	**Blanche & Merino (1989)** was one of the first comprehensive narrative reviews of SA in L2 learning. Focusing on foreign language skills of adult learners, the researchers reported that studies on the accuracy of SA are "somewhat contradictory" (p. 326) and vary based on skill domains and students' cultural backgrounds. The authors drew on their findings to offer suggestions for future research (e.g., examine the effects of learners' age and personality, the amount of instruction received, and task types; use SA questionnaires to capture students' learning progress) as well as implications for teachers (e.g., regular administration of SA for improving accuracy). Several subsequent studies built on those suggestions (e.g., AlFallay, 2004[1]; Baleghizadeh & Masoun, 2013; Butler & Lee, 2006; Heilenman, 1990; Patri, 2002).	**A1, A3, A4, C3**
1989	Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing, 6*(1), 1–13. https://doi.org/10.1177/026553228900600103	In this influential paper, **Oscarson (1989)** described and justified the value of adopting SA in L2/FL education and addressed the potential of SA for promoting student learning. The paper also offered examples of different types of SA.	**B1, B3, C4**
1990	Heilenman, L. K. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing, 7*(2), 174–201. https://doi.org/10.1177/026553229000700204	Using a method called the split-ballot procedure, **Heilenman (1990)** examined 232 French-learning college students' response biases in their SA. The learners tended to overestimate in all of the domains that were assessed (e.g., grammar, vocabulary), but this tendency was more apparent among students who were relatively new to French learning (cf., Blanche & Merino, 1989).	**A3, A4, C3**
1993	Swain, M., & Hart, D. (1993). Self-assessment in two French immersion programmes. *Applied Linguistics, 14*(1), 25–42. https://doi.org/10.1093/applin/14.1.25	Following Bachman & Palmer (1989), **Swain & Hart (1993)** also examined the validity and reliability of SAs, but this study concerned Grade 8 students who were in two types of French immersion programs ($N = 26$): namely, early and late immersion programs. Two types of benchmarks were also used for SA: one was the perceived French proficiency against their Francophone peers; the other was perceived task difficulty in French. Although the correlations between SAs and objective measures were generally low in both immersion programs, the students' SA was more accurate when the benchmark for SA was set against specific tasks than when it was set against native peers.	**A3, C2**

[1]*Note.* Authors' names are shown in small capitals when the study referred to appears elsewhere in this timeline.

1996	Moritz, C. E. F. (1996, March 24). *Students' self-assessment of language proficiency*. 18th AAAL Conference, Chicago, USA [Paper presentation]. https://eric.ed.gov/?id=ED399771	This conference paper by **Moritz (1996)** was one of the first studies in applied linguistics to examine the process by which L2 learners respond to SA items, rather than examining the assessment of learning outcomes. Twenty-eight French-learning college students engaged in a think-aloud protocol as well as a semi-structured interview while and after responding to SA items. Moritz found that the students used varied reference points when evaluating their French abilities, including using Social Category (e.g., comparing their abilities to those of their classmates), Meaningful Other Category (e.g., judging their abilities against those of a native-speaker acquaintance), and Autobiographical Category (e.g., comparing their current performance to their own past performance). Also, see Swain & Hart (1993) mentioned above.	**A3, A4, C3**
1998	Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experimental factors. *Language Testing, 15*(1), 1–20. https://doi.org/10.1177/026553229801500101	**Ross (1998)** was one of the first quantitative reviews (a meta-analysis) of SA in L2 learning, complementing the narrative review by Blanche & Merino (1989). The review, which is the first part of Ross's two-part study, found that SA is generally accurate (the average correlation was 0.65), but it identified substantial variabilities in effect sizes across studies. The average effect sizes were larger in receptive skills (listening and reading) than in productive skills (speaking and writing). The second part of Ross (1998) is a report on an empirical study conducted in an adult language training program in Japan. The students' SA accuracy was found to be mediated by their experience with the language learning tasks that were assessed by the SA.	**A3, A4, C3, A5**
2002	Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing, 19*(2), 109–131. https://doi.org/10.1191/0265532202lt224oa	**Patri (2002)** examined the effect of peer feedback on the accuracy of students' SA and peer assessment (PA) of an English oral presentation task in comparison with the teachers' assessment ($N = 56$, Hong Kong college students). All participants received a training session in which they learned the criteria for SA and PA. The results indicated that peer feedback, which was offered before the SA and PA were administered, contributed to making the accuracy of the PA similar to that of the teachers' assessment. However, the peer feedback did not have the same effect on the accuracy of SA (cf., Blanche & Merino, 1989).	**A3, C3**
2002	Sullivan, K., & Lindgren, E. (2002). Self-assessment in autonomous computer-aided second language writing. *ELT Journal, 56*(3), 258–266. https://doi.org/10.1093/elt/56.3.258	In this case study of four adult learners of English in Sweden, **Sullivan & Lindgren (2002)** used a computer logging program, J Edit, with the goal of promoting students' self-assessment of and self-reflection on their essay writing. In line with Oscarson (1989), this method assisted the students in reflecting on and revising their essays. The participants reported that the method provided them with useful insights into their writing.	**B2, B3, C3**
2003	Luoma, S., & Tarnanen, M. (2003). Creating a self-rating instrument for second language writing: From idea to implementation. *Language Testing, 29*(4), 440–465. https://doi.org/10.1191/0265532203lt267oa	Inspired by Oscarson (1989) and his successive works, a self-rating instrument was developed as part of DIALANG, an internet-based diagnostic language assessment system aligned with the Common European Framework of Reference for Language (CEFR). **Luoma & Tarnanen (2003)** focused on the development of the writing portion of the self-rating instrument. They described the trial-and-error process of developing the self-rating instruments based on a usability study conducted among six adult learners of Finnish as their L2.	**A5, C3**

(*Continued*)

(*Continued*)

Year	References	Annotations	Theme
2004	AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System, 32*(3), 407–425. https://doi.org/10.1016/j.system.2004.04.006	In a study of 78 Saudi Arabian college students, **AlFallay (2004)** examined the influence of the students' psychological and personality traits on the accuracy of the SA and PA of their English oral skills, in relation to the teachers' assessment (TA). (See future research suggestions made by Blanche & Merino, 1989.) The students tended to overestimate their peer's performance compared with that of their own. Students with low self-esteem most accurately self-assessed their own oral ability, whereas students with higher instrumental motivation self-assessed their oral skills the least accurately.	**A3, A4, C3**
2005	Little, D. (2005). The Common European Framework and the European Language Portfolio: Involving learners and their judgments in the assessment process. *Language Testing, 22*(3), 321–336. https://doi.org/10.1191/0265532205lt311oa	Because the CEFR and its by-product, the English Language Portfolio (ELP), were not designed specifically for young learners (YLs), age-appropriate adjustments are necessary before using them with YLs. In response to the growing number of English-as-a-second language (ESL) children in Ireland, **Little (2005)** reported on an in-progress project in which he and his team developed an ESL curriculum aligned with the CEFR (the English Language Proficiency Benchmark) and revised ELP checklists that essentially served as self-assessment for young ESL learners. See also Luoma & Tarnanen (2003) for a study describing a process of SA item development based on the CEFR for adults.	**A5, C1, C2**
2005	Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing, 22*(3), 337–354. https://doi.org/10.1191/0265532205lt312oa	Similar to Little (2005), **Hasselgreen (2005)** described another adaptation process of the CEFR for YLs using two projects (the Bergen 'can-do' project and the National Testing project) in Norway. She discussed a number of challenges, among them maintaining "the integrity of the CEFR levels" while also accounting for "the particular characteristics of children and younger teenagers" (p. 351).	**A5, C1, C2**
2005	Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing, 22*(1), 59–92. https://doi.org/10.1191/0265532205lt297oa	**Malabonga et al. (2005)** reported on two studies, the first of which concerns SA. The SA study examined the accuracy of college students' SA when choosing an appropriate starting level on a computer-adaptive test called the Computerized Oral Proficiency Instrument (COPI). COPI aligns with the ACTFL Proficiency Scale as well as the Stimulated Oral Proficiency Interview (SOPI). Fifty college students learning Arabic, Chinese, or Spanish participated in the study. The results indicated that the SA was reliable (e.g., students' SA results were stable). Their SA was also highly correlated with their actual performance in both COPI and SOPI, consistent with Bachman & Palmer (1989) and Ross (1998). Additionally, the students could accurately choose their COPI starting level.	**A3, C3**
2006	Brantmeier, C. (2006). Advanced L2 learners and reading placement: Self-assessment, CBT, and subsequent performance. *System, 34*(1), 15–35. https://doi.org/10.1016/j.system.2005.08.004	While Heilenman (1990) identified a response bias among college students' SA, particularly among less-experienced students, **Brantmeier (2006)** examined the accuracy of advanced college students' SA in reading and found that it was not accurate enough for use as a placement test or a predictor of their subsequent performance. Brantmeier suggested the need for more investigations of factors that influence the results of SA even for advanced learners.	**A3, C3**

2006	Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessment among Korean elementary school students studying English. *Modern Language Journal, 90*(4), 506–518. https://doi.org/10.1111/j.1540-4781.2006.00463.x	Similar to Little (2005) and Hasselgreen (2005), **Butler & Lee (2006)** focused on SA for young language learners (in this case, English-as-a-foreign language learners [EFL] in South Korea, ages 9–10 and 11–12). The researchers examined the validity of two types of administrations of SA in relation to a standardized external test and a teacher's judgment. The types of SA were a holistic SA (referred to as 'off-task' SA) and a contextualized SA ('on-task' SA; items were designed for specific tasks and were administered immediately after the tasks). The children could self-assess their performance more accurately in the on-task condition (cf. Ross, 1998). Moreover, the on-task condition was less influenced by the children's attitudes and personality factors. Age was also a factor, in that older children self-assessed their L2 performance more accurately than younger children.	**A3, C1**
2008	Dlaska, A., & Krekeler, C. (2008). Self-assessment of pronunciation. *System, 36*(4), 506–516. https://doi.org/10.1016/j.system.2008.03.003	**Dlaska & Krekeler (2008)** examined the accuracy of advanced adult German learners' (N = 46) SA of their pronunciation of select German sounds compared with expert raters' judgments. The study found that, although the agreement was relatively high (the agreement coefficient was .85), the learners were stricter than the raters (c.f., Heilenman, 1990). The paper also suggested some causes of difficulties that even advanced learners have when self-assessing pronunciation.	**A3, C3**
2009	De Saint Léger, D. (2009). Self-assessment of speaking skills and participation in a foreign language class. *Foreign Language Annals, 42*(1), 158–178. https://doi.org/10.1111/j.1944-9720.2009.01013.x	Focusing on the learning potential of SA (Oscarson, 1989), **De Saint Léger (2009)** conducted an exploratory study in which advanced French-learning college students (N = 32) self-assessed their speaking performance multiple times during a semester. The students' self-perceived proficiency (fluency, vocabulary, and overall confidence, in particular) increased over time. The interview data revealed that the learners thought that their goal awareness increased as a result of SA, but it remained unclear if their increased awareness led to actual linguistic improvement.	**B2, C3**
2009	Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75–100. https://doi.org/10.1177/0265532208097337	**Matsuno (2009)** highlighted the role of culture in students' SA responses. Employing multifaceted Rasch measurement, Matsuno investigated the accuracy of Japanese college students' SA and PA on their English writing, using teacher assessments for comparison. Similar to AlFallay (2004), students, high achievers in particular, were overly critical in their self-ratings; the author attributes this finding to the high value placed on modesty in Japanese culture. When assessing their peers' writing performance, students were less critical. Students' PA ratings were more internally consistent than their SA ratings and were independent of their own writing performance (e.g., higher achievers were not necessarily more critical in their PA). The author suggested the possibility of using PA as a replacement for teacher assessment but cautioned against using SA for such purposes. See Patri (2002) for a similar recommendation.	**A3, C3**

(*Continued*)

(*Continued*)

Year	References	Annotations	Theme
2010	Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing, 27*(1), 5–31. https://doi.org/10.1177/0265532209346370	Building on De Saint Léger (2009), **Butler & Lee (2010)** conducted an intervention study to directly examine the effectiveness of SA on learners' learning as well as on their attitudes. Different from De Saint Léger (2009), the study focused on young learners (ages 11–12, *N* = 254). The researchers found that the repeated use of SA improved not only the accuracy of the children's SA responses but also their language performance (measured by a standardized test) and confidence (although the effect sizes were relatively low). Additionally, depending on learning/teaching contexts (e.g., differences in SES backgrounds), both the students and their teachers perceived the effectiveness of SA differently, and their perceptions influenced how they implemented SA.	**B2, C1**
2011	Glover, P. (2011). Using CEFR level descriptors to raise university students' awareness of their speaking skills. *Language Awareness, 29*(2), 121–133. https://doi.org/10.1080/09658416.2011.555556	Unlike previous studies concerning the development of CEFR-based SA (Hasselgreen, 2005; Little, 2005; Luoma & Tarnanen, 2003), **Glover (2011)** examined how learning about the Common Reference Levels (CRL) of the CEFR helped Turkish university students (pre-service English teachers, *N* = 62) self-assess their speaking abilities. Knowing about the CRL helped the students describe their speaking abilities critically with greater detail and increase their confidence and perceived learning.	**B2, C3**
2012	Brantmeier, C., Vanderplank, R., & Strube, M. (2012). What about me? Individual self-assessment by skill and level of language instruction. *System, 40*(1), 144–160. https://doi.org/10.1016/j.system.2012.01.003	Using modified questions from DIALANG (cf., Luoma & Tarnanen, 2003), **Brantmeier et al. (2012)** examined the accuracy of SA among 276 Spanish-learning college students. When SA was administered as a criterion-referenced instrument tailored to the course objectives, the students could accurately self-assess their four skills, although there were some variations across skill domains and proficiency levels (cf., Ross, 1998). The SA also accurately predicted advanced learners' performance across the skills.	**A3, C3**
2013	Baleghizadeh, S., & Masoun, A. (2013). The effect of self-assessment on EFL learners' self-efficacy. *TESL Canada Journal, 31*(1), 42–58. https://doi.org/10.18806/tesl.v31i1.1166	**Baleghizadeh & Masoun (2013)** examined the role of repeated use of SA on English-learning Iranian college students' self-efficacy (*N* = 57). Adapted SA items from Blanche & Merino (1989) were used. The treatment group enhanced their self-efficacy as a result of using SA, which is consistent with findings by De Saint Léger (2009) and Butler & Lee (2010).	**B2, C3**
2014	Ashton, K. (2014). Using self-assessment to compare learners' reading proficiency in a multilingual assessment framework. *System, 42*, 105–119. https://doi.org/10.1016/j.system.2013.11.006	**Ashton (2014)**, focusing on 439 secondary school students (ages 12–15) learning German, Japanese, or Urdu, examined the accuracy of SA in relation to their teachers' assessment and a reading test. Three factors of learners' self-rated reading proficiency (factors related to personal communication, higher order cognitive functions, and locating specific details for comprehension) predicted the students' reading performance across the three language groups. However, depending on the learning contexts (heritage speakers or classroom learners), the students' SA response patterns (overestimating or underestimating their ability) differed (c.f., Brantmeier et al., 2012; Heilenman, 1990; Ross, 1998). Linguistic factors (e.g., Latin or non-Latin scripts) may also have been responsible for the way that SA represented the students' learning progression.	**A3, C2**

2014	Brown, N. A., Dewey, D. P., & Cox, T. L. (2014). Assessing the validity of can-do statements in retrospective (then-now) self-assessment. *Foreign Language Annals, 47*(2), 261–285. https://doi.org/10.1111/flan.12082	Using a retrospective method (students respond to SA after a given intervention), **Brown et al. (2014)** examined the reliability and predictive validity of SA based on ACTFL can-do statements. Thirty-six intermediate to advanced college learners of Russian completed the SA and the ACTFL Oral Proficiency Interview (OPI) before and after an internship in Russia. The study found that the SA had a high degree of reliability. The SA item difficulty levels mapped well with the ACTFL scales (e.g., superior-level items were more difficult than advanced-level items, etc.) although the means of items representing each ACTFL level were not significantly different. Finally, significant gains in both OPI and SA were obtained after the internship, while correlations between OPI and SA were not significant. Overall, the authors concluded that the ACTFL can-do items can be reliably used to obtain learners' perceived gains over time (cf., Malabonga et al., 2005).	**A3, C3**
2014	Lappin-Fortin, K., & Rye, B. J. (2014). The use of pre-/posttest and self-assessment tools in a French pronunciation course. *Foreign Language Annals, 47*(2), 300–320. https://doi.org/10.1111/flan.12083	In an intermediate-level college French course ($N = 48$), **Lappin-Fortin & Rye (2014)** investigated the accuracy of the students' SA focusing on pronunciation. Pretest and post-test scores, as well as SAs administered with the pre- and post-tests, were examined. Their analyses indicated that the students' SA was reasonably accurate in relation to both pre- and post-test scores judged by experts, but the accuracy of SA varied depending on the components; students more accurately assessed their pronunciation of liaisons than their pronunciation of vowels and prosody. In general, the students overestimated their performance (cf., Ashton, 2014; Blanche & Merino, 1989; Heilenman, 1990; Dlaska & Krekeler, 2008).	**A3, C3**
2015	Edele, A., Seuring, J., Kristen, C., & Stanat, P. (2015). Why bother with testing? The validity of immigrants' self-assessed language proficiency. *Social Science Research, 52,* 99–123. https://doi.org/10.1016/j.ssresearch.2014.12.017	**Edele et al. (2015)** investigated the validity of two types of SA (estimates of general ability and concrete performance) in adult immigrants' L1 and L2 in relation to language tests. In total, 1,300 9th-grade students in Germany whose L1 was either Russian or Turkish participated in the study. The study found that SA of concrete performance correlated more highly with objective test scores than SA of general ability but that both types of SA were systematically biased against certain groups (e.g., boys and Turkish-origin students overestimated their abilities while students with higher cognitive abilities underestimated their abilities) (cf., Brantmeier et al., 2012; Butler & Lee, 2006; Ross, 1998).	**A3, C2**
2015	Jang, E.E., Dunlop, M., Park, G., & van der Boom, E. H. (2015). How do young students with different profiles of reading skills mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing, 32*(3), 359–383. https://doi.org/10.1177/0265532215570924	**Jang et al. (2015)** examined how holistic diagnostic feedback (HDF) was processed, perceived, and used by young learners ($N = 44$, ages 11–12) in Canada. The HDF reports indicated the children's level of mastery and their self-assessment proficiency level. The children were also invited to respond to the goal indicated in the report, followed by individual conferences with the teacher. The study found that the way that children processed the HDF was influenced by their perception of their ability as well as their orientation to learning. Although the study did not focus on the effect of SA per se, it shows that the SA helped the children facilitate deeper reflections on their abilities even though they had difficulties accurately self-assessing their abilities (cf., Butler & Lee, 2010).	**B2, C1**
2015	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*(4), 283–302. https://doi.org/10.1080/09658416.2015.1099659	**Kissling & O'Donnell (2015)** examined the effect of using the ACTFL proficiency guidelines (ACTFL PGs) for oral production on students' language awareness and self-efficacy. The study was conducted among 13 college students of foreign language (intermediate- to advanced-level Spanish learners) over a semester. The students' self-assessment narratives revealed that the ACTFL PGs assisted the learners in being more aware of different aspects of their speech and better able to articulate their weaknesses and strengths (cf., Brown et al., 2014; Malabonga et al., 2005 for studies concerning ACTFL).	**B2, C3**

(*Continued*)

(*Continued*)

Year	References	Annotations	Theme
2015	Suzuki, Y. (2015). Self-assessment of Japanese as a second language: The role of experiences in the naturalistic acquisition. *Language Testing, 32*(1), 63–81. https://doi.org/10.1177/0265532214541885	**Suzuki (2015)** examined the role of experiences (length of residence in the target country and reading experiences) on adult learners' SA accuracy among 63 advanced Japanese learners with Chinese backgrounds in Japan. The data indicated that the learners' experiences played a significant role in their accuracy of SA in an immersion context, a finding that is consistent with previous studies conducted in classroom contexts (see Ross, 1998, and Butler & Lee 2006); there was a tendency for experienced learners to underestimate their abilities, whereas the reverse tendency was found among less-experienced learners (i.e., the Dunning-Kruger effect).	**A3, C3, A5**
2016	Babaii, E., Taghaddomi, S., & Pashmforoosh, R. (2016). Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing, 33*(3), 411–437. https://doi.org/10.1177/0265532215590847	**Babaii et al. (2016)** examined the effect of providing Iranian college students ($N = 29$) with training (giving them scoring criteria for L2 speaking and a chance to practice using the criteria) on their SA responses. The results showed that the training improved the accuracy of SA in relation to the teachers' assessment, leading to narrowing mismatches between the learners' and the teachers' evaluations (cf., Butler & Lee, 2010). Moreover, the students generally had positive views of the effectiveness of the training for improving their SA results; They believed that SA provided an opportunity to raise their self-awareness and had the potential for positive long-term effects.	**A3, B2, C3**
2016	Butler, Y. G. (2016). Self-assessment of and for young learners' foreign language learning. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 291–315). Springer.	Focusing on young learners, **Butler (2016)** discussed the two orientations of SA research (i.e., assessment of learning and assessment for learning) and reviewed both theoretical and empirical studies concerning the use of SA among young L2 learners (cf., Oscarson, 1989). This paper also examined and classified major types of SA by paying attention to five dimensions that characterize existing SAs (i.e., domain setting, scale setting, goals setting, the focus of assessment, and the method of assessment, p. 291) for assisting educators and researchers in developing SA for young learners.	**A1, B1, C1**
2016	Dolosic, H. N., Brantmeier, C., Strube, M., & Hogrebe, M. C. (2016). Living language: Self-assessment, oral production and domestic immersion. *Foreign Language Annals, 49*(2), 302–316. https://doi.org/10.1111/flan.12191	**Dolosic et al. (2016)** examined the accuracy of SA among 24 adolescents (ages 14–18) conducted before and after they joined a French language summer camp. Oral production tasks indicated that the students improved their ability over the course of the summer camp. The study also found that, in comparison to the actual oral production performance, students could self-assess their performance by the end of the summer camp, while they could not do so before the camp (cf., Butler & Lee, 2006; Brantmeier et al., 2012).	**A3, C2**
2016	Huang, S.-C. (2016). Understanding learners' self-assessment and self-feedback on their foreign language speaking performance. *Assessment & Evaluation in Higher Education, 41*(6), 803–820. http://dx.doi.org/10.1080/02602938.2015.1042426	**Huang (2016)** examined how college students in Taiwan ($N = 50$) self-assess or offer self-feedback on their speaking performance. The students were asked to listen to, transcribe, and analyze their own recorded speeches as well as to make statements of actions for future improvement. Qualitative analyses revealed that the learners' self-feedback was "multifaceted" (p. 803) in that it contained not only their reflections on the speech performance at hand but also their learning history and predictions for the next step. The author indicated that "the self-feedback went largely beyond most teachers' feedback capacity and bore great potential for learning and instruction" (p. 803) (cf., Kissling & O'Donnell, 2015)	**B2, C3**

2016	Tigchelaar, M. (2016). The impact of peer review on writing development in French as a foreign language. *Journal of Response to Writing, 2*(2), 6–36. https://scholarsarchive.byu.edu/journalrw/vol2/iss2/2	**Tigchelaar (2016)**, in her semester-long intervention study, compared the effect of peer reviews and self-reviews on intermediate-level college students' writing (French-learning students, *N* = 55). As far as holistic writing scores were concerned, none of the groups (including a control group) significantly improved over time. However, the type, length, and criticalness of the comments differed between the peer and self-reviews, and uptakes differed as well (cf., ALFALLAY, 2004; MATSUNO, 2009; PATRI, 2002).	**B2, C3**
2016	Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition*, *19*(1), 122–140. https://doi.org/10.1017/S1366728914000832	Following HEILENMAN (1990), LAPPIN-FORTIN & RYE (2014), SUZUKI (2015), and ÜNALDI (2016), **Trofimovich et al. (2016)** examined college L2 speakers' (*N* = 134) bias in self-assessing the accentedness and comprehensibility of their speech. The study confirmed the Dunning-Kruger effect irrespective of the learners' language background and perceived task difficulty. A subset of the original participants (56 students) showed that the discrepancies were associated with phonological dimensions (segmental and suprasegmental dimensions) but not with dimensions concerning lexicon, grammar, and discourse.	**A3, A4, C3**
2016	Ünaldi, I. (2016). Self- and teacher assessment as predictors of proficiency levels of Turkish EFL learners. *Assessment and Evaluation in Higher Education, 41*(1), 67–80. https://doi.org/10.1080/02602938.2014.980223	This case study conducted by **Ünaldi (2016)** examined the accuracy of SA (using items from the DIALANG project) among Turkish learners of English at college (*N* = 239) (cf., BRANTMEIER ET AL., 2012; LUOMA & TARNANEN, 2003 for studies using DIALANG). SA was reasonably highly correlated with both teachers' judgments and objective test scores. Consistent with HEILENMAN (1990), LAPPIN-FORTIN & RYE (2014), SUZUKI (2015), and others, the study also found that students with higher proficiency tended to underestimate their abilities whereas students with lower proficiency tended to overestimate their abilities (i.e., the Dunning-Kruger effect).	**A3, C3**
2017	Tigchelaar, M., Bowles, R. P., Winke, P., & Gass, S. (2017). Assessing the validity of ACTFL can-do statements for spoken proficiency: A Rasch analysis. *Foreign Language Annals, 50*(3), 584–600. https://doi.org/10.1111/flan.12286	**Tigchelaar et al. (2017)** examined the construct validity of NCSSFL-ACTFL can-do statement using college Spanish learners (*N* = 382) (cf., BROWN ET AL., 2014; KISSLING & O'DONNELL, 2015; MALABONGA ET AL., 2005 for ACTFL validation studies). A Rasch analysis identified 15 misfitting items (out of 50 items). The authors attributed the misfitted results to: (a) vague descriptions; (b) examples that did not match the students' learning experience; and (c) single items that assessed multiple skills. The authors highlighted the importance of item development that is suited for the target learner population.	**A3, C3**
2017	Wang, W. (2017). Using rubrics in student self-assessment: Student perceptions in the English as a foreign language writing context. *Assessment & Evaluation in Higher Education, 42*(8), 1280–1292. https://doi.org/10.1080/02602938.2016.1261993	**Wang (2017)** explored Chinese college students' views of the use of rubrics in SA in writing. Eighty students' reflective journals and interviews with six select students were analyzed qualitatively. The results indicate that the rubric was a useful guide for facilitating self-regulated learning. The data also identified some factors affecting the effectiveness of the rubrics, including categories covered and the structural formats (e.g., analytic vs. holistic); the wording of the rubric and score range; learners' domain knowledge; and the duration of the intervention (cf., BABAII ET AL., 2016; BUTLER & LEE, 2010).	**B2, C3**

(*Continued*)

(*Continued*)

Year	References	Annotations	Theme
2018	Butler, Y. G. (2018a). Young learners' processes and rationales for responding to self-assessment items: Cases for generic can-do and five-point Likert-type formats. In J. Davis et al. (Eds.), *Useful assessment and evaluation in language education* (pp. 21–39). Georgetown University Press.	Building on Moritz (1996), **Butler (2018a)** is one of the few studies looking into students' process of responding to SA, among young learners in Japan in this case. As part of a larger study, Butler compared English-learning young learners' ($N = 31$) processes and rationales when they respond to two different SA formats: dichotomous (can-do statements) and five-point Likert-type formats. Retrospective interview data along with students' SA responses were analyzed based on Higgins et al.'s (1986)[2] stages of SA processing. The results show that both individual factors (e.g., aspiration and self-efficacy) and social-environmental factors influenced the learners' SA responses, with complex relationships among these factors. Unique age effects were also found.	**A2, A3, A4, C1**
2018	Butler, Y. G. (2018b). The role of context in young learners' processes for responding to self-assessment items. *Modern Language Journal*, *102*(1), 242–261. https://doi.org/10.1111/modl.12459	Similar to Butler (2018a), **Butler (2018b)** also examined young learners' process of SA (also see Moritz, 1996). In response to previous findings that contextualization of SA administration is important (Edele et al., 2015), this paper compared young learners' SA responses ($N = 31$, English-learning children in Japan) in two implementation conditions: an after-task condition (task-specific items were used) and a generic condition (decontextualized items were used). The study found that students depended on a variety of relevant incidents and reference points in the generic condition, whereas in the after-task condition students focused on the task-at-hand and used the task requirements as the reference point. **Butler (2018b)** also found that older children (ages 10–12) tended to underestimate their performance compared with younger children (ages 8–9), consistent with previous studies conducted in other subject domains such as math (Andrade, 2019[3]).	**A2, A3, A4, C1**
2019	Hung, Y. (2019). Bridging assessment and achievement: Repeated practice of self-assessment in college English classes in Taiwan. *Assessment & Evaluation in Higher Education, 44*(8), 1191–1208. https://doi.org/10.1080/02602938.2019.1584783	Subscribing to social cognitive theory, **Hung (2019)** conducted an intervention study to see if the repeated use of SA can effectively influence college students' speaking abilities in Taiwan ($N = 97$). SA items in a guided format (as opposed to a rating format) were administered five times throughout the semester. Actual speaking performance and accuracy of self-assessment of speaking performance both improved over time, while self-assessing oral vocabulary and grammar remained somewhat difficult. An open-ended questionnaire and interviews with the students also showed that they had positive perceptions of the effectiveness of using SA repeatedly for improving their speaking abilities (cf., Butler & Lee, 2010).	**B2, C3**

[2]Higgins, E. T., Strauman, T., & Klein, R. (1986). Standards and the processes of self-evaluation: Multiple effects from multiple stages. In R. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (pp. 23–59). The Guilford Press.

[3]Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education, 4*, article 87. https://doi.org/10.3389/feduc.2019.00087

2019	Liu, H., & Brantmeier, C. (2019). 'I know English': Self-assessment of foreign language reading and writing abilities among young Chinese learners of English. *System, 80*, 60–72. https://doi.org/10.1016/j.system.2018.10.013	**Liu & Brantmeier (2019)** examined the accuracy of SA responses in relation to objective measures in reading and writing among English-learning secondary school students (N = 106, ages 12–14) in China. The students' SA was significantly correlated with both reading and writing tests, although the correlation for writing was lower, supporting Ross (1998). Given these findings, the authors suggested the potential value of incorporating SA into the L2 learning curriculum.	**A3, C2**
2019	Ma, W., & Winke, P. (2019). Self-assessment: How reliable is it in assessing oral proficiency over time? *Foreign Language Annals, 52*(1), 66–86. https://doi.org/10.1111/flan.12379	Following previous studies concerning the use of ACTFL can-do statements (cf. Brown et al., 2014; Kissling & O'Donnell, 2015; Malabonga et al., 2005, Tigchelaar, 2019; Tigchelaar et al., 2017), **Ma & Winke (2019)** investigated whether college students can reliably use the can-do statements to evaluate their learning gains over time (two-year period) in comparison to their OPI scores (learners of Chinese, N = 80). Overall, students tended to underestimate their performance, but students at the Novice and Advanced proficiency levels were more accurate than their Intermediate counterparts. No difference in the accuracy of SA was found between the first and the second year.	**A3, C3**
2019	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	**Summers et al. (2019)** examined the validity and reliability of the ACTFL can-do statement (see Tigchelaar, 2019; Tigchelaar et al., 2017). Ninety-two adult learners in an intensive English language program responded to select ACTFL can-do statements and took a placement test. The result indicated that: (a) the actual item difficulties of the can-do statements were aligned well with the intended difficulties specified in the ACTFL proficiency guidelines and that (b) the SA reliably could discriminate between students. However, given that the SA only moderately correlated with the placement test, the authors noted that users should be cautious about using SA as a replacement for a placement test.	**A3, C3**
2019	Sweet, G., Mack, S., & Olivero-Agney, A. (2019). Where am I? Where am I going, and how do I get there? Increasing learner agency through large-scale self assessment in language learning. In P. Winke & S. Gass (Eds.), *Foreign language proficiency in higher education* (pp. 175–195). Springer.	**Sweet et al. (2019)** described how Basic Outcomes Student Self Assessment (BOSSA) was developed for college students learning a foreign language and examined its efficacy using large-scale data (more than 10,000 students in ten languages). The mixed methods study found that BOSSA helped the students increase their self-awareness (of what they can do with the target language), learner agency, and engagement; in other words, BOSSA facilitated learner-centered teaching and learning. The correlations between the students' SA and ACTFL ratings were all significant but varied across skill domains and semesters. (cf., Brown et al., 2014; Kissling & O'Donnell, 2015; Ma & Winke, 2019; Malabonga et al., 2005, Tigchelaar, 2019; Tigchelaar et al., 2017)	**A3, A5, B2**
2019	Tigchelaar, M. (2019). Exploring the relationship between self-assessment 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.	**Tigchelaar (2019)** analyzed the accuracy of college students' (N = 216) self-assessed spoken abilities in relation to their ACTFL scores received on their Oral Proficiency Interview (OPI) (cf., Brown et al., 2014; Kissling & O'Donnell, 2015; Malabonga et al., 2005, Tigchelaar et al., 2017 for validation studies of ACTFL can-do statements). **Tigchelaar (2019)** showed moderate to strong validity of SA in oral abilities, although the results varied depending on the types of numeric scales employed (e.g., ordinal scale from 1 to 9, nonequal-interval scales, etc.) and the statistical analytical techniques used for the analysis.	**A3, C3**

(*Continued*)

(*Continued*)

Year	References	Annotations	Theme
2021	Lenkaitis, C. A. (2021). Virtual exchanges for intercultural communication development: Using can-do statements for ICC self-assessment. *Journal of International and Intercultural Communication, 14*(3), 258–274. https://doi.org/10.1080/17513057.2020.1784983	ACTFL can-do statements cover proficiencies not only in communication but also in intercultural communication (ICC). **Lenkaitis (2021)** applied ACTFL can-do for ICC for the first time to evaluate the effectiveness of a six-week virtual exchange with college students in other countries ($N = 106$). The can-do ratings increased after the intervention in seven of ten categories in ACTFL can-do for ICC (cf., Brown et al., 2014; Kissling & O'Donnell, 2015; Ma & Winke, 2019; Malabonga et al., 2005, Sweet et al., 2019; Tigchelaar, 2019; Tigchelaar et al., 2017).	**A5, C3**
2021	Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing, 38*(2), 189–218. https://doi.org/10.1177/0265532220932481	In one of the most comprehensive reviews since Ross (1998), **Li & Zhang (2021)** conducted a meta-analysis of 67 studies concerning the correlation between students' SA of L2 learning and their language performance. The analysis revealed a significant but moderate correlation (.466), suggesting that SA has potential but has room for improvement with respect to validity. Significant moderating effects included clear and detailed criteria, formats (computer-adaptive formats were better), training, the number of SA items, and reliability of SA.	**A3, A6, C4**
2021	Hosseini, M., & Nimehchisalem, V. (2021). Self-assessment in English language teaching and learning in the current decade (2010-2020): A systematic review. *Open Journal of Modern Linguistics, 11*(6), 854–872. https://doi.org/10.4236/ojml.2021.116066	In another recent review of SA, **Hosseini & Nimehchisalem (2021)** took a qualitative approach as opposed to Li & Zhang's (2020) quantitative meta-analysis above. The review also focused on the most recently published studies (2010–2020) on SA in English language teaching (ELT), not L2 in general. The paper discusses critical issues concerning the use of SA both as a measurement tool and as a learning tool. The authors proposed a view of SA that combines and highlights both its measurement and learning aspects.	**A6, B4, C4**
2021	Mäkipää, T. (2021). Students' and teachers' perceptions of self-assessment and teacher feedback in foreign language teaching in general upper secondary education – A case study in Finland. *Cogent Education, 8*(1), 1978622. https://doi.org/10.1080/2331186X.2021.1978622	In this case study, **Mäkipää (2021)** qualitatively examined Finnish secondary school students' and their teachers' perceptions of the use of SA and teachers' feedback in foreign language courses. Interviews with nine students (varied in the target languages) and ten teachers revealed that how SAs were used in class varied substantially across the teachers. The data revealed discrepancies in the views of students and teachers, with students reporting that they often received insufficient guidance on how to use SA, whereas the teachers believed that they had explained the criteria to the students (cf., Jang et al., 2015).	**A2, B2, C2**
2021	Wind, A. M. (2021). Co-development of self-assessment and second language writing from a complex dynamic systems theory perspective: A single case study. In G. Tankó & K. Csizér (Eds.), *DEAL 2012: Current explorations in English Applied Linguistics* (pp. 229–262). Eötvös Loránd University.	Based on complex dynamic systems theory, this single case study examined an adult student's accuracy of SA and the linguistic complexity of her writing. Employing a time-series analysis, **Wind (2021)** found that the student's accuracy of SA generally improved over time but not in a linear fashion. While SA positively correlated with cohesion, it negatively correlated with vocabulary and grammar (cf., Luoma & Tarnanen, 2003; Matsuno, 2009; Ross, 1998).	**A3, C3**