Social perspectives on language testing

Papers in honour of Tim McNamara
Carsten Roever, Gillian Wigglesworth

Foreword

Tim McNamara has had an outsized influence on Applied Linguistics, and especially on language testing. No single volume can do justice to Tim's whole body of work but this volume brings together diverse papers to honour his achievements in some of the core language testing areas he has worked in: language testing and social justice, citizenship testing, tests of spoken interaction, and language testing and lingua franca, among others.

Tim entered Applied Linguistics the way many of us do, in a roundabout way. After a first degree at the University of Melbourne, he trained as a language teacher and subsequently became a teacher trainer in the UK where he was also part of the team working on the original development of IELTS. When he returned to Melbourne to undertake his PhD, his dissertation focussed on the development of a test of English for health professionals, which later grew into the Occupational English Test, an instrument still used today to assess the language skills of a wide variety of health professionals. He taught in the Applied Linguistics program at the University of Melbourne for over 30 years, during which he co-founded (with Alan Davies) the Language Testing Research Centre, and was involved in the revamp of TOEFL, which eventually led to TOEFL iBT.

Tim was instrumental in introducing Rasch measurement to the field of language testing, and has always had a special interest in performance-based testing and interview-based language tests.

Tim's interests, however, despite his expertise in language testing, are much broader, and his name is probably most closely associated with work on the social dimension of language testing with his research in this area informed by his interest in sociolinguistics, identity and language contact. Added to this is his engagement with postmodernism and poststructuralism, which contributed greatly to his writings on language testing in the context of migration and citizenship and his broad thinking of what language tests measure, can measure, should measure, and his critical enquiry into their use.

Throughout his career, Tim taught a variety of subjects in both the linguistics undergraduate degree, and the applied linguistics postgraduate degree as well as supervising numerous PhD theses to completion. Tim is highly regarded by students at all levels, as well as staff in the School, and colleagues across the world, all of whom continue to speak fondly of him and maintain their contact with him.
Contents

List of contributors .................................................................................................................. 13

The Social Dimension of Tests

Bernard Spolsky
The slow recognition of the social impact of language testing ........................................... 17

Elana Shohamy, Alastair Pennycook
Extending fairness and justice in language tests .................................................................... 29

Kellie Frost
Language testing in immigration policy: Transitioning from fairness to social justice ........ 43

Susy Macqueen, Kerry Ryan
Test mandate discourse: Debating the role of language tests in citizenship ....................... 55

Cathie Elder, Ute Knoch, Owen Harradine
Language requirements for Australian citizenship: Insights from a senate enquiry ............. 73

Tests and Testing

Joseph Lo Bianco
Talking to the Pollies: Academic researchers and public officials ....................................... 89

Noriko Inashita
Peer interaction assessment: Overview of research and directions ...................................... 105

Lyn May
"Whose performance?" Challenges in defining and operationalising interactional competence in paired speaking tests ......................................................... 121

Ana Maria Ducasse
Raters and the social dimension of scale revision for Spanish L1 Graduate Writing Skills: the case of the Saber Pro ............................................................... 131
Angela Scarino
Teacher judgment of student writing performance in second language learning ......................................................... 143

Learner, Teacher and Stakeholder Views

Tomoyasu Akiyama
Investigations of past and present stakeholders’ values using Messick’s validity framework: A case of introducing EFL speaking tests into a high-stakes test .................................................................................. 159

Jason Fan, Yan Jin
Working towards professional standards: How do students and teachers view English testing practice in China? .................................................. 173

Kathryn Hill
Students as partners in classroom-based language assessment .................................................. 187

The Future of Language Testing

Barry O’Sullivan
Considering validity ..................................................................................................................................... 199

John Pill, Luke Harding
A most engaging scholar: Tim McNamara and the role of language testing expertise .................................................. 217

Paul Gruba
The challenge of theory: Social media and language assessment .................................................. 229

Henry Widdowson, Barbara Seidlhofer
Accounting for communication .................................................................................................................. 243

List of contributors

Tomoyasu Akiyama
Bunkyo University, Japan

Joseph Lo Bianco
The University of Melbourne

Ana Maria Ducasse
RMIT University

Cathie Elder
The University of Melbourne

Jason Fan
The University of Melbourne

Kellie Frost
The University of Melbourne

Paul Gruba
University of Melbourne

Luke Harding
Lancaster University

Owen Harradine
The University of Melbourne

Kathryn Hill
La Trobe University

Noriko Iwashita
University of Queensland

Yan Jin
Shanghai Jiao Tong University

Ute Knoch
The University of Melbourne

Susy Macqueen
Australian National University

Lyn May
Queensland University of Technology

Barry O’Sullivan
British Council

Alastair Pennycook
University of Technology, Sydney

John Pill
Lancaster University

Carsten Roever
The University of Melbourne

Kerry Ryan
Swinburne University of Technology

Angela Scarino
University of South Australia

Barbara Seidlhofer
University of Vienna

Elana Shohamy
Tel Aviv University

Bernard Spolsky
Bar-Ilan University

Henry Widdowson
University of Vienna

Gillian Wigglesworth
The University of Melbourne
Paul Gruba

The challenge of theory: Social media and language assessment

Pausing for a moment here to reflect on the work of Tim McNamara, I know that I have been fortunate to be his student, colleague and friend. Tim helped me develop a thesis exploring the role of video in listening, mentored me through a circuitous career, and most recently, counselled me through some difficult times. At the core of our relationship is mutual respect that, importantly, comes a joy in sharing new ideas, critiques and observations about life. Over the years, my own work has concerned the role of technology within educational, assessment and social uses of language. As I move across fields, Tim’s work on the challenge of theory (McNamara, 2015) often helps me anticipate nagging questions in my research: Where is the basis for your study? What are the implications for what you are doing? Where is this leading? Such questions, and the challenges that they presage, continue to percolate through recent studies to do with technology, language and society (Bruhn, 2015).

To spark discussions with Tim about the work that I now do, I can imagine making a bold, even outrageous, claim: Language testing, as we know it, will cease to exist within a generation. Transdisciplinary elements of computational linguistics, cloud computing and artificial intelligence can be brought together to produce comprehensive language use profiles; for years, individuals who are ‘always on’ their devices have produced a stream of data over their lifetimes of interaction in study, work and play across a global network of connected devices (Statista, 2018a). The result of their collective interactions has resulted in Big Data, or the aggregation of global computer-based information, and it will form the linchpin of future analyses of the ways in which individuals interact in a variety of contexts, times and modalities (Klous & Wieland, 2016). In this developing scenario, I believe, any assessment of language proficiency may well simply become just one part of much larger, and more detailed, part of the profile that already exists for every one of us who spends time online (Confessore, 2018). Indeed, ‘we are Big Data’ (Klous & Wieland, 2016), and part of a growing response across the social sciences towards ubiquitous computing (Foster, Ghani, Jarmin, Kreuter, & Lane, 2017), now firmly situated in a post-humanist era (Pennycook, 2018). My aim in this chapter is to intertwine three themes that draw on aspects of Tim’s work: the social dimensions in language assessment.
(McNamara & Roever, 2006), the challenge of theory (McNamara, 2015), and recent work on transdisciplinarity (The Douglas Fir Group, 2016) that will, I believe, inform any response to my imagined bold claim.

Published in the same year that Facebook was released, key concepts from McNamara and Roever (2006) provide a sound basis for future developments. Briefly, in line with what they have asserted, future researchers will need to understand that language tests are indeed value laden instruments of social policy. Those same researchers will need to draw from social theory and cultural studies, rather than psychometrics, to further develop language testing research. Importantly, as social media further mature and are ever more deeply implicated in our contemporary world, it is crucial that social pragmatics and interactional competence remain central to defining constructs inherent in language proficiency.

By any measure, it is clear that social media use is on the rise. Available in a wide variety of forms, including microblogs, photo-sharing platforms and social networks themselves, it is estimated that there will be no less than three billion active social media users by 2021. The average young adult now spends no less than eight hours a day online (Statista, 2018a). By contrast, face-to-face communication amongst most of us in the industrialized world is now a mere 90 minutes per day (Statista, 2018a).

Research in social media across the disciplines, of course, is also on the rise. A brief survey demonstrates a growing scholarship concerning the role of social media within areas that include social theory (Fuchs, 2017), second language acquisition and learning (The Douglas Fir Group, 2016; Thorne & May, 2017), pragmatics (Herring, Stein, & Virtanen, 2013; Hoffmann & Bublitz, 2017), multilingual research methods (Zhravleva, de Bot, & Haug Hilton, 2016), discourse (Herring, 2015; Zappavigna, 2018) and, of course, language assessment itself (Arisle & Burston, 2017; Gruba, 2014; Li & Tsai, 2015).

As concerns of validity would be important to any claim that I make (Kane, 2016), can any warrant be found regarding the need to consider the use of social media in language assessment? Three points underpin an early defence of such consideration:

1. that the effective use of social media is integral to academic and professional literacies (Malinowski & Kramsch, 2014; Thorne, Sauro, & Smith, 2015), and that those who can use it well are seen to be strong influencers in realms of government, politics and business (Hennessy, 2018);
2. that poor proficiency in social media use can harm authentic participation, for example, in government, medical and institutional support services (Leppänen, Möller, Norroby, Stehr, & Kytölä, 2015);
3. that assessors, long situated in 'pencil and paper' views of education, must come to better 'recognize' contemporary modes of communication and what it now means to learn in their design of new assessment instruments (Douglas & Hegelheimer, 2007; Jewitt, 2005; Kress, 2009, 2013).

With social media now at the heart of interactions between intimate individuals all the way through to diverse global audiences, I believe that our field will move increasingly towards the new directions made possible through the use and analysis of Big Data.

One magic in reading the autobiographical work of McNamara (2013) is to see how language proficiency influences self-identity, social relationship and accomplishment. The use of selfies, blogs and mobile devices are now clearly implicated in the way we see ourselves (Rettberg, 2014). To advance our discussions here, let’s take a cue from studies in the cultural constructs of identity (Urrieta & Noblit, 2018) and work in study abroad (Beaven & Spencer-Oatey, 2016). Imagine a proficient ESL international graduate student who enrols at an Australian university, and who perhaps runs into difficulties adjusting to her new life in Australia during her first semester. The university often uses social media platforms, particularly the microblogging site Twitter, to alert students to enrolment, library and academic support services (Lemon, 2013) but our imagined student does not pay attention to tweets and misses crucial information (Gammon & White, 2011; Sandel, 2014). Although her subjects are somewhat grounded in sociocultural theory and participation is strongly encouraged, her shyness precludes her from classroom interactions and other face-to-face social interaction (Sawir, Marginson, Deurnt, Nyland, & Ramia, 2007). Halfway through the semester, she encounters difficulties in adjusting, but is hesitant to seek campus-based counselling services. Going online, though, she may use odd phrasing that causes her to be mocked and ignored in ways similar to those in reported studies (cf., Li, Zhu, & Jiao, 2018; Stommel & Mejman, 2011).

At present, most universities use language tests as part of the selection process based on the claims of test developers that “the student concerned can or cannot cope without additional help with the communicative tasks facing him or her in the study setting, such as following lectures, participating in class discussions, reading cognitively and linguistically sophisticated texts, and writing in a formal style” (McNamara & Ryan, 2011, p. 162). In light of decreasing attendance at university lectures and work on flipped classrooms (Kim, Park, Jang, & Nan, 2017), the validity of the current language assessment designs may be questioned by tertiary students; to my knowledge, neither university admissions criteria nor the required language test include any reference to the social media proficiency
that our students may well need to engage in official university business. Further, students may wonder why language tests require them to demonstrate oral interaction skills but provide no means for demonstrating their performance in crucial online interactions. Though I am not suggesting that tests alone would resolve many such issues, language assessment regimes and scores are now ever more focused on real world actions and outcomes (Frost & McNamara, 2018; McNamara & Roever, 2006) in ways that have caught the attention of governments such that “test constructs are increasingly dictated as a function of policy” (McNamara & Ryan, 2011, p. 175). Tests, in many ways, signal to learners and others what is important, which requires our attention, and what is valued across contexts. With the work of McNamara (2014) in mind concerning the implications of language testing, I would argue that the tests used for university entrance lack recognition of a range of larger sociocultural and educational forces, including an exponential rise in social media use, and thus run the risk of ignoring factors that now shape contemporary education (Kress, 2013).

To further build concepts into our ongoing discussion, I draw on Talmy (2010) to explore contrasts across two metaphors of language assessment (Tab. 1).

Briefly, with reference to Tab. 1, concepts that frame language assessment will shift with the rise of social media use. The primary endeavour of language testing will shift in purpose as we embrace Big Data, and indeed the status of the data itself will be influenced. The sources of data for our assessment will be the first problem we encounter. Should we gather our own data, through calibrated and valid instruments, or gather it from social media?

I think that we will come to gather information from candidates not just on set occasions (e.g., sitting for a test) in category variables (e.g., responses to items and scores) but rather we will harvest vast amounts of data that is longitudinal (e.g., over a lifetime) and continuous (e.g., generated by ubiquitous computing and being “always on”). Our analyses will move to examine time-based slices of interaction over a lifespan, chosen at times by candidates themselves or nominated by those who need a real-time and up-to-date profile of language ability. Dozens of language proficiency scores will be produced for individuals for purposes as specific as an ability to navigate translingual spaces, the strength and reach of global social networks, and time spent learning online. At present, the ride sharing company Uber calculates the rating of a driver on no less than an average of 500 trips (Uber, 2018a). Single scores, valid solely with reference to academic language proficiency, will appear dated and be insufficient to assess the abilities of a candidate to interact effectively through social media.

Other concepts will evolve, too. Candidates, now seen in single page reports, will be made much more visible through extensive lifetime profiles that constantly

| Tab. 1: Contrasting concepts in two metaphors of language assessment. Based on Talmy, 2010, p. 132 |
|-----------------------------------------------|---------------|
| **Language assessment as the development, use and effects of high-stakes tests** | **Language assessment as the periodic evaluation of online social interaction over a lifetime** |
| Status of the endeavour | Creation and validation of reliable tools and procedures for gathering information | A transferable and dependable profile of candidate language use |
| Status of the data | Data are collected at relatively brief set times that provide a snapshot of ability; scores indicate ability at a singular point of time and are valid no more than two years; Individual data is set within overall test results | Big Data are the culmination of an average of eight hours per day of online interaction over decades of ubiquitous social computing; scores result from longitudinal data that indicate temporal and situational variations |
| Candidate voice and interlocutors | Candidates are largely silent when responding to set items; ideal trained interlocutors provide standardized interactions | Candidates are made visible through a lifetime profile constructed through Big Data; interlocutors, numbering in the thousands, include individuals and AI bots |
| Bias, security and ethics | Test development and rating strives to minimize data contamination; test security is paramount; ethics are aligned with a non-binding code of practice | A range of algorithms, applications and platforms are implicated in Big Data sets; candidates control access and privacy to personal data sets; ethics are situational and negotiated informed by law |
| Analytic focus | Product-oriented with specific focus on particular times, item sets, and purpose | Process-oriented with varied emphases on time, tasks, context and interlocutors |

grow and flex. Importantly, ethics and legal concerns will be discussed, if not legislated, across large swaths of human endeavours far beyond the contemporary code of ethics of language testers (ILTA, 2018). Because language testing is currently seen to be a ‘weak profession’ that does not come under the same scrutiny as the ‘stronger’ areas of law or medicine, for example, an awareness of ethical considerations must grow in the language testing community (Georgakopoulou, 2017). In short, I think we—or perhaps more accurately, future language assessors—will be asked to seek information pertaining to individual language
abilities who will be trained in a range of transdisciplinary skills through a sea of continuous variables. Our willingness to honour such requests will inevitably raise ethical issues. Throughout such changes, returning to McNamara (2015) reminds us to ask an important question: where will our theoretical sources be located? My response, in contrast to the possibility that we will be taking cues from the humanities and postmodern theory, is that many of our sources will be at home in social informatics, as outlined in Stokols (2018).

An imagined rise in the use of social media and Big Data in language testing leads to yet another provocation: Future language assessors will be employed by large technology companies as they alone with have the depth of resources needed to work effectively with Big Data. Already one of the most industry-facing and lucrative areas of applied linguistics, language test development will be increasingly funded by companies who understand full well the significant social and commercial value of assessing individual abilities, interactions and social networks; language testers, too, know that that their efforts are instruments of social and political control (Frost & McNamara, 2018; McNamara & Roevek; 2006) in ways that will have a deeper resonance in the years ahead.

At present, there are a great number of online language assessments that appear simply by using the key phrase ‘online language tests’, for example. Many online instruments appear to have been produced by (currently) unknown companies, and yet several others are supported by well-known sponsors, universities and government agencies (Winke & Isbells, 2017). Though noting that Kunnan (2018) offers a framework for evaluating such language assessments, I would like to sidestep the many issues of design that face computer-based testing (including a lack of research; see, for example, Chapelle & Voss, 2016) in preference for a pivot towards a wholesale shift in our field towards Big Data. To pinpoint issues, I sketch out points of discussion to indicate, perhaps, how differing constructions of language assessment may influence our emerging conceptualization in Tab. 2.

My thoughts set out in Tab. 2 attempt to sketch the points of a trend that sees language testing go ever more online and, in the process, become more diffuse, less controlled and move into areas not yet fully understood. Against our developing views of language in a posthumanist era (Pennycook, 2018), we can readily see the power of the technology companies to sell “thousands of data points” that enables advertisers to “make increasingly precise guesses about what you wanted, what you feared and what you might do next: Quit your job, for example, or have an affair, or get a divorce” (Confessore, 2018, n.p.) If advertisers can already do this with Big Data, then why wouldn’t such predictive abilities be attractive to language assessors? Although companies are now legally constrained from gathering and selling data because of the European Union General Data Protection (European Union, 2018), for example, many people never read obscure and dated ‘privacy policies’ when they access online companies and social media sites (Turow, 2018). Language assessment specialists need not remain on the sidelines as concerns rise. Given their deep understanding of the matters the ways data and results affect the legal rights of individuals, as Lo Bianco (2018) has noted, language testers are in a unique position to inform policy discussions. Here, then, we can make transdisciplinary contributions as suggested by McNamara (2015).

<table>
<thead>
<tr>
<th></th>
<th>High stakes; wholly online;</th>
<th>Low stakes; wholly online;</th>
<th>Assessment of language across a lifetime of online social interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proctoring and security</td>
<td>Time bound in a single location; secured sitting including identification checks</td>
<td>Multiple sittings across locations no security and no identification checks</td>
<td>Average 6–8 hours a day for decades; identification verified by patterns of usage across time</td>
</tr>
<tr>
<td>Technology; modality</td>
<td>Secure computer; single screen</td>
<td>Personal computer, mobile devices; multimodal</td>
<td>Personal computer, mobile devices; multimodal</td>
</tr>
<tr>
<td>Opportunities for interaction with people and/ or bots</td>
<td>Limited; pre-determined by set protocols</td>
<td>Non-existent opportunities with people; limited interaction with bots</td>
<td>Continual, persistent and longitudinal</td>
</tr>
<tr>
<td>Outcome</td>
<td>Score with use-by-date based on set criteria for performance</td>
<td>Multiple scores that can be averaged over time based on multiple sets of criteria for performance</td>
<td>Section of a comprehensive profile; continually updated and modified performance indicators</td>
</tr>
<tr>
<td>Domain</td>
<td>Intended use</td>
<td>Intended use</td>
<td>Actual use in the target domain</td>
</tr>
<tr>
<td>Threats to validity</td>
<td>Actual performance</td>
<td>Lack of security</td>
<td>Transferability to face-to-face communication</td>
</tr>
<tr>
<td>Threats to reliability</td>
<td>Flaws in design and instrumentation</td>
<td>Flaws in design and instrumentation</td>
<td>Fluctuations in performance</td>
</tr>
</tbody>
</table>
Clearly, our futures will include much more interaction with robots (Herring, 2016) and indeed we can expect that future work on "testing language use and social context could use an infusion of testing experts with a strong knowledge of discourse analysis" (McNamara & Roever, 2006, p. 253). Such discourse analysis, as language assessment shift towards Big Data, will necessarily involve computer-mediated discourse analysis, multimodal discourse analysis and a range of emerging analytical theories and methods coming out of computational linguistics (Tannen, Hamilton, & Schiffrin, 2015). In these analyses, a greater focus on social pragmatic knowledge will also be needed (McNamara & Roever, 2006); indeed, such a focus has found a place in studies of online interaction (Herring, 2015). Work on fluency, for example, from a range of perspectives in applied linguistics (de Jong, 2018) may be able spur thinking about what 'interaction' may come to mean in a wide range of online spaces that include how we interact in videoconferences, within massively multiplayer gaming environments, across microblogging platforms, in discussion forums and throughout social networks sites. An understanding of interaction and pragmatics, initially grounded in face-to-face settings, will widen to online and transdisciplinary sources.

Perhaps dystopian, if not prescient, the comedic first episode of the third season entitled 'Nosedive' of the British science fiction series Black Mirror imagines a time when people rate each other from one to five stars in each and every interaction; such ratings are not benign but rather have an immediate impact on the rise and fall of the socioeconomic status of individuals (Netflix, 2016). Though I immediately laughed aloud at this episode, questions soon arose about its implications for language assessment: Would continual ratings of social interaction be a dream or a nightmare? What if communicative competence was grounded in an online interaction across the lifetimes of billions of multilingual people?

Our response to such questions can be based in the continuous and unrelenting present of today: that is, informed by our sensitivity and understanding of rating, we can begin to investigate the experiences of mutual ratings produced by the approximately 5.6 billion people who now use peer-to-peer ridesharing (Statista, 2018b). Ratings, as we know, have real world consequences: soon, the use of Uber in Australia and New Zealand will be restricted to passengers who score above four on a five-point rating scale (BBC, 2018); language skills play a part in the ratings of drivers as they are urged to "keep the conversation polite, professional and respectful" with passengers (Uber, 2018b).

How does my initial outrageous claim, further provocations and sketched concepts help to situate the work of McNamara? For me, Tim's work has always been a source of inspiration, of opening, of possibility. There is a sense, when I have read his work over the years, in which I find myself asking the 'what if' question and seeking to identify the implications of pushing the boundaries of language testing and, indeed, the wider field of applied linguistics. My view of language use is informed by research on technology, and through that lens the implications of Tim's work can be seen clearly: assessment and social theory interact, and thus the relevance of language testing lies its relation to society.

In my own efforts to make sense of such an implication, I often return to McNamara (2015) and its co-published article from Kramsch (2015). Both scholars ask us, as applied linguists, to consider the role of theory in our field. Thinking of the future, Kramsch (2015) writes that applied linguistics must engage deeply in the "struggle between gaining the technical recognition of institutional and corporate sponsors by striving to improve the practice on one hand and on the other, gaining symbolic scholarly legitimacy by refining the theory" (p. 457). For me, that struggle will be first be played out in the ways that language assessors come to use and adopt technologies, and, secondly, come to contribute their understandings of power, interaction and language across transdisciplinary boundaries.

In the years ahead, the work of McNamara will continue to offer clues on how we can shape our contribution to other fields, and necessarily signal that there are key struggles ahead as we face challenges of theory and development and come to better understand that "there is a lot of rethinking, and debating, to do" (McNamara, 2015, p. 476) in light of the rise of Big Data. Many of those debates, centered on the kinds of problems we seek to address, our sources of theory, and the transdisciplinary contributions we can make, will reference Tim's enormous contribution to our dynamic field.

References


Lemon, N. (2013). @Twitter is always wondering what's happening: Learning with and through social networks in higher education. In M. Pätät & B. Pätät (Eds.), *Social media in higher education: Teaching in Web 2.0* (pp. 237–261). Hershey, PA: IGI Global.


