Annual Review of Applied Linguistics (2020), **40**, 113–118
doi:10.1017/S0267190520000021

CAMBRIDGE UNIVERSITY PRESS

SHORT PAPER

Reflect, Revisit, Reimagine: Language Assessment in *ARAL*

Carol A. Chapelle*

Iowa State University
*Corresponding author. E-mail: carolc@iastate.edu

Abstract

This 40th anniversary of *ARAL* also marks the 40-year anniversary of a significant uptick in research on language assessment, and hence there is much to reflect on and revisit within this period—and still scope for imagining the future. Pre-1980, language assessment had a long history, but Spolsky (1995) designated the late 1940s as a time of professionalization, which continued through the following decades. By the 1970s, language testers were gradually organizing into an academic community with an annual international conference, regional conferences, journals, and scholarly books. The new academic community not only developed and used language tests but also investigated the validity of their interpretations and uses. Canale's (1987) paper in *ARAL* provides an enduring frame of reference for reflection on the concerns of the academic community, which he introduced as the *what*, *how*, and *why* of language assessment.

Four Decades Revisited

The first issue of the *ARAL* in 1980 contained one article on language testing titled, "Language Testing Research (1979–1980)." It reflected contemporary discussion in the field, treating language proficiency as a trait that influences performance across a variety of measures (Oller, 1980). Others had challenged the hypothesis that the language construct should be theorized as unidimensional and invariant across different measures; psychometric statistical tools were being marshaled to serve in the quest to better understand language proficiency through the analysis of test performance, and connections between such research and second language acquisition were being made (Bachman, 1988).

Before the late 1970s, language testing books had provided guidance about methods for language testing, but the work of Oller and of Bachman and Palmer expanded the discussion of testing methods. Oller (1979) had proposed testing methods such as cloze and dictation would elicit evidence of the language proficiency trait, whereas Bachman and Palmer (1982) found that the ability elicited by tests of different methods could be shown to measure at least somewhat different constructs. From Oller's perspective, when "measurement methods have large amounts of non-random variance associated with them which are not also associated with the traits that the tester desires to measure, an undesirable situation occurs" (Oller, 1980, p. 131). Test methods have come to be

recognized to be of utmost importance and should be designed, not to attempt to eliminate their effects, but to plan for the desired effects (Douglas, 2000).

In the 1990s, the multiple dimensions of language assessment evident today began to bloom. Douglas' (1995) "Developments in Language Testing" described the feeling that language testing had come of age with a coherent professional community that shared discussion and debate about a theory, methods, data analysis, and even professional standards. *ARAL* welcomed articles that helped to shape the discussion in language assessment while inviting others in applied linguistics to join in. In addition to Douglas' article summarizing developments in the area, a second one at the end of the decade again took stock of development in language testing in the 1990s (Kunnan, 1999).

ARAL's articles in the 1990s reflected the developments within each of the *what*, *how*, and *why* of language testing. An article in the 1990 issue picked up the discussion about the construct of language proficiency, reviewing the theory and research that helped to shed light on discourse competence as one component (Shohamy, 1990). In the 1998 issue titled, "Foundations of Second Language Teaching," one article was devoted to assessing each of the four language skills of listening, reading, speaking, and writing. The division of language proficiency into four skills reflected the editor's interest in compiling knowledge about language assessment that would be useful for language teaching at that time. Each of the four articles included not only the construct but also the assessment methods, because of the connection between *what* is tested and *how* the assessment is carried out. The topic of test methods, however, transcended individual skills and was therefore developed in articles about performance assessment, alternative assessment, and technology use in language assessment. Picking up the *why* of language assessment, McNamara's 1998 *ARAL* article focused on policy and social considerations in language testing for the first time. In contrast to the 1980s, when psychometric methods from psychology were used with little comment, a 1999 article presented a different perspective on validation (Chapelle, 1999).

The final issue of the first decade of the 2000s marked the high point for emphasis on language assessment in *ARAL* with a special issue dedicated to the role of language assessment as an instrument of language policy. In the introduction to the issue, Spolsky (2009) explained the enduring connection between testing and privilege in society. He sketched the central controversy created by tests intended to provide equal access to privilege by setting clear criteria for advancement and maintaining quality standards in education and workplace settings. The same tools designed to maintain equity are also viewed as instruments that reinforce the privilege of members of society whose background and experience give them an advantage (Spolsky, 2009).

The policy lens on language assessment in the special issue provided an opportunity to expand on Canale's (1987) introduction of the *why* of language assessment and McNamara's 1998 article on policy and social aspects of language assessment by covering a broad range of language test uses and their value implications. Speaking to the central controversy, the articles explain the implications and issues associated with test use in domains including education, immigration, and workplace decision-making as well as in locations including the United States, Iran, Wales, and Scotland. The policy perspective also spawned attention in the profession to language assessment literacy among all test users, because others, not testers, are frequently responsible for decisions about test use.

Other articles throughout the early 2000s continued to build on the range of issues reviewed through the previous decade, including the application of discourse

https://doi.org/10.1017/S0267190520000021 Published online by Cambridge University Press

perspectives on language to assessment and the use of technology in language assessment. The connection between test constructs and social issues was developed in an article on "Assessing English as a Lingua Franca" (Elder & Davies, 2006). In short, the articles in the first decade of the 2000s lay the groundwork for the issues to be picked up in language assessment in the future.

Any follower of the language assessment threads through the first three decades of *ARAL* has to wonder what happened to them in the 2010s. Nearly mirroring the slender treatment of assessment in the 1980s, the 2010s issues contain only four or five articles whose focus is clearly marked as assessment. Did language assessment diminish in its recognized importance? Did it split off from applied linguistics to create a separate discipline? Or did it become so embedded in other applied linguistics research and practice as to become nearly invisible?

It would be difficult to argue that language assessment is no longer important, but some language testers might say that it is a separate discipline from applied linguistics. Language assessment today has a multifaceted identity spanning theoretical, methodological, and practical issues throughout government, education, and business domain within and across national borders. Nevertheless, one tentacle of its reach remains within applied linguistics, and some of the articles in the 2010s hint at this fact by situating assessment within other issues of applied linguistics.

Several of the articles in the 2010s build upon the groundwork of the previous decades and step toward imagining the future of language assessment in applied linguistics. Several articles examine testing methods in view of how they are used to meet goals in applied settings. These include one on uses for task-based language assessment (Norris, 2016) and one on potential learner impacts from the use of CEFR-based self-assessments, peer assessments, and teacher assessments in education in Europe (Little & Erikson, 2015). In addition, three articles of interest appear in the 2017 issue on child second language acquisition (Philp, Borowczyk, & Mackey, 2017). Each of these articles is able to demonstrate the strengths and limitations of assessment methods for providing the specific performance data required to serve a particular use (Bailey, 2017; Fortune, & Ju, 2017; Foster & Wigglesworth, 2016). All demonstrate the significance of test use for evaluating assessments. The thread from 1999 about validation research, which has gained momentum throughout the past decade, was picked up in "Mixed Methods Research in Language Testing and Assessment" (Jang, Wagner, & Park, 2014) and promises to be central going forward.

The Future Imagined

Each of the three areas identified by Canale—the *what*, *how*, and *why* of language assessment—will remain central to a future reimagined area of language assessment. Each is being affected by the use of evolving technologies and the many forces of globalization. In part because of the expansion of language assessment, two other wh-questions, *who* and *where,* need to be added as core questions to emphasize the social and political dimensions of language assessment, and the knowledge of the people involved.

With respect to the *what* of language testing, constructs are the basis for score interpretation, and, therefore, they need to be specified in a manner that serves in test development and validation research. Throughout previous *ARAL* articles, language test constructs have remained an enduring theme as underlying assumptions have evolved about how language constructs should be defined, their connection to assessment

https://doi.org/10.1017/S0267190520000021 Published online by Cambridge University Press

methods, and the appropriate means for their justification. Readers have seen language constructs discussed in relation to teaching skills, classroom needs, research objectives, technology influences, and policy impacts, to name a few perspectives. Throughout these articles, it is evident that construct interpretations are anything but given entities to be discovered; rather they are, as the term "construct" suggests, created by test developers and researchers for particular purposes. A constructivist view is important, particularly as language use is so often mediated by technology in ways that necessitate reanalysis of assumptions about language constructs. Going forward, language testers will need to continue to define, evaluate, and revise language constructs in technically and politically aware terms.

The *how* of language assessment refers to test methods, or the procedures for eliciting samples of performance used to make inferences about test-takers' abilities and future performance. The articles in the past decade forecast a productive view of test methods through their analysis of methods for particular uses in contrast to past syntheses of methods organized around skills (e.g., listening or speaking). Test-taking method as the organizing principle for synthesis allows testers to examine the effectiveness of certain characteristics of test methods to advance language assessment and contribute to other areas of applied linguistics. A good goal for the next decade would be to cull existing knowledge about commonly used configurations of test methods to offer the field a summary of their strengths and weaknesses for particular uses. This project would lay groundwork sorely needed for exploring the new test method characteristics made available through the use of technology.

The *why* of language assessment refers to the justification of language test use within particular contexts for certain purposes, taking into account their value implications and social consequences. In the 1980s, Canale distinguished the ethical justifications for testing from the validation of a test as a measure of a particular construct. Today and going forward, both the construct validity justifications and those associated with the value implications and social consequences of testing are integrated within a framework for validation. Such a framework, referred to as argument-based validity (Chapelle, 2021; Kane, 2013) encompasses claims to be made about the construct that the test is intended to measure in addition to the intended social consequences of test use. Leaders in language assessment have provided a strong foundation for taking into account the social dimensions in evaluating the validity of test use (Shohamy, 2001; Bachman & Palmer, 2010; McNamara & Roever, 2006). Recent books by leaders in the field further advance philosophical discussions of the values inherent in validation (Fulcher, 2015; Kunnan, 2018). With this groundwork laid, the future of validity investigations should integrate the technical with the value implications and social consequences of language testing.

The expanding uses of language assessments and complexity of options they present mean that *who* is creating, using, and interpreting tests is critical. For example, online tools make it possible for a variety of people to access data about learners' and test-takers' performance from which they draw inferences. In other words, when students use technology for their routine academic work, they inadvertently produce data that are gathered by researchers, teachers, and testers wanting samples of performance. The use of such data to make inferences about students' abilities and future performance is assessment. The 2019 *ARAL* article "Recent Contributions of Data Mining to Language Learning Research" (Warschauer, Yim, Lee, & Zheng, 2019) illustrates the need for users of such data to be aware of the principles of assessment they are working with when they make such interpretations. In short, the importance of

https://doi.org/10.1017/S0267190520000021 Published online by Cambridge University Press

assessment and expanded access to performance data mean that all applied linguists need to have an understanding of language assessment. In this context, language assessment literacy across the expanding pool of language-assessment users is an area of current and future importance (Kremmel & Harding, 2020), as forecast by the 2009 *ARAL* article "Developing Assessment Literacy" (Taylor, 2009).

The *where* of language assessment refers to the impact of place on assessment use within a global environment, where media, travel, migration, and technology bring speakers of different languages into contact. Situations of language contact create opportunities and needs for language assessment in learning, certification, and gatekeeping. The ability of language users is assessed, either implicitly or explicitly, at almost every point of contact as they carry on conversations, write emails, apply for immigration and citizenship, seek acceptance in higher education, complete graduation requirements, and try for employment or advancement in their work. Language assessments are also used throughout the process of language learning in a magnified level of importance in online and blended learning, in which technology can be used to provide more regular assessment and precise individualized feedback than what a teacher can manage for a large class. Expanded sites and formats for assessment call for renewed consideration of security as it pertains to the validity of, and different cultural perspectives on, testing. Moreover, the expanded reach of aspiring test developers suggests the need for prospective test-users throughout the world to have sufficient knowledge to judge claims about new tests.

Conclusion

Overall, ARAL articles have afforded a window on language testing and assessment as well as its integral connection to other areas of applied linguistics. They have captured the essence of the preoccupations in the 1980s, the boom of the 1990s, and the social turn of the 2000s. In the 2010s, they began to hint at some of the issues going forward. The considerable activity on the horizon for language assessment in the future may prove to be a challenge for *ARAL* to keep up with. However, the new format introduced in 2016, including review articles that provide state-of the-art overviews of the field, and other types of articles, including position pieces and empirical papers will be a good home for summarizing, arguing, and showcasing what is new in language assessment.

References

Bachman, L. (1988). Language testing: SLA research interfaces. *Annual Review of Applied Linguistics*, *9*, 193–209.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, *16*(4), 449–465.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Bailey, A. L. (2017). Progressions of a new language: Characterizing explanation development for assessment with young language learners. *Annual Review of Applied Linguistics*, *37*, 241–263.

Canale, M. (1987). The measurement of communicative competence. *Annual Review of Applied Linguistics*, *8*, 67–84.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272.

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. Sage Publishing.

Douglas, D. (1995). Developments in language testing. *Annual Review of Applied Linguistics*, *15*, 166–187.

Douglas, D. (2000). *Assessing Language for Specific Purposes*. Cambridge University Press.

https://doi.org/10.1017/S0267190520000021 Published online by Cambridge University Press

Elder, C., & Davies, A. (2006). Assessing English as a lingua franca. *Annual Review of Applied Linguistics*, *26*, 282–304.
Fortune, T., & Ju, Z. (2017). Assessing and exploring the oral proficiency of young Mandarin immersion learners. *Annual Review of Applied Linguistics*, *37*, 264–287.
Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*, 98–116.
Fulcher, G. (2015). *Re-examining language testing*. Routledge.
Jang, E., Wagner, M., & Park, G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics*, *34*, 123–153.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.
Kremmel, B., & Harding, L. (2020) Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. *Language Assessment Quarterly*, *17*(1), 100–120.
Kunnan, A. (1999). Recent developments in language testing. *Annual Review of Applied Linguistics*, *19*, 235–253.
Kunnan, A. J. (2018). *Evaluating language assessments*. Routledge.
Little, D., & Erickson, G. (2015). Learner identity, learner agency, and the assessment of language proficiency: Some reflections prompted by the common European framework of reference for languages. *Annual Review of Applied Linguistics*, *35*, 120–139.
McNamara, T. (1998) Policy and social considerations in language assessment. *Annual Review of Applied Linguistics*, *18*, 304–319.
McNamara, T. & Roever, C. (2006). *Language testing: the social dimension*. Blackwell Publishing.
Norris, J. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics*, *36*, 230–244.
Oller, J. (1979). *Language tests at school*. Longman.
Oller, J. (1980). Language testing research (1979–1980). *Annual Review of Applied Linguistics*, *1*, 124–150.
Philp, J., Borowczyk, M., & Mackey, A. (2017). Exploring the uniqueness of child second language acquisition (SLA): Learning, teaching, assessment, and practice. *Annual Review of Applied Linguistics*, *37*, 1–13.
Shohamy, E. (1990). Discourse Analysis in Language Testing. *Annual Review of Applied Linguistics*, *11*, 115–131. doi:10.1017/S0267190500001999
Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language if language tests*. Pearson Education.
Spolsky, B. (1995). *Measured words*. Oxford University Press.
Spolsky, B. (2009). Editor's introduction. *Annual Review of Applied Linguistics*, *29*, vii–xii.
Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, *29*, 21–36.
Warschauer, M., Yim, S., Lee, H., Zheng, B. (2019). Recent Contributions of Data Mining to Language Learning Research. *Annual Review of Applied Linguistics*, *39*, 93–112.

Cite this article: Chapelle, C. A. (2020). Reflect, Revisit, Reimagine: Language Assessment in *ARAL*. *Annual Review of Applied Linguistics*, *40*, 113–118. https://doi.org/10.1017/S0267190520000021