International Journal of Language Studies
Volume 9, Number 4, October 2015, pp. 133-143

Language testing: The state of the art
(An online interview with James Dean Brown)

James Dean BROWN, University of Hawai'i at Mānoa, USA
M. A. SALMANI NODOUSHAN, Iranian Institute for Encyclopedia Research, Iran

In this interview, JD Brown reflects on language testing/assessment. He suggests that language testing can be seen as a continuum with hard core positivist approaches at one end and post modernist interpretive perspectives at the other, and also argues that norm referencing (be it proficiency, placement, or aptitude testing) and criterion referencing (be it diagnostics, progress, or achievement testing) fall on this continuum. He further suggests that evaluation is done at the level of program or course but that assessment is focused on the classroom, and then argues that both assessment and evaluation exploit measurement and testing albeit to different effects. He then comments on his views about high-stakes and low-stakes testing as well as washback, and finally expresses serious concerns about the impacts of language policy on language testing by calling the current NS models into question. Relating his concerns to validity issues, he suggests that language testers need to consider other options to the NS model to serve the needs of speakers of other Englishes.

Keywords: Portfolios; Assessment; Evaluation; Criterion Referencing; Norm Referencing; Dynamic Assessment

The Interview

MASN:[1] How would you define language testing/assessment today?

JDB:[2] I would define language testing/assessment as a subfield of applied linguistics in which a variety of different kinds of people work on the issues of testing and assessment from perspectives ranging widely from norm-referenced (proficiency, placement, and aptitude testing, including complex statistical analyses and theories of

[1] Mohammad Ali Salmani Nodoushan
[2] James Dean Brown

ISSN: 2157-4898; EISSN: 2157-4901

validity) to criterion-referenced (diagnostics, progress, and achievement testing, including classroom and curriculum notions of task-based assessments, portfolios, conferences, self- and peer-assessments, as well as continuous, differential, and dynamic assessments). In short, what unites language testers is our interest in applying testing and assessment to languages. But what divides us are our specializations within language testing which range from hard core positivist and advanced statistical orientations to postmodernist interpretive perspectives, and everything in between.

MASN: What is the *sine qua non* of language testing in the 21st century? Reliability? Practicality? Validity? All of them? What else?

JDB: The focus in theory development during the past couple of decades has been on validity. However, the focus in the real world of people having to actually develop and use language tests is now (and always has been) on practicality with a dash of reliability (and sometimes validity) thrown in for good measure.

MASN: What kind of interface do you see between language testing, language teaching methodology, linguistics, educational psychology, ESP, and education?

JDB: From my point of view, language testing has always drawn on linguistics, teaching methodology, educational psychology, ESP, education, statistics, psychometrics, and other fields for ideas and techniques to serve our ultimate purposes, which are to serve the testing and assessment needs of language teaching and learning.

MASN: How would you define 'evaluation', 'assessment', 'measurement', and 'testing'? How would you relate them? What is the nature of the interface (if any) that you see between them?

JDB: These terms cause much confusion in our field because they are used in different ways by different "experts." To solve this problem for myself and my students, I have tried to use them consistently in the following ways. To me:

Evaluation is a word that I never use alone. Instead, I only use it in the phrases *program* (or *course*) *evaluation*, which I define as the processes of determining the value and ways to improve the curriculum of a particular language course or program.

Measurement, to me, includes all forms of testing and assessment, but also questionnaires, observations of various kinds, and any other ways we quantify, code, or describe the behaviors of language students.

Testing focuses on the summative or formative direct or indirect observation of the language behaviors of language students for feedback (whether numerical or verbal) and decision making purposes.

Assessment includes all sorts of testing and other forms of measurement but focuses on the processes and purposes of determining the language performance, progress, and achievement of individual students in language teaching and learning situations, most often to promote learning or for grading purposes.

Notice how very careful I have been in defining these concepts. They are really slippery. Perhaps an easier way to think of these terms is the way my father (who was a sailor during World War II) taught me to think about the difference between a ship and a boat. He said, "A boat fits on a ship, but a ship doesn't fit on a boat." By analogy, *program evaluation* and *assessment* are two different types of ships that serve two different purposes at the program and classroom levels, respectively. On each of those ships, you can fit the boats of measurement and testing, but in different ways.

Using these terms consistently the way I do does not make my definitions the only ones. However, I hope that, one way or another, I have made it clear how I think about these terms. Other people see them differently, but of course, they are wrong (JD laughs here).

MASN: How would you distinguish high-stakes from low-stakes testing?

JDB: To me, *high-stakes decisions* are relatively important ones that have serious implications and consequences in terms of the decisions we are making about people's lives. While the tests involved tend to be administered on a single occasion, they usually include many items to help insure reliability. For example, the iBT TOEFL tends to be high-stakes because a student's scores on this test can determine whether they will go to university (or at least whether they will attend a top-notch university); since it is administered on a single occasion, the test designers include many items to make this high-stakes decision as reliable as possible.

In contrast, *low-stakes decisions* do not have such grave

implications and consequences, and yet we instinctively try to make them as reliable in the aggregate as possible by using many such low-stakes quizzes and tests over a longer period of time. For example, a teacher will administer numerous relatively low-stakes quizzes and tests before using the aggregated (and more reliable) information for assessing and grading each student's progress and achievement over an entire term.

MASN: What is your definition of washback? Why is it important?

JDB: To me, *washback* is the effect (positive or negative) that testing can have on the language teaching and learning associated with it. Washback is important because it can serve as a carrot or stick in shaping the attitudes of students and motivating them to learn or acquire the language. For example, when I was teaching in China way back in 1980-1982, the students initially did not like our communicative language teaching, especially pair and group work, preferring instead the grammar-translation-memorization teaching that had worked in China for "thousands of years." We were able to turn their attitudes around by, among other things, testing them in interviews, pairs, groups, and other productive language use formats that created positive washback.

At the same time, most of our Chinese students knew that they were going to have to take what was at that time exclusively a multiple-choice TOEFL before they would be able to go to an English speaking country for post graduate work. As a result, they wanted to memorize thousands of multiple-choice items from TOEFL preparation books. Because we believed that "if you learn English, your TOEFL score will go up" and that memorizing thousands of multiple-choice items was a terrible way to learn English (and a colossal waste of time), we felt that the TOEFL was having a negative washback effect on our students and program.

MASN: What is the importance of codes of ethics in language testing? What ethical issues do you find vital in language testing?

JDB: As far as I am concerned, ethics are important in everything we do. However, to me, *codes of ethics* are only important insofar as they make us reflect on the details of ethical behavior in certain domains, like language testing, and lead us to discovering new ways of thinking about that behavior. At the heart of all ethical behavior is what we call the golden rule: *treat others as you would have them*

treat you. Unfortunately, because human beings sometimes have difficulty empathizing and seeing things from another person's point of view, some people find this simple rule difficult to apply. I suppose that is why codes of ethics have been formulated to help such people act ethically.

MASN: What are the implications of language testing (as you see it) for social policy?

JDB: There are many language testing issues that have social policy implications, but one of the most important from my perspective is the role of the native speaker. As it stands today, most language tests, whether NRTs for proficiency or placement testing or CRTs for diagnostic or achievement testing, are based on the native speaker (NS) model, that is, the idea that the English that non-native speakers (NNS) should be studying is that of the NS of English. Thus, for listening comprehension, NNSs are presented with NSs giving lectures or conversing in pairs, and the NNSs must show that they can comprehend NSs. Or for speaking, the English of NNSs is judged for pronunciation, syntactic accuracy, fluency, or whatever, in terms of how well it approximates the NS model of English. Indeed, the very notion of validity for English tests is typically grounded in this NS model. And, that has serious implications for social policy. For example, the NS model implies: that Inner-circle British and/or American Englishes are better than the Outer-circle Englishes of say India, Singapore, or the Philippines; that NNSs are somehow broken and need to be fixed by turning them into NSs; and that, by extension, NNSs are somehow inferior to NSs (notice that *non*-native speakers are *non*-, implying that they lack something, and what they lack of course is nativeness).

As a result of such attitudes, (a) many people value NS teachers over NNS teachers, even if the NS teachers have no teacher training or experience while the NNS teachers are well-trained and very experienced, which of course, makes zero sense, and (b) students around the world are set up to fail (by the impossible dream of becoming a NS) because most of them will never become (nor do they need to become) anything close to native. The reason for this is that most NNSs simply do not have the time and practice necessary to become even native-like. After all, NSs develop their native abilities over many decades with *tens of thousands of hours of very heavy input* in English, while most students of EFL, or even ESL,

study a few hours per week for 3-6 years getting at best *hundreds of hours of weakly reinforced input*. Do you see how all of that has implications for social policy?

Never mind that we have no idea what the *NS* is in the *NS* model. The various World Englishes communities have argued for years that English is not one thing. As a result, it is very difficult to define what a native speaker is. Is a NS anyone who only speaks English? Or is the status of NS restricted to educated people who only speak English? And, which English are we talking about? British? American? Australian? Singaporean? And, even if we decide on say British English, which of the *many* dialects in the UK are we talking about? Equally important, who should serve as a model of the sainted status of NS? Are we talking about the Queen of England here? Or JD Brown? Or George W. Bush?

Nonetheless, we go ahead and build tests of *overall English language proficiency* (whatever that is) based on something we cannot define. We do this by presenting NNSs with comprehension or production problems that require them to manipulate NS English. We pilot such items and then only keep those items that spread the NNSs out along a continuum. So in reality, the NS model in proficiency testing is based on the sorts-of-stuff-that-discriminates-among-NNSs-in-terms-of-how-well-they-work-with-NS-language (whatever that is). Do you see the problem?

We also need to ask ourselves who these tests based on the so-called NS model are appropriate for? If we *could* define the NS model, which we can't, I suppose it could be argued that such tests might be appropriate for some NNSs—in particular, those who are in fact going to study in or immigrate to an English speaking country. However, the vast majority of students of English around the world will *not* study at a university at home, much less abroad[3], nor immigrate to an English speaking country. For the majority of students, then, teaching and testing based on a NS model is a great disservice. We know that, in the real world, most such learners are much more likely to communicate in English with other non-native speakers (NNSs) of English (Japanese with Koreans, Farsi speakers with Malay speakers, etc.) than they are to communicate with NSs.

[3] According to *OECD* (2013, p. 15), on average across the many countries they surveyed, less than 40 percent of people over 25 years old have completed tertiary diplomas or degrees of any kind anywhere.

Wouldn't such learners be better served by being taught some relatively manageable and learnable form of English like English as an international language (EIL) or English as a lingua franca (ELF) based on their actual local needs to communicate in English?

In short, because of the NS model approach to language testing, such people are being tested using a model of communication that they will never be able to achieve and most likely will never need to use. How's that for social policy?

MASN: What are your predictions for the future directions of language testing?

JDB: Language testing in the past several decades has rightly focused and continues to focus on validity issues in all their richness. But, the fatal flaw in all this is that language tests are largely designed to test the NS model of proficiency. If we show that these tests are valid for testing the NS model, even if we do it extremely well, are we adequately showing that their scores are valid—for whom and for what purposes? Or are we completely forgetting Messick's (1996) notions of *implications* and *consequences*—especially in terms of reasonable expectations for what our students can learn and their actual uses of English in the real world?

I believe that there are many alternative approaches we could be using for testing English language proficiency—approaches that could usefully replace the NS model because they are achievable and reflect the real-world needs of most students. For example, we might judge learners' receptive reading and listening abilities in terms of their ability to comprehend the English of other NNSs, and there productive writing and speaking skills in terms of intelligibility. In Brown (forthcoming), I argue that there are at least 14 approaches to testing English proficiency, six of which are top-down approaches that focus on variations within the language for different sorts of English (including the traditional NS model approach, as well as what I call the truth-in-advertising, multiple world Englishes, English as a lingua franca, global standard English, and functional approaches) and eight of which are bottom-up approaches that focus on how persons vary in their English in the real world (including what I call the effective communicator, scope of proficiency, scale of range, intelligibility, resourcefulness, symbolic competence, intercultural communication skills, and performative ability approaches) (for more on this, see Brown, forthcoming)

My point is that there are other ways of looking at English language proficiency and that they are quite different from the NS model that has been *the* model for English proficiency throughout my career. I am certainly not arguing here for one approach over the others, but I *am* pointing out that language testers should consider other options to the NS model and that some of these approaches might singly or in combination be more appropriate and effective for various groups of students and testing purposes—even for *most* groups and purposes.

I could go on and on (and I do in Brown, 2012, 2014, in press; McKay & Brown, 2015). But you asked about my predictions. I predict that language testing will necessarily face up to the fact that English proficiency can be defined in a variety of different ways that have little or nothing to do with NSs. Thus there will not be one validity for the so-called overall English language proficiency (whatever that is), but rather different very clearly defined and labelled validities for different purposes and groups of students taking the various English tests around the world.

MASN: What are the implications of your view of language testing for research and training?

JDB: In language testing, we have tended to adopt a psychometric model for what testing should be. We also tend to love our statistics. Honestly, I'm never happier than when I'm doing Rasch analysis, or using Generalizability theory, or building a structural equation model. But, given that the entire structure of English language proficiency is built on the false NS-model premise, it really doesn't matter how enamoured we are with our validation strategies or fancy statistics, does it?

MASN: Do you have any recommendations for language testing specialists?

JDB: I presume you are talking about the younger folks who are just starting out and may profit from such advice. Honestly, I only have four thoughts I would like to pass along, and they have more to do with being professional applied linguistics researchers than with language testing specifically: First, I think it is crucial to follow your own instincts and explore where your personality takes you in the field. Your interests and curiosity will help you to carve out a corner of the field that is yours and is therefore different from the work of everyone else. Second, it is very important, at least for me, to work

every single day on my research and writing, maybe only an hour or two, but every single day. That way, I do not have to rush through things, I can be thoughtful and careful, and I can progress slowly and steadily. Third, it is crucial to constantly ask questions, lots of questions, and then be sure to answer them as honestly as you can regardless of how those answers may match or mismatch your preconceptions. In fact, if something does not turn out the way you expected it to, you should pay special attention to that. If an answer you get seems anomalous, don't ignore or bury that information. Follow that anomaly up with new questions and answer those. I personally have found repeatedly that it is just such anomalies that have led to the most interesting and important things I have discovered in my research. And finally, try as best you can to enjoy your work. If you are not enjoying it, if you are not eager to sit down and work each day, maybe you should find something else to do, something you do enjoy, right?

MASN: Thank you very much for accepting this interview invitation. It means a lot to me and the readers of the journal. You are an Icon, and it was a huge honor for me to be given this opportunity to conduct this interview. Thank you.

JDB: Thank you for the invitation to do this interview. You must have put a lot of thought into your questions. In fact, I have enjoyed thinking about these issues and responding.

As for being an *icon*, I'm not sure how I feel about that. After all, to my granddaughter, I am just grandpa JD who tickles her and makes her laugh, and to my wife, I am just JD who forgets where the keys are and makes her laugh. Being called an *icon* makes me feel old, and serious, and maybe near to the end of the road. In reality, I think I still have a lot of life left in me. In fact, I find myself learning about our field and writing articles and books at a faster rate now than ever before. Please stay tuned. I think I still have more to say.

The Authors

James Dean ("JD") Brown (Email: brownj@Hawaii.edu) is Professor of Second Language Studies at the University of Hawai'i at Mānoa. He has spoken and taught in many places ranging from Brazil to Venezuela. He has published numerous articles and books on language testing, curriculum design, research methods, and connected speech. His most recent books are: *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and*

Pacific languages (2012 from NFLRC); *New ways in teaching connected speech* (2012 from TESOL); *New ways of classroom assessment, revised* (2013 from TESOL); *Practical assessment tools for college Japanese* (2013 with K. Kondo-Brown from NFLRC); *Mixed methods research for TESOL* (2014 from Edinburgh University Press); *Cambridge guide to research in language teaching and learning* (2015 with C. Coombe from Cambridge University Press); *Teaching and assessing EIL in local contexts around the world* (2015 with S. L. McKay from Routledge); *Introducing needs analysis and English for specific purposes* (in press 2016 from Routledge), and two others that are currently in the works.

Mohammad Ali Salmani Nodoushan (Email: dr.nodoushan@gmail.com) is associate professor of Applied Linguistics at the Iranian Institute for Encyclopedia Research, Tehran, Iran. His main areas of interest include politeness and pragmatics. He has published over 50 papers in international academic journals, including *Teaching and Teacher Education*, *Speech Communication*, and *TESL Canada Journal*, and has (co-)authored five academic books. He is Editor-in-Chief of the *International Journal of Language Studies*, and sits on the editorial boards the *Journal of Asia TEFL*, *Journal of Linguistic and Intercultural Education*, *Asian EFL Journal*, and *Journal on English Language Teaching*. He is also a reviewer for a number of international journals, including *Journal of Pragmatics*, *TESOL Quarterly*, *Pragmatics and Society*, *Australian Journal of Linguistics*, and *The Journal of Politeness Research*.

References

Brown, J. D. (2012). EIL curriculum development. In L. Alsagoff, S. L. McKay, G. W. Hu, & W. A. Renandya (Eds.), *Principles and Practices for Teaching English as an International Language,* (pp. 147-167). London: Routledge.

Brown, J. D. (2014). The future of world Englishes in language testing. *Language Assessment Quarterly, 11*(1), 5-26.

Brown, J. D. (forthcoming). World Englishes and international standardized English proficiency tests. In C. Nelson, B. B. Kachru, & Z. G. Proshina (Eds.), *Handbook of World Englishes* (2nd ed.). Malden, MA: Wiley-Blackwell.

McKay, S. L., & Brown, J. D. (2015). *Teaching and assessing EIL in local contexts around the World.* New York: Routledge.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*, 241-256.

OECD (2013). *Education at a glance 2013: OECD indicators.* The Organisation for Economic Co-operation and Development. Retrieved July 25, 2015 from http://www.oecd.org/edu/eag2013%20(eng)--post-B%C3%A0T%2013%2009%202013%20(eBook)-XIX.pdf