Published version: Pill, J., Frost, K., & Macqueen, S. (2024). Fairness and justice in language testing: The challenge of Tim McNamara's legacy. *Language Testing*. https://doi.org/10.1177/02655322241277268

Fairness and justice in language testing: The challenge of Tim McNamara's legacy

The Virtual Special Issue of *Language Testing* that we have brought together through this editorial is intended as a tribute to Tim McNamara and his substantial contribution to the field of language testing and assessment. We seek to remind long-standing readers of *Language Testing* of the different aspects of Tim's work that have been represented in the journal, and to introduce this work to new readers, noting both the ongoing relevance to our field of the issues he considered important and the value to us of revisiting and reflecting on them. We estimate that the fundamental challenges posed in Tim's research and writing will continue to resonate for years to come, even as contexts and perspectives inevitably change. (We agreed, as guest editors, to refer to our subject by his given name "Tim", making this editorial more personal than academic style generally requires.)

The 17 articles authored or co-authored by Tim McNamara in the journal *Language Testing* from 1990 to 2019 represent the breadth of his scholarship in the field of language testing and assessment. Among them, we see his unwavering commitment to untangling an inherent tension at the heart of language testing – as both a practice and a field of expertise – between fairness, an internally derived property of tests, and justice, which he saw as pertaining to test uses and consequences. Tim's articles in *Language Testing* consistently highlight the potential for measurement theory and methods not only to address practical concerns and enhance test fairness, but also to progress theoretical understandings of the socially situated, co-constructed nature of language abilities and language performances, and, in so doing, to support more just approaches to language

teaching and assessment. Across his contributions in the journal, we see threads that he developed elsewhere, including synergies between innovations in language testing and in applied linguistics, which have informed his wider scholarship and remain relevant to current concerns in language testing. His pioneering work in the areas of Rasch measurement and specific-purpose language testing, most notably in the Occupational English Test (OET), is indicative of his awareness of the professional and wider societal norms and values that underlie judgments of the legitimacy or otherwise of language practices. His work provides insights into the uses of language tests to enact professional identities and domains of expertise, including the potential for test constructs and uses to perpetuate implicit social and cultural biases. A further characteristic of Tim's scholarship is how he could weave, with great clarity and integrity, ideas from a range of disciplines through his work, both in the single papers collected here and in his career-long agenda that focused on, among other things, forging a critical social path for measurement theory and assessment practice.

In the six papers from *Language Testing* that we have selected for this Virtual Special Issue, we see Tim combine measurement methods, philosophy, social theories, and the practical know-how of pedagogy, just as he did in his wider scholarship. What is most obvious in revisiting his work across the years is how so much of it remains current, posing still-relevant challenges for the language assessment community. The six papers herein are organised to illustrate three key areas of Tim's work: (i) his work on the social dimensions of language testing; (ii) his work on Rasch measurement, and (iii) his work on performance assessment, particularly in the context of the Occupational English Test (OET). What is so valuable about Tim's work is not that it continues to be relevant merely in the challenges it raises, but also that it provides such useful practical and theoretical tools with which to address fundamental questions. These challenges and tools can be observed in the following papers included in this Virtual Special Issue (with the full references for the papers given in the Appendix):

Year	Co-author	Title	Focus
(i) Social dimensions of language testing			
2001	-	Language assessment as social practice: challenges for research	Insistence on the social character of language constructs
2012	Kathryn Hill	Developing a comprehensive, empirically based research framework for classroom-based assessment	Legitimisation of assessment practices in learning contexts and a framework for them as the object of validation research
(ii) Rasch measurement			
1998	Brian Lynch	Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants	Innovations in Rasch analysis and its application in language testing
2012	Ute Knoch	The Rasch wars: The emergence of Rasch measurement in language testing	
(iii) Performance assessment			
1990	-	Item Response Theory and the validation of an ESP test for health professionals	Impetus to broaden the test construct to include what matters
2016	John Pill	How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals	Importance of engaging informants who understand the domain

The social character of language constructs

Questions of social identity were a long-standing interest of Tim's, from his early research on the multiple group memberships of Israeli immigrants in Melbourne to the culmination of his life's work in his book *Language and Subjectivity* (2019). Tim's academic background in philosophy and his enduring interest in the relationship between social values and language aligned naturally with the

work of validity theorist, Samuel Messick. As is well known, Messick's philosophically grounded conceptualisation of validity emphasised the values inherent in assessment constructs, and the context and consequences of assessment. Tim admired the depth and challenge of Messick's proposal that the notion of validity encompass not just the evidential basis for the interpretation and use of a test but also the consequences. Throughout his career, Tim returned to Messick's progressive matrix setting out the facets of validity (Messick, 1989, p. 20) as a way to understand score meanings and uses in various policy contexts, and, later, to distinguish between achieving fairness, a technical quality of testing, and justice, a societal one (McNamara & Ryan, 2011). Like Messick, Tim had an enduring interdisciplinary curiosity. He sought out theoretical perspectives, such as those of Judith Butler, Jacques Derrida and Michel Foucault, that would explain, in robust social terms, what tests are, and what they do. In a sense, this purpose was energised by diverse perspectives on *language* as an object of study – in the humanities on the one hand, and in the social sciences on the other.

Tim's 2001 paper "Language assessment as social practice: Challenges for research" is in this vein. The paper is significant as one of his first articulations for a language testing audience of the social character of language constructs which was to drive much of his later work. Indeed, in the paper, Tim urges the field to re-evaluate its practices and assumptions, as a necessary process of evolution. The paper appeared in the special issue of *Language Testing* on "alternative assessment", for which Tim was also guest editor (McNamara, 2001). Tim explains in his editorial that the issue originated in an invitation to convene a colloquium on alternative assessment at the 2000 conference of the American Association for Applied Linguistics. In typical boundary-pushing style, he decided to interpret the theme broadly, beyond school contexts, where the concept of "alternative assessment" originated, extending the notion to encompass "alternatives or challenges to the current mainstream in language testing research both at the level of theory and at the level of practice" (p. 329). In his paper, Tim draws together two notions of *performance*. First, he refers to the familiar concept of performance testing, in which he problematises the idea that an individual's language competence is readily available through eliciting and observing a jointly and socially constructed performance. He then extrapolates

the tester's conundrum to Judith Butler's notion of *performativity* which (in our overly simple terms) holds that social identity performances nurture a belief in the performer that they reflect an "inner essence" (p. 339). Applying Messick's (1989) matrix, Tim poses the question: "What if the act of testing itself constructs the notion of language proficiency?" (p. 339). Although the paper is now over 20 years old, the question remains relevant, particularly as we witness operationalised constructs that are increasingly dictated by the constraints and affordances of technology and the values that lie therein. Finally, Tim sets out a critique of managerialist impositions (such as accountability assessment practices) that work contrary to the needs of teachers and learners, and calls for an alternative approach to assessment that legitimises the needs of teachers and classroom practices. These concerns, too, have only increased in relevance over time. Teachers in the Australian context (and likely in any country drawn into the comparative discourses around large-scale international testing regimes such as the Programme for International Student Assessment [PISA]) are increasingly required to de-emphasise the kind of "constructively critical reflection" (p. 344) that Tim proposes in this paper, in favour of pseudo-experimental methods to generate data points with which to gauge students' progress.

Co-authored with Kathryn Hill, the 2012 paper titled "Developing a comprehensive, empirically based research framework for classroom-based assessment*",* is a thought-provoking companion to "Language assessment as social practice" (2001). Here, the authors heed Tim's prior call to legitimise the needs of teachers and give serious attention to assessment practices in classrooms. In contrast to the broad theoretical challenges he issues in the earlier paper, attention here is on the minutiae of classroom practice, gathered by Hill in her ethnographic doctoral study on Indonesian language learning in Australian schools (2012). The practices include classroom interactions, teachers' notes, rubrics and a range of other artefacts and insights. From this rich documentation, the authors construct a multiperspectival framework for understanding assessment in instructional contexts, from corrective feedback in everyday routines to summative reporting over weeks, through the eyes of teachers, students and their observers. Their contribution is not just for researchers of classrooms.

They also remind the broader language testing community, with its tendency to focus on large-scale testing practices, about the legitimate complexity of assessment in instructional sites. Their resulting framework continues to have much to offer researchers examining assessment practices in instructional contexts, as they shift with the changes in technology that have occurred since this paper was published.

Rasch measurement in language testing

Tim's work in the area of Rasch measurement can be situated within his wider preoccupation with bridging a tension or gap between the two areas of expertise he viewed as constituting the field of language testing: expertise in language and language use on the one hand, and expertise in statistics and measurement on the other. As he sets out in a commentary piece in *Language Testing* (McNamara, 2011), language testers typically come to the field from one or the other side, and rarely both. Tim himself entered the field in the late 1980s with a background in English teaching, and thus associated himself more with the "language" side of language testing at the outset of his career, but worked very quickly to establish robust expertise in measurement, which was, in his view, integral to a deep understanding of language testing, as a discipline and as a social practice. This importance of understanding measurement principles and practices was manifested in his teaching and his scholarship over the decades ahead, not only in language testing, but in applied linguistics generally. His interest in Rasch measurement in particular, which formed a key part of his doctoral project on the OET in the late 1980s, emerged in response to the psychometric challenges created by the role of human judgments in performance-based testing for professional purposes, and was developed over a series of papers in the 1990s. In one of these papers, co-authored with Brian Lynch and published in *Language Testing* in 1998 ("Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants"), the authors engage with Bachman et al. (1995) in a dialogue over the respective value of Generalizability Theory and the Rasch model for enhancing test design and fairness in language testing. This signals Tim's ongoing interest in

understanding the technical and theoretical aspects of measurement, both as a means of promoting best practice, and of shedding light on the complexities inherent in any language performance. The 1998 paper, together with others exploring applications of Rasch measurement in *Language Testing* (McNamara, 1990, see further below; McNamara & Lumley, 1997), illustrates the value of the Rasch model in identifying and compensating for the impact on score outcomes of rater inconsistencies and rater severity, as well as the effects of different item and task types.

In a subsequent 2012 paper, "The Rasch wars: The emergence of Rasch measurement in language testing", based on a survey of articles published over three decades (1980s, 1990s and 2000s) across four research journals in the field (*Language Assessment Quarterly*, *Language Testing*, *Melbourne Papers in Language Testing* [now *Studies in Language Assessment*], and *Assessing Writing*), Tim and Ute Knoch characterise a shift from early debates around the appropriateness of Rasch measurement in the field of language testing, to the ultimate acceptance of its utility for addressing validity concerns, especially those related to performance-based language assessments. In early debates, as the authors describe, questions were raised over the appropriateness of Item Response Theory (IRT) models, including Rasch, particularly the assumption of a single measurement dimension and its implications for language test constructs; at the same time, even the proponents of IRT viewed the one-parameter Rasch model as an oversimplification, especially concerning item properties, compared to two- and three-parameter models. The authors highlight the role of differences between British and US traditions in language testing in the 1970s and 1980s that drove these early debates, and they situate Australia as somewhere Rasch-based research particularly flourished, perhaps because both language testing and applied linguistics were nascent disciplines at the time. Tim, as the founder of the first applied linguistics program at the University of Melbourne in 1987, and co-founder of the Language Testing Research Centre at the same institution in 1990, was a significant figure in bringing Rasch measurement into the mainstream in language testing. Interestingly, as is outlined in the article, Tim himself variously occupied these three traditions, physically and intellectually, as he moved between Australia, the UK, and the US in the foundational years of his career, establishing deep

personal and interdisciplinary ties, which he then built on and maintained over his time in the field. The tracing of the history of debates around Rasch measurement reflects a thematic trend across Tim's scholarship; expertise and disciplinary practices in language testing are closely connected to the people whose work generated shifts in the field, not only to their names, but also to their backgrounds and the wider contexts in which they were educated and through which their perspectives and values were shaped. In "The Rasch wars" paper, the situated account of the evolution in thinking about Rasch measurement is thus emblematic of Tim's wider interest in foregrounding the values and ideologies underlying test constructs and language testing practices, which was the focus of much of his work.

Assessment of specific-purpose language performance

Tim's long engagement with the OET, which involved substantial contributions to theoretical and practical developments in the field of specific-purpose language testing, is illustrated through his earliest publication in *Language Testing* (in 1990) and some of the most recent in the same journal, over 25 years later. He developed this specific-purpose English language test for migrating healthcare professionals seeking to train and practise in the increasingly multicultural context of Australia in the late 1980s as a project for the federal government, and, as already noted, this work also came to form his doctoral thesis, submitted at the University of Melbourne. Initial publications presented practical validation studies for the component sub-tests, using early datasets from live tests and investigating the merits of Item Response Theory (IRT) and the Rasch model. The previous section has foregrounded how Tim took a leading role in introducing these statistical approaches to the field of language testing (see also McNamara, 1996). The innovation promoted in the 1990 article ("Item Response Theory and the validation of an ESP test for health professionals") is how Rasch measurement can be used – beyond its purpose to investigate a test's reliability – as a way to validate a test, by showing the extent to which (for example) assessment criteria complement each other and indicate unidimensionality, this then being interpreted as evidence of a common underlying construct.

It is instructive, from today's viewpoint, to consider the criteria identified in the 1990 article to assess the speaking and writing sub-tests of the OET. In these criteria, Tim strove to operationalise a broader construct of communicative competence than that found in many other tests of that time or, we might observe, in more contemporary assessment schemes. For example, the holistic criteria of *overall communicative effectiveness* (for speaking) and *overall task completion* (for writing) sought to capture more than just the sum of the analytic linguistic criteria, recognising a wider scope of performance (of "getting the job done") required in the healthcare domain. The term *intelligibility* was used as a criterion in the speaking sub-test (rather than *pronunciation*) and *appropriateness of language* was foregrounded, indicating the importance to the test construct of patients' and co-workers' expectations in real-world healthcare contexts, where, in Australia and other contexts where the OET was used, interactions would often occur between participants from different linguistic and cultural backgrounds. One speaking criterion, *comprehension* (of the interlocutor), did not survive test revisions, which changed the assessment scheme from a semantic differential format – a scale of six points with defined extremes (e.g., "rich, flexible" vs. "limited" being used for the speaking criterion *resources of grammar and expression*; 1990, p. 75) – to fuller descriptors for each of six levels of performance. Nevertheless, the overall definition of the assessment criteria for these performance tests for healthcare practitioners remained largely unchanged in the operational test until the late-2010s (see below). Tim's initial rating scales were progressive, pursuing the construct of communicative competence.

As a complement to Tim's first publication in *Language Testing*, we turn to a later article (2016), co-authored with John Pill, that also considers the OET, its assessment criteria, and challenges arising in the rating of specific-purpose language tests: "How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals". This article appeared in the *Language Testing* special issue on "authenticity in LSP testing" (Elder, 2016) comprising research papers from an Australian Research Council (ARC) funded project that developed and trialled a revised set of OET speaking criteria to capture more effectively the language and

communication skills required by healthcare professionals in interaction with patients (and others). The interdisciplinary project, involving applied linguists and clinical educators in medicine, nursing and physiotherapy, empirically addressed Tim's longstanding questioning (see also Jacoby & McNamara, 1999) of the value of assessment criteria that do not adequately reflect the performance features that matter to participants in a particular domain, that is, their *indigenous assessment criteria* for the context and task at hand. Applying the expanded set of speaking criteria that was developed in the research project to be more relevant to the test's healthcare settings and to encompass aspects of "clinical communicative competence" (Pill & McNamara, 2016, p. 218), the authors report on a standard-setting exercise that engaged with persistent methodological and theoretical concerns in specific-purpose language testing. These included capturing and formalising evaluations made by domain "insiders" (in this case, practitioners in different healthcare professions), who provided a holistic perspective of test performances. They also judged test takers' readiness to participate in the domain and determined whether trained raters' (i.e., English language teachers') ratings of the same performances and using the revised speaking assessment criteria were sufficiently (authentically) aligned with those insider evaluations. The commensurability of insider and outsider perspectives in specific-purpose language testing, and, consequently, the capacity of raters engaged in operational testing to be effective proxies representing the views of others may often be taken for granted in standard-setting studies and test administration, but here Tim sets out the theoretical conundrum and attends to its practicalities. The article illustrates how procedures used in language test development inevitably entail human judgement and the understanding of contextual factors, once again showing Tim's attention to interrogating the connections between technical and social systems in his published work.

The new speaking assessment criteria developed in the ARC project were subsequently implemented in the operational version of the OET in 2018. This revision can be viewed as one reason for the test being taken up by further professional regulators in several countries in recent years. For example, the test is now used to measure the English language and professional communication skills

of physicians trained outside the US who seek certification to train and practise in that country, taking the place of the substantial practical assessment ("Step 2 Clinical Skills") in the United States Medical Licensing Examination (Mladenovic et al., 2023). Tim was rightly proud of this achievement for the test that he established and continued to support throughout his career. The OET has become one of very few genuinely specific-purpose language tests in widespread use at the global level, as demand continues to grow for skilled migrants to join the healthcare sector workforce in the more economically developed countries where English is used (OET, n.d.).

Legacy and challenge

The title given to our Virtual Special Issue editorial plays on the title Tim used for an article discussing the importance of the work of Samuel Messick: "Validity in language testing: The challenge of Sam Messick's legacy" (McNamara, 2006). Tim first presented the content of that article in the inaugural Messick Memorial Lecture at the Language Testing Research Colloquium in 1999, following Samuel Messick's death in 1998. It was subsequently published (in 2006) with a postscript, in which Tim exemplified how Messick's work continued to influence thinking and research in educational measurement and, consequently, also in language testing. This publication also spearheaded what has become something of a tradition for *Language Testing* to publish high-calibre, potentially high-impact Messick lectures with broad appeal (see Randall et al., 2024, for a recent example). Because of the breadth and depth of Tim's work, we anticipate the ongoing stimulation and provocation of his thinking and research in language testing for many years to come. In this editorial, we have tried to demonstrate why Tim's work might be seen to embrace the goals of "fairness and justice in language testing" and to set out some of the enduring challenges for our field that we recognise in his articles in *Language Testing*.

Whether it is the first encounter or a return visit to a familiar source, we encourage readers to engage with the six articles we have selected here, illustrating the range and focus of Tim's scholarship. We appreciate the clarity and precision of his writing and recognise the underlying experience and

thought necessary to reach this transparency of expression. We also note how the interdisciplinary reciprocity characteristic of Tim's work in language testing remains pertinent to ongoing concerns in our field, especially regarding the societal impacts of language testing practices (see, e.g., Randall et al., 2024). Thus, while reading, we should also be open to the challenge of his legacy: the fundamental role of the social in language testing, the value of evidence and being open to new ways of measuring and interpreting it, and the need for fidelity to context and its inhabitants.

Acknowledgements

The authors thank the editors of *Language Testing* for their invitation to bring together this Virtual Special Issue to reflect the contribution of Tim McNamara to the journal.

Authors' note

The three guest editors were each fortunate to know Tim McNamara through their connections with the University of Melbourne and the Language Testing Research Centre. John Pill first met Tim while employed at the OET Centre in Melbourne. His subsequent doctoral studies at the University of Melbourne were co-supervised by Tim, Cathie Elder and Robyn Woodward-Kron, and funded through the ARC research project on OET speaking criteria described in the editorial. Kellie Frost was first a student of Tim's Language Testing course in the Master of Applied Linguistics at the University of Melbourne, then one of his many doctoral students, and subsequently his colleague in the Linguistics and Applied Linguistics department at the University of Melbourne. Susy Macqueen also took Tim's Language Testing course, and her first tutoring job at the University of Melbourne was with Tim.

References

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-257. https://doi.org/10.1177/026553229501200206

Elder, C. (Ed.). (2016). Authenticity in LSP testing (Special issue). *Language Testing*, *33*(2).

Hill, K. M. (2012). *Classroom-based assessment in the school foreign language classroom*. Peter Lang. https://doi.org/10.3726/978-3-653-01984-1

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes*, *18*(3), 213-241. https://doi.org/10.1016/S0889-4906(97)00053-7

McNamara, T. F. (1996). *Measuring second language performance*. Addison Wesley Longman.

McNamara, T. (2001). Rethinking alternative assessment [Editorial]. *Language Testing, 18*(4), 329-332. https://doi.org/10.1177/026553220101800401

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*(1), 31-51. https://doi.org/10.1207/s15434311laq0301_3

McNamara, T. (2011). Applied linguistics and measurement: A dialogue. *Language Testing, 28*(4), 435-440. https://doi.org/10.1177/0265532211413446

McNamara, T. (2019). *Language and subjectivity*. Cambridge University Press. https://doi.org/10.1017/9781108639606

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*(2), 140-156. https://doi.org/10.1177/026553229701400202

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly*, *8*(2), 161-178. https://doi.org/10.1080/15434303.2011.565438

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). American Council on Education; Macmillan.

Mladenovic, J., van Zanten, M., & Pinsky, W. W. (2023). Evolution of Educational Commission for Foreign Medical Graduates Certification in the absence of the USMLE Step 2 Clinical Skills Examination. *Academic Medicine, 98*(4), 444–447. https://doi.org/10.1097/ACM.0000000000005051

OET. (n.d.). Who recognizes the OET Test? https://oet.com/test/who-recognises-oet

Randall, J., Poe, M., Slomp, D., & Oliveri, M. E. (2024). Our validity looks like justice. Does yours? *Language Testing, 41*(1), 203-219. https://doi.org/10.1177/02655322231202947

Appendix

References of the articles constituting the VSI

1. McNamara, T. F. (1990). Item Response Theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), 52-76. https://doi.org/10.1177/026553229000700105
2. Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180. https://doi.org/10.1177/026553229801500202
3. McNamara, T. (2001). Language assessment as social practice: challenges for research. *Language Testing*, *18*(4), 333-349. https://doi.org/10.1177/026553220101800402
4. Hill, K., & McNamara, T. (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing, 29*(3), 395-420. https://doi.org/10.1177/0265532211428317
5. McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. Language Testing, 29(4), 555-576. https://doi.org/10.1177/0265532211430367
6. Pill, J., & McNamara, T. (2016). How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals. *Language Testing*, *33*(2), 217-234. https://doi.org/10.1177/0265532215607402