Lang. Teach. (2009), **42:4**, 491–503 © Cambridge University Press
doi:10.1017/S0261444808005570 First published online 12 December 2008

The ecology of assessment

Carl Falsgraf Center for Applied Second Language Studies,
University of Oregon, USA
falsgraf@uoregon.edu

Assessment has long been beset by a dichotomy between what Lynch (2003: 4) termed the positivist and interpretivist orientations. The former views assessment as a technical exercise involving the quantification of learners' knowledge or cognitive abilities; the latter considers assessment as the humanistic endeavor of portraying learners' qualitative development and subjective experiences. This presentation argues that, although these distinctions may be useful heuristics, differing types of assessments are more profitably viewed as filling particular niches within an ecosystem. The speech also reports on technical developments that place control over assessment results in the hands of learners, which allows them to assemble evidence from a variety of sources to substantiate proficiency claims. These developments should help promote the paradigm of assessment as ecology and help the field move beyond the uneasy dichotomy that now predominates.

1. The ecology of assessment

The need for context in interpreting information about student performance will be a key theme throughout this presentation. In the interest of providing context for interpreting the information here, I would like to provide some background on my interest, training, and biases related to language assessment.

After a career teaching English and Japanese as foreign languages, I received my academic training in linguistics in the American functionalist tradition with a focus on SLA. Instead of following a traditional academic path, however, I became what I half-seriously wrote on the 'occupation' line of forms as 'educational engineer'. In this capacity, I developed curricula, professional development programs, and assessments to support foreign language teachers in the U.S.

My bias, therefore, sees assessment as an integral part of the language learning enterprise, inextricably entwined with teacher improvement, curriculum, standards, and methodology. This orientation is somewhat at odds with the high-stakes testing traditions in the U.S. and much of Asia, which share a belief that human ability can be quantified and that major decisions about test takers' lives can be made on the basis of this result. In the American tradition, the predominant testing paradigm has been historically associated with

Revised version of an invited talk presented on 18 October 2007 to the series Assessing Language Learning, University of Wisconsin–Madison.

https://doi.org/10.1017/S0261444808005570 Published online by Cambridge University Press

psychometrics and intelligence testing, though politicians and the general public seem much more enthusiastic about the enterprise than psychometricians themselves, who are aware of its limitations. The American public often views schools as factories whose output is educated students, as measured by standardized tests. Although neither the American psychometric tradition nor the Asian Confucian tradition have taken deep root in Europe, our discussion here may still be germane to the general discourse regarding the relationships among various assessment types or perhaps serve as a source of satisfaction that other parts of the world are beginning to see things the European way.

Lynch (2003) labels the psychometrically oriented American tradition the 'positivist' approach and a more holistic and contextualized orientation the 'interpretivist' approach. The American positivist orientation describes an array of methods, attitudes, and assumptions deriving from the core belief that psychological states and processes can be measured through sampling and statistical analyses. The purpose of assessment is to accurately quantify these states and processes. The interpretivist orientation, also with its own subculture and assumptions, derives from the core belief that complex psychological processes can only be understood in the socio-cultural context and that the purpose of assessment is the improvement of student learning. The often unproductive debate between the more extreme elements of these two camps is about false choices: 'Do you want to make a choice about a person's future on the basis of an empirically validated test, or will just anybody's opinion do?' ask the positivists. 'Do you trust the judgment of a teacher who knows the whole child or a faceless statistician in a large corporation?' counter the interpretivists.

The debate is not just academic. Interest groups such as Fair Test (http://www.fairtest.org) lobby against large-scale, high-stakes testing. Financial interests such as Pearson (http://pearsonassess.com) and Educational Testing Service (http://ets.org) influence the political process to promote the use of their commercial products. Educational policy and many millions of dollars hang in the balance. Of greater interest to this audience, I presume, is the fate of the language teaching enterprise. I will argue that we do not have to make the false choice debated by positivists and interpretivists nor do we even have to compromise. Rather, by employing the power in information technology to gather, process, and display data, we can both capture the richness of language growth in context and provide reliable data on student performance to demonstrate to ourselves, our students, and those who pay our salaries that we are indeed accomplishing our mission.

After a brief description of the dominant paradigm in American public education, focused on high-stakes psychometric testing, accountability, and empirical evidence, I will examine the interpretivist critique of this paradigm. I will describe the ecology of assessment in the real world and end with a proposal, illustrated by examples, of how the power of information technology can be used to gather, process, and display data, allowing us to move beyond the false choices of the past.

1.1 The American positivist assumption

The roots of psychometrically based educational testing lie in the philosophy, technology, and politics of the early 20th century. Philosophically, positivism rests on the belief that, given adequate information, all natural phenomena can be described objectively and accurately. Technically, testing relies on the statistical techniques developed to measure intelligence.

https://doi.org/10.1017/S0261444808005570 Published online by Cambridge University Press

Politically, it satisfies the requirements of a government demanding better educated troops for World War I, of factories needing better educated workers, and of a public expecting top value for its tax dollar (Giordano 2005). The positivist assumption holds that one can infer an individual's overall ability in something (i.e., language proficiency) from a limited sample of behaviors in a controlled situation (i.e., a test) and that this is essential to certifying successful individual learning and improving education. Then, as now, the positivist assumption is associated with national security, commercial interests, and conservative politics and is generally opposed by teacher unions, liberals, and social reformers (Giordano 2005).

Psychometricians, defined here broadly as those engaged in developing and verifying tests, strongly adhere to the positivist assumption. They generally see the 'assessment problem' as resulting from insufficient rigor in the development of tests and from the inappropriate use of tests: '[u]nfortunately, many tests are written, administered, and interpreted by people who lack adequate training' (Allen & Yen 1979: 1).

Positivists focus on the mathematically measurable trait of reliability rather than the relatively subjective notion of validity. The superior reliability of psychometrically verified tests argues for their use to make high stakes decisions:

> When we make 'high stakes' decisions . . . we had better be careful that the scores upon which we base our decisions are sufficiently reliable that our resulting decisions are reasonable. (Thompson 2002: ix)

> A language test without validation research is like a police force without a court system: Unfair and dangerous. (McNamara 2007: 280)

Psychometricians focus on critiquing existing techniques and improving the reliability of testing instruments. As long as a society chooses to use tests for making important decisions such as college admissions and school funding, we are indebted to them for giving us the best information possible given the current state of understanding about cognition and statistical modeling. A critical piece of any psychometrician's training is understanding the limits of inferences that can be drawn from a sample of data. The political exigency of putting a number on student learning, however, often leads to results being interpreted 'by people who lack adequate training' (Allen & Yen 1979: 3).

I have personal experience on both sides of this epistemological quagmire. As the developer of the Standards-based Measurement of Proficiency (STAMP; http://casls.uoregon.edu/stamp.php and http://avantassessment.com/products/about_stamp.html), I was horrified to receive a call from a teacher whose administrators threatened her with dismissal because her students did poorly on the test. As the director of a resource center responding to the needs of schools around the country, however, I often find myself pushing our test developers to create a test faster or cheaper, knowing full well that doing so will limit the reliability of the information produced but not limit the conclusions schools will draw from that information. Anyone developing a test in the real world – with limited budgets, unavoidable error, and a public desiring pat answers – faces a similar dilemma: How good does a test have to be? When is a bad test better than no test at all?

1.2 The interpretivist critique

The interpretivist tradition takes as its basic assumption that learning is complex, individualized, and dynamic and that the purpose of assessment is to improve teaching

https://doi.org/10.1017/S0261444808005570 Published online by Cambridge University Press

and learning. Contextualized performance assessments that provide meaningful feedback to students and teachers are, therefore, the appropriate instrument for documenting learning.

Interpretivists reject the psychometric program on two counts: validity and educational efficacy. The validity argument states that psychometric standardized testing measures ability, a supposedly fixed attribute. Learning and performance, however, are not fixed attributes, therefore undermining one of the core assumptions underlying the validity claims of standardized tests. Because psychometric measurement requires standard procedures, testing represents a distinct context either from the classroom, where learning occurs, or the 'real world', where learning may be applied and demonstrated (Gipps 1994).

Arguing from a Vygotskyan perspective, Poehner & Lantolf (2005) distinguish between 'static assessment' and 'dynamic assessment'. Static assessment is characterized by the positivist assumption that a relatively stable knowledge state exists and can be inferred through techniques that elicit and analyze evidence of that knowledge. Dynamic assessment, on the other hand, accounts for the results of an intervention, providing information on latent abilities that represent learning in progress. Applying the principles of dynamic assessment to language testing, Poehner & Lantolf (2005) argue for an interactive process between learner and expert that elicits latent linguistic abilities that would be missed in a traditional testing situation.

The second critique of the positivist paradigm holds that standardized testing does not improve education, while formative performance assessments do. Bergan et al. (1991) found that kindergartners engaged in a formative measurement and planning system scored significantly higher in reading, math, and science measures. An examination of Israeli eleven-year-olds showed that intrinsic motivation increased when learners received regular, meaningful feedback (Butler 1988). Portuguese students engaged in daily reflective self-assessment performed better in mathematics (Fontana & Fernandes 1994).

The effect of standardized testing on learning is limited and often detrimental. Concerns about the distortion of the curriculum to accommodate high-stakes testing are longstanding. Gipps (1994), for example, reports on the educational effects of standardized testing on the British education system in the early 1990s. Watanabe (2004) reports on the negative washback of standardized English tests on English language education in Japan. Furthermore, washback, positive or negative, seems to be limited to changes in curriculum content, not teaching methodology (Stecher, Chun & Barron 2004). The case for high-stakes, summative testing alone as a means of raising standards, therefore, is not supported by research, while contextualized formative assessment has been shown to improve learning.

1.3 The ecology of assessment revisited

This presentation has purposely characterized the positivist and interpretivist positions in somewhat extreme terms in order to clearly explain each position. In reality, few test developers are without qualms about the issues of validity and educational efficacy raised above, and few teachers see no role whatsoever for external testing. The eminently reasonable compromise is to clearly state the intended uses of any assessment method, be clear about the limitations of that method, and use assessments only for the purposes for which they were developed.

https://doi.org/10.1017/S0261444808005570 Published online by Cambridge University Press

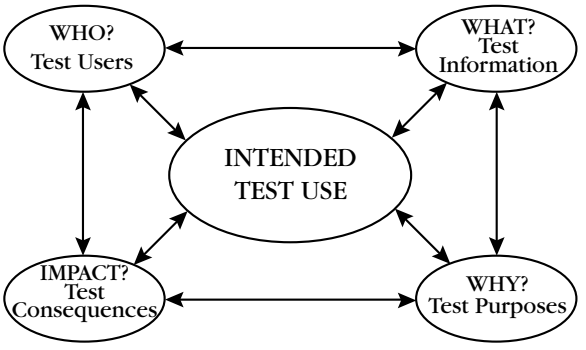


Figure 1 Components of intended use (from Norris 2000).

John Norris (2000) argues for moving from a focus on testing to a focus on assessment, leading to a holistic view that recognizes the appropriate use and interpretation of an array of assessment methods that range from achievement tests to alternative assessments. Each instrument should be accompanied by a statement of intended test use that takes into account the 'who', 'what', 'why', and 'so what/impact' of assessment, as illustrated by Norris' (2000) diagram in Figure 1.

I call this orientation 'the ecology of assessment' and have titled this paper accordingly. This view represents, I believe, a more profitable approach than the mutually exclusive assumptions of the extreme positivist and interpretivist positions. An ecosystem is not a compromise, but rather mutually dependent relationships in constant flux, adapting to changes in the environment. The testing environment includes the needs of the students, curricular changes, and community demands for accountability. Under the ecosystem model, zero-sum questions such as 'Which is better, the Standards-based Measurement of Proficiency (STAMP) or the Advanced Placement (AP) test?' are transformed into more profitable formulations such as 'How can we use STAMP to give us information that will help students eventually pass the AP test?' In this ecosystem, formative assessment provides meaningful feedback to students, which has been shown to improve motivation and learning.

However, this ecosystem is far from the Garden of Eden. In fact, a healthy testing ecosystem is likely to exist in only a few isolated places for the following reasons:

1) Assessment literacy. Teachers generally do not receive training in assessment practices, giving them insufficient knowledge and background to make rational decisions about choosing appropriate instruments and interpretations.
2) In-class assessment. Lack of training also leads to questionable in-class assessment. Classroom assessment generally encourages superficial and rote learning. Teachers do not trust each other's judgments, and learning is subordinate to grading (Black & Wiliam 1998).
3) Rarity of formative assessment. Classroom assessment is overwhelmingly summative, though occasionally dressed in formative clothing. Even when feedback is given, that feedback is often summative in nature (e.g., 'Great job! You got a 95.').

https://doi.org/10.1017/S0261444808005570 Published online by Cambridge University Press

4) Accountability. Political pressure for accountability ensures that external, high-stakes tests will be required to document school and student accountability.

On closer examination, the assessment ecosystem looks like a desert. Changing teachers' belief systems to acknowledge assessment as a crucial component of learning, not just evaluation, is the key to a long-term solution. Assessment OF learning must transform into assessment FOR learning. Such a shift is unlikely to occur until new teachers have experienced a litany of graded in-class tests punctuated by the occasional state-mandated external test. Furthermore, teachers' knowledge regarding assessment practices must be significantly upgraded, likely requiring changes to licensure requirements. In other words, this desert has no rain in the immediate forecast.

2. The power of information technology

Information technology can gather, process, and display information better and faster than human beings. Although we may be a long way off from the day when teachers have a belief system and knowledge base conducive to a healthy assessment ecosystem, the tools for providing students with useful feedback through technology is at hand. This speech concludes optimistically with a description of two tools with promise to significantly improve the quality and accessibility of performance data available to students and teachers.

2.1 The Standards-based Measurement of Proficiency

The Standards-based Measurement of Proficiency (STAMP) is an online, psychometrically validated proficiency test available in Chinese, French, German, Hebrew, Japanese, and Spanish. Arabic, Hindi, Korean, Persian, Swahili, Urdu, and Yoruba are currently under development. Based on the American Council on the Teaching of Foreign Languages (ACTFL) Performance Guidelines (http://actfl.org), STAMP is a summative assessment with a twist: it provides timely and detailed feedback to teachers and students.

STAMP is delivered entirely online, requires no downloads or special software, and works on almost any computer configuration. The reading and listening sections of STAMP are in a multi-stage semi-adaptive format, so items are somewhat tailored to students' proficiency level, resulting in a summative judgment related to the ACTFL scale. Items are based on authentic or semi-authentic texts in a selected-response format and validated using Rasch analysis. Speaking and writing sections consist of open-ended prompts to which students respond either orally or in writing. At least two trained graders rate these texts according to scoring guides visible to students, as shown in Figure 2.

For the purposes of this paper, the important innovation of STAMP is the reporting structure. Because the test is delivered online, scores for reading and listening are accessible immediately to the student and teacher, and scores for speaking and writing are available within two weeks. General scores are reported, as depicted in Figure 3, with explanations of the criteria the learner has met.

https://doi.org/10.1017/S0261444808005570 Published online by Cambridge University Press

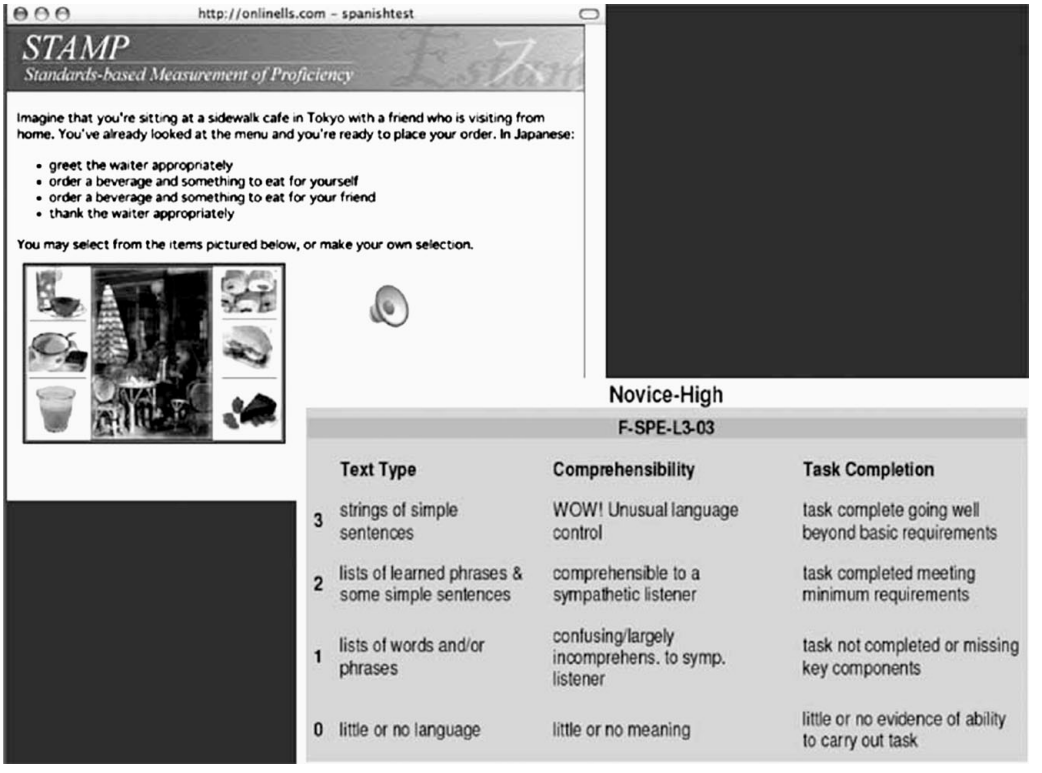


Figure 2 STAMP speaking prompt with scoring guide.

The real power of the reporting system, however, lies in the hyperlinks to student work. By clicking on one of the 'sample' hyperlinks under writing and speaking, teachers and students can view or listen to that student's productions along with the external graders' evaluations. This reporting structure provides an invaluable teaching opportunity. Because the evaluation is done externally, the teacher truly becomes a coach, helping the learner develop strategies to achieve the proficiency necessary to do better the next time. Thus, while STAMP does make a summative judgment of general proficiency in the manner of positivist, psychometrically validated standardized tests, it also provides students and teachers with the tools to engage in a formative discussion of strengths, weaknesses, and goals.

Figure 4 shows a writing prompt and a student's response from an actual report page, illustrating how teachers and students have access to the testing context, the student's response, and the external raters' evaluation. The report structure facilitates a detailed diagnosis of strengths and weaknesses as well as advice on how to improve.

STAMP occupies a niche in the American assessment ecosystem not unlike that of DIALANG in Europe. Based on the ACTFL Performance Guidelines, on which almost all state standards are also based, STAMP helps teachers and schools orient instruction towards proficiency standards. Although providing less diagnostic feedback than DIALANG, STAMP does give access to actual student productions along with the external evaluators'

https://doi.org/10.1017/S0261444808005570 Published online by Cambridge University Press

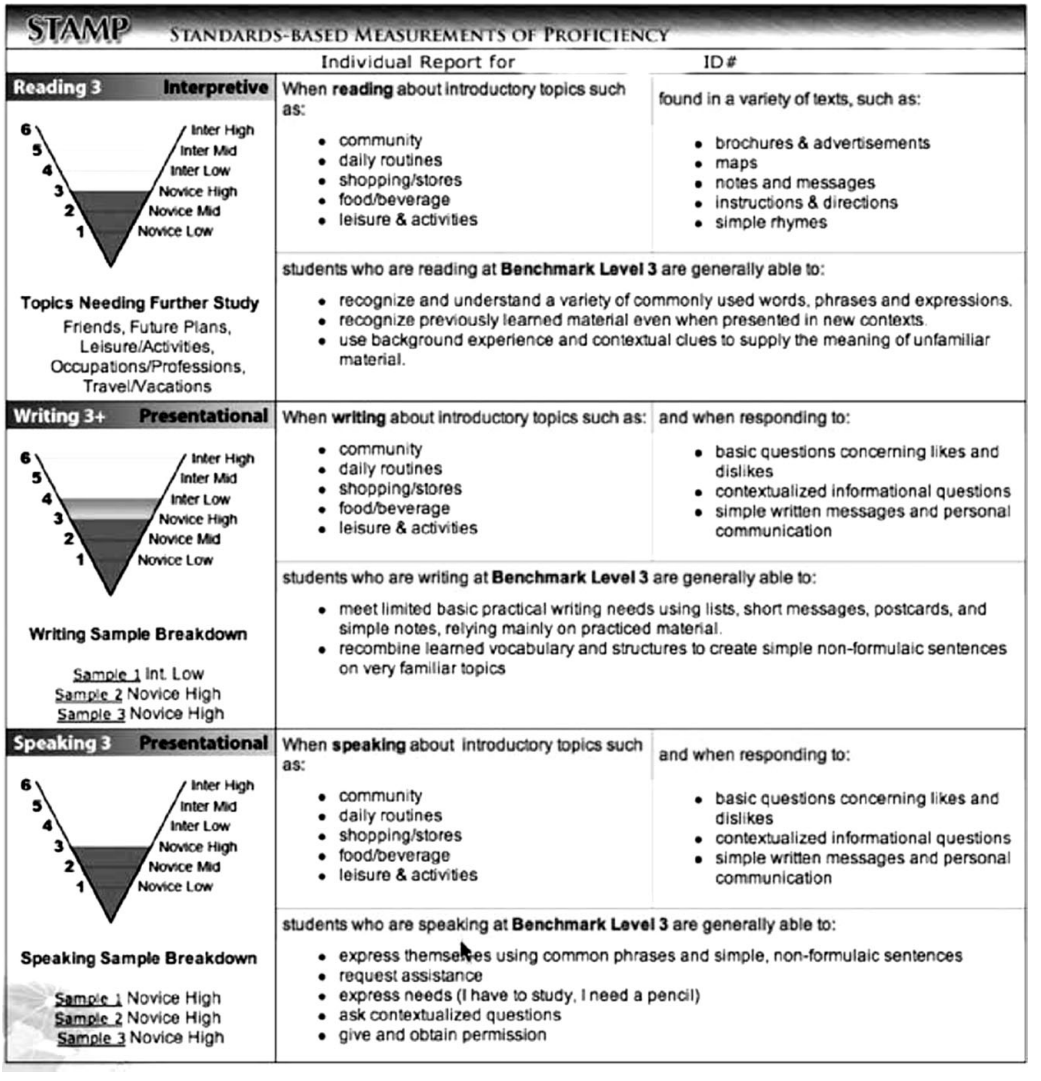


Figure 3 STAMP reports page.

judgments following the scoring guide. The prototypical use of STAMP, therefore, is for program evaluation and teacher-mediated formative feedback, as opposed to DIALANG's automated feedback to individual learners primarily on receptive skills.

2.2 LinguaFolio Online

STAMP may have a niche in the assessment ecosystem as a one-time testing event, yet its utility is limited to providing a once-a-year snapshot of student performance that may or may not accurately reflect learners' 'real' ability, and it certainly cannot document progress

https://doi.org/10.1017/S0261444808005570 Published online by Cambridge University Press

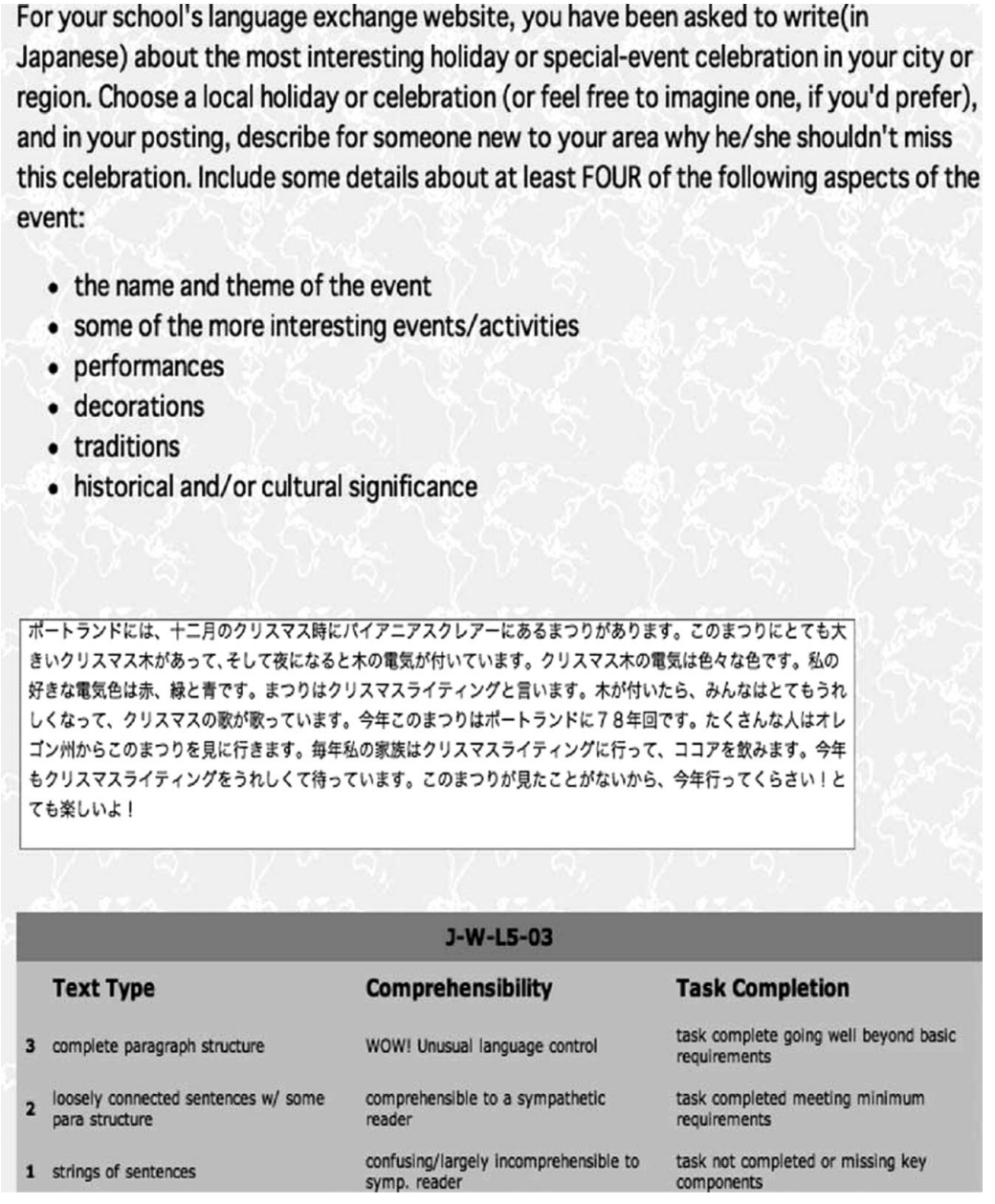


Figure 4 Japanese writing sample (accessible through report page).

over time. LinguaFolio Online, however, represents a fundamentally different approach to assessment by placing the burden of proficiency proof on the student, not on an external test developer or even the classroom teacher, although both have important roles in the LinguaFolio Online paradigm.

https://doi.org/10.1017/S0261444808005570 Published online by Cambridge University Press

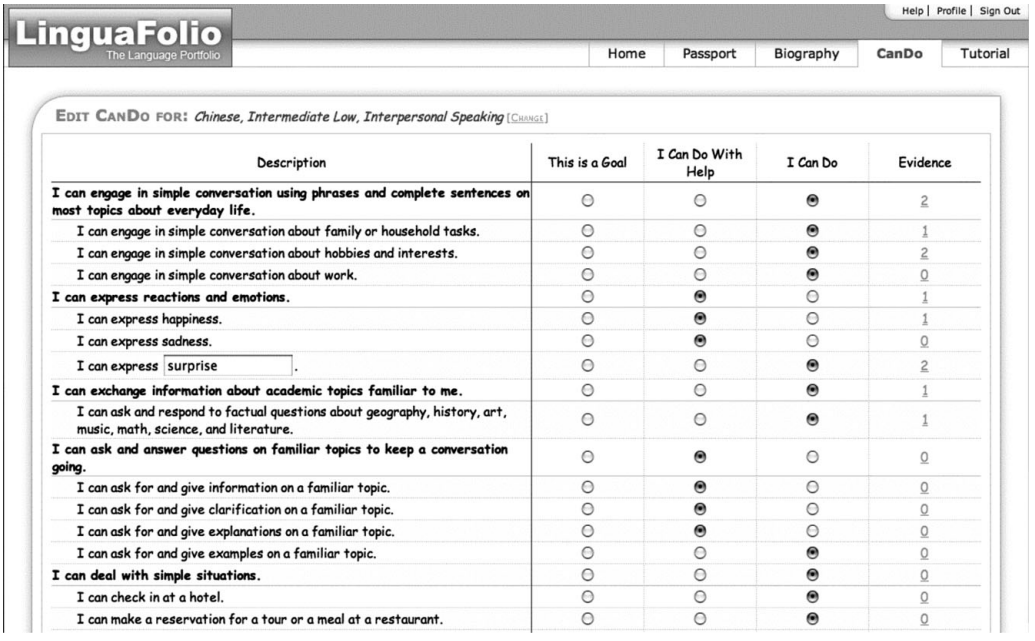


Figure 5 LinguaFolio Online CanDo checklist.

Based on the European Language Portfolio (http://www.coe.int/portfolio), LinguaFolio was developed by the National Council of State Supervisors for Languages (NCSSFL), a professional organization of chief foreign language administrators in each state (http://ncssfl.org). After five years of piloting the paper-and-pencil version, NCSSFL collaborated with the Center for Applied Second Language Studies (CASLS) and the National Foreign Language Center to develop an online version of the instrument, called LinguaFolio Online.

Learners can create a LinguaFolio Online account at any time, free of charge. Students are then responsible for evaluating their own abilities according to a set of CanDo statements derived from the Common European Framework of Reference for Languages (Council of Europe 2001) and tied to the ACTFL/ILR guidelines widely used throughout the United States. As with the European Language Portfolio, learners self-assess according to various CanDo statements and submit evidence to support those claims. The online version allows learners to attach evidence tied to specific statements, greatly enhancing ease of evaluation. A learner claiming 'I can write a simple thank you note', for example, could upload an email or scanned image of a handwritten thank you note. Any external consumer (e.g., employer, university) could evaluate the veracity of the claim for themselves by clicking the 'evidence' button next to the CanDo statement. Figure 5 shows the CanDo checklist where this information is submitted and accessed. Students can also document and display external test scores (e.g., ACTFL OPI, STAMP, Advanced Placement tests) and certificates of completion, providing external verification of overall proficiency. Finally, a Biography section prompts learners to reflect on their language learning history, their

https://doi.org/10.1017/S0261444808005570 Published online by Cambridge University Press

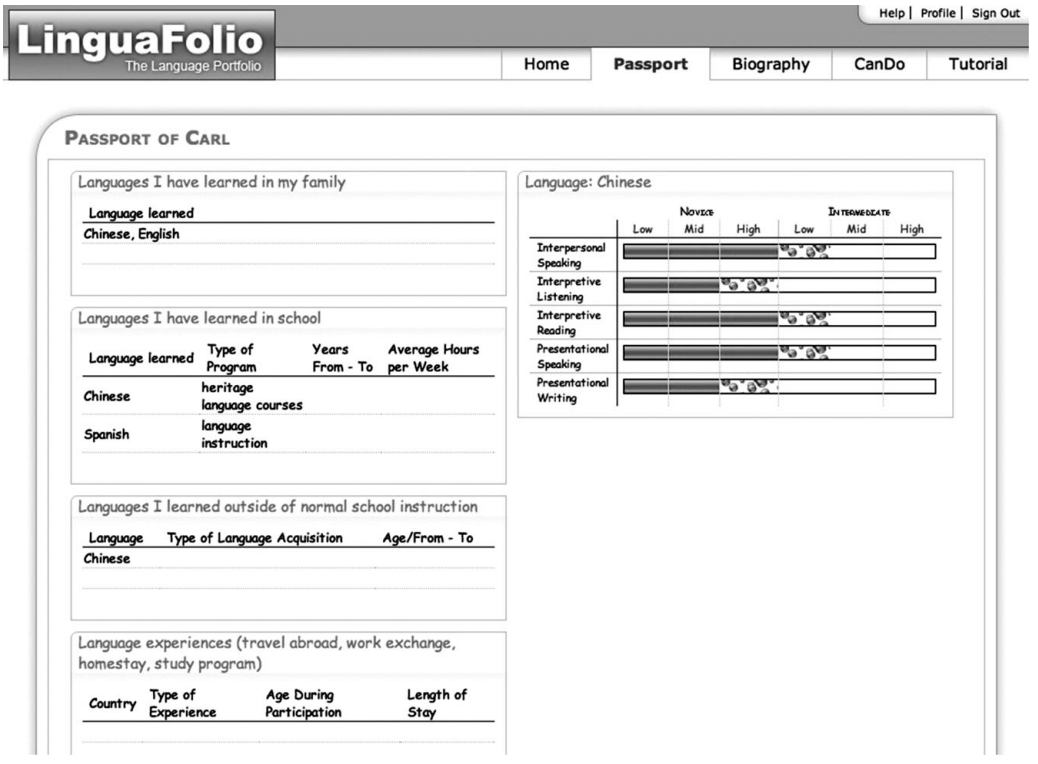


Figure 6 LinguaFolio Online passport.

growing interculturality, motivations, and language learning goals. A sample from the Passport page, which summarizes student reports, is shown in Figure 6. The solid line represents full proficiency at a level, operationalized as all CanDo statements being checked. The bubbles represent a level at which the student is currently working towards proficiency.

Based on a set of standards (CanDo statements), the LinguaFolio Online system represents significant improvements over traditional portfolios. Students own their LinguaFolio Online accounts, and, because it is online, the system is completely portable. Highly mobile Americans can keep a record of their language learning experiences as they change locales and move from primary to secondary to tertiary education. LinguaFolio Online also documents learning that occurs outside the formal education system.

The student bears the burden of proof, but the evaluator (e.g., employer, college admissions officer, teacher) bears the burden of interpretation. Different purposes require different types of evidence. Certifying that an airline pilot possesses sufficient English skills to communicate with air traffic control clearly requires more compelling evidence than determining whether to place a student in an intermediate or advanced language class. Each evaluator must determine the level of proof required for the decision at hand.

Most important to this discussion, LinguaFolio Online represents a departure from the positivist/interpretivist dichotomy and offers a third way of looking at assessment. The need

https://doi.org/10.1017/S0261444808005570 Published online by Cambridge University Press

for valid summative judgments is addressed by allowing the posting of standardized test scores. Within the same tool, learners create a rich, qualitative portrait of their motivations, goals, and progress. Learners, educators, and the public need not make a false choice between rigorous, verifiable, external, summative judgments and contextualized, longitudinal, qualitative, formative information.

2.3 Transcending old paradigms through technology

Internet and networking technologies have democratized and decentralized control over information. People commonly book their own flights, find their own homes, and conduct their own research online rather than relying exclusively on travel agencies, realtors, or libraries. Similarly, tools like STAMP and LinguaFolio Online can decentralize control over information regarding student performance. In the past, this control rested primarily with testing companies and teachers. Now, students will be able to assemble evidence of their abilities themselves and share that evidence with anyone they would like.

Just as travel agents and libraries have adjusted their practices to accommodate this new reality, language teachers and test developers will also need to redefine their roles in assessment. Teachers have traditionally served as referees – enforcing rules, making judgments, handing out the occasional red card. As students gain control over their own evaluation, teachers will become more like advocates to help learners assemble evidence that best supports their case for proficiency. Certainly, some will feel uncomfortable with this transfer, just as travel agents and realtors once predicted chaos when consumers received access to once privileged information. Will some students cheat? Will some employers make bad decisions? Almost certainly. As the system matures, however, the social sanctions against cheating and the obvious consequences of choosing an unqualified employee will place checks and balances on the worst abuses.

3. Conclusion

The technical means for collecting, organizing, and displaying diverse types of evidence alone will not change language teaching for the better or even improve the measurement of student performance. These outcomes depend on the development of assessment literacy on the part of teachers, students, and the general public. The washback effect of having collections of evidence available is hard to predict. Will people continue to focus on the raw summative score, or will ready access to actual student performances lead to increased awareness of the complexity of language development? Will businesses concentrate on only a narrow band of performance indicators ('I can negotiate', 'I can write a business letter', etc.), or will human resource managers become familiar with the general proficiency scale? The answers most likely depend more on professional development and public awareness efforts than they do on assessment tools themselves. Nonetheless, the existence of tools that puts assessment information in the hands of learners themselves is an important first step in placing value on language proficiency and in helping the wider society interpret and understand that value.

https://doi.org/10.1017/S0261444808005570 Published online by Cambridge University Press

References

Allen, M. J. & W. M. Yen (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing.
Bergan, J. R., I. E. Sladeczek, R. D. Schwarz & A. N. Smith (1991). Effects of a measurement and planning system on kindergartners' cognitive development and educational programming. *American Educational Research Journal* 28, 683–714.
Black, P. & D. Wiliam (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice* 5.1, 7–32.
Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology* 58, 1–14.
Cheng, L., Y. Watanabe & A. Curtis (eds.) (2004). *Washback in language testing: Research contexts and methods*. Hillsdale, NJ: Lawrence Erlbaum.
Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
Fontana, D. & M. Fernandes (1994). Improvements in mathematics performance as a consequence of self-assessment in Portuguese primary school pupils. *British Journal of Educational Psychology* 64, 407–417.
Giordano, G. (2005). *How testing came to dominate American schools: The history of educational assessment*. New York: Peter Lang.
Gipps, C. (1994). Developments in educational assessment: What makes a good test? *Assessment in Education: Principles, Policy & Practice* 1.3, 283–296.
Lynch, B. K. (2003). *Language assessment and programme evaluation*. Edinburgh: Edinburgh University Press.
McNamara, T. F. (2007). Language assessment in foreign language education: The struggle over constructs. *Modern Language Journal* 91.2, 280–282.
Norris, J. M. (2000). Purposeful language assessment: Selecting the right alternative test. *English Teaching Forum* 39.1. http://exchanges.state.gov/forum/vols/vol38/no1/p18.htm.
Poehner, M. E. & J. P. Lantolf (2005). Dynamic assessment in the language classroom. *Language Teaching Research* 9, 1–33.
Stecher, B., T. Chun & S. Barron (2004).The effects of assessment-driven reform on the teaching of writing in Washington State. In Cheng et al. (eds.), 53–72.
Thompson, B. (2002). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.
Watanabe, Y. (2004). Teacher factors mediating washback. In Cheng et al. (eds.), 129–146.

CARL FALSGRAF is director of the Center for Applied Second Language Studies (CASLS) and the Oregon Chinese K-12 and Undergraduate Flagship. He holds a doctorate in linguistics and has been teaching and conducting research on language education for over twenty years. His classroom experience includes teaching ESL, Japanese, and a variety of graduate courses in pedagogy and methodology. He served on the American Council on the Teaching of Foreign Languages Executive Board and as President for the Pacific Northwest Council for Languages. His interests include proficiency assessment, standards-based approaches to language education, and data-driven professional development.

https://doi.org/10.1017/S0261444808005570 Published online by Cambridge University Press