Towards 'policy responsible' language assessment: Framing the Language Testing Research Centre's contribution in languages other than English[1]

Cathie Elder
University of Melbourne

The importance of viewing tests in their policy contexts is now widely recognized in our field. The Language Testing Research Centre (LTRC), as a self-funding Centre, responds by necessity to policy shifts and the language testing initiatives it has undertaken over the years offer insights into broader societal and more local institutional imperatives. The Centre's language testing activities also have the potential to influence policy, whether directly or indirectly. This paper offers a historical overview of the Centre's various projects in the languages (other than English) arena over a 30-year period, describing the diverse orientations adopted (i.e. to inform, enact or evaluate policy) with particular reference to three cases illustrative of different policy trends. I speculate about the policy impact of each case, highlighting the complexities encountered and some of the factors favouring and inhibiting policy uptake. Insights from these particular cases are then linked to the research literature on policy impact to propose recommendations for 'policy responsible' language testing.

Keywords: test impact, policy impact, languages other than English

Introduction

There is now widespread recognition that language tests need to be viewed in their policy contexts and that validation efforts should extend to interrogating language test use in such contexts. However, the ways our field talks about policy and the place of tests within the complex world of policy making seem, for the most part, to be quarantined from trends in the broader policy literature. Language testing scholars tend to characterize

[1] This paper is a revised version of my acceptance speech for the Cambridge/ILTA Distinguished Achievement Award. It was delivered on 17th June 2021 at the 42nd Language Testing Research Colloquium *'Assessment in Multilingual Contexts: Models Practices Policies & Challenges'*.

Email address for correspondence: caelder@unimelb.edu.au

policy mechanisms in linear terms, with policies seen as generating language tests, which in turn serve as vehicles for translating policy into action. Test impact is likewise, more often than not, cast as the influence of high stakes testing activity on teachers, learners and other stakeholders, with the test positioned at centre stage, rather than as part of a more complex web of sociopolitical interactions that both shape and are shaped by the policy environment.

In this paper I will reflect on what is known in the policy literature as the 'discursive web' and consider how this can illuminate the various interactions between policy and testing practice that have played out in projects conducted at the Language Testing Research Centre (LTRC), the University of Melbourne, the place where my career as a language tester was forged. It seems fitting in this anniversary issue to outline the Centre's extensive work in the languages arena and, drawing on examples of particular projects, to reflect on the multifaceted nature of its policy contribution as well as the factors that may have enhanced or constrained its policy impact.

I will move from these specific examples to a broader characterisation of the role of language testing activity within the policy 'web' proposing some principles for what I will term 'policy responsible' language testing.

Literature review

Language tests in their policy context

Discussions of policy in the language testing literature have been led by a small group of notable scholars such as Spolsky, Shohamy and McNamara. Spolsky (1995, 2001) was the first in our field to give sustained attention to the powerful institutional forces shaping the testing industry in the US and the decisions surrounding the creation of the TOEFL test. He showed that these forces were not necessarily benign, arguing for greater caution in the interpretation and uses of test scores and less reliance on test score alone for high stakes decision-making. Shohamy (e.g., 2001, 2003, 2008) cast her net more widely, exposing the power of tests and assessment standards as *de facto* policy instruments to perpetuate structural inequities and to constrain or distort what teachers teach and learners learn. McNamara (e.g., 2009, 2011), like Shohamy, emphasised the inherently political nature of language testing, linking this to the values dimension of Messick's formulation of test validity. He claims that our field's narrow focus on the technical qualities of our tests can blind us to the ideological and often discriminatory policy constructs which underlie them and to the social consequences of their use.

Given the influential role of policy in determining both the shape of tests and their potentially insidious impact, Knoch and Macqueen (2020) (see book review by Hill, this issue) propose that the assumptions underlying assessment-related policies be themselves subject to systematic review and that such review be an integral part of the test validation process. Focussing on tests for professional purposes, they lay out an argument-based validity framework which lists the kinds of evidence needed to support or refute a range of policy assumptions. Chalhoub-Deville (2009), drawing on her experiences with the NCLB policy in the US, invokes social impact assessment (SIA) as the basis for proactive policy intervention. She advocates that test developers and researchers join forces to map the potential impacts of high stakes educational reform policies and to allocate responsibilities for evaluating these impacts before such policies are implemented.

The above-mentioned scholars and indeed many of the policy-related articles in our field tend to focus on high stakes language tests. These tests are generally characterised as part of a chain in which policies are formulated by those in authority, then tests are put to the service of policy edicts and in turn impact on test takers and other stakeholders, often, as is pointed out, with less than optimum consequences. While the focus on tests as instruments of policy is valuable, it is important to recognise that policies are not fixtures, but themselves emerge from dynamic networks of intersecting interests which are subject to change or contestation by various parties. The capacity of language testers to intervene within such networks is rarely discussed. Historical overviews of policy evolution, like those of Spolsky (1995), Menken (2008) and Bunch (2011) help us better understand the processes entailed in implementing policy-driven testing regimes, but the role of language testers in responding to these regimes is seldom emphasised. While some authors (e.g., see Dwyer 2002) offer sensible recommendations for combatting the negative consequences of government policies, these are not usually directed to language testing professionals. In cases where language testers are invited to review policy, for example in relation to test accommodation provisions in the US (e.g., Stansfield et al. 2000), their findings are generally confined to commissioned reports and we seldom hear of their policy impact. And when language testers do intervene to bring about change, such as the evidenced based change to a university admissions policy reported by Ginther and Yan (2017), the processes involved in such interventions are hidden from view (but see Knoch, this issue, for a notable exception). Deygers and Malone (2019) comment on the lack of dialogue between policy makers and language testers, attributing this to the divergent concerns and priorities of the different parties. A lack of reciprocal understanding may be one reason for our field's limited influence in the policy-making arena (Deygers, 2021).

In sum, more attention needs to be paid to how language testing expertise can be harnessed not simply in enacting policy mandates via our test instruments, but also in using our expertise to shape policy development in various governmental or institutional contexts. Documenting the processes and outcomes of such activities and their impact in local policy environments, I will argue, is of value not just for the historical record, but also for building the 'policy literacy' (Lo Bianco, 2001, 2019) needed for our profession to better engage with future policy agendas.

Current conceptualizations of policy and policy impact

Current thinking amongst scholars of policy and planning eschews the narrow notion of policy as text, instead characterising it as a 'discursive web' (Goldberg 2006). This metaphor stems from post-structuralist conceptualisations of policy (e.g., Ball, 1993; Gale, 1999) and chimes with the Foucaultian view of discourse as a practice or action. Gale (1999) emphasises the intertextual nature of policy and planning activities, involving both pre-existing and emerging discourses or 'policy ensembles' that compete with or complement one another to different degrees and influence how policy is taken up.

Lo Bianco and Aliani (2013) elaborate on this notion of policy as an interdiscursive struggle in their exposition of the chequered history of languages policy in Australia, which is the context addressed in this paper. They point to the marked contrast between asserted realities of policy rhetoric as manifest in successive government policy documents spanning the last three decades and the lived realities of schools, ultimately resulting in consistent failure to achieve desired language learning outcomes or to meet the language learning needs and aspirations of the majority of its citizens. They attribute this failure not just to the reductionist nature of many policy documents but to the multiple points of potential disruption and subversion enacted at different levels of policy intervention. These levels are characterised as different discourses in a model of policy and planning that incorporates *intention, rhetoric,* and *experience*. *Intention* is about the official policy text, which is essentially a formal plan for a future new social or educational order (which itself arises from a particular set of social and historical circumstances) issued by bodies with the authority to allocate resources and manage implementation. *Rhetoric* is about the interpretation of the policy – the discussion and debates that occur amongst those (government bureaucrats, academics, community representatives and others) on whom these bodies depend if their authority is to be legitimised, confirmed and operationalised. *Experience* is about how policy, once enacted, is taken up or indeed contested by those for whom it is designed, whether this be language teachers and learners or other parties. These levels are not linear, with policy generating rhetoric then rhetoric generating experience but rather, as other policy scholars have argued, form a dynamic network or "discursive web", such that they are mutually constituting and interacting over different time frames. Lo Bianco & Aliani (2013) and Slaughter et al. (2019) show how these dynamic interactions play out in the Australian context, where successive top-down policy prescriptions for language learning programs in schools have been met by lukewarm or intermittent commitment from school principals, teachers and their students resulting in token language teaching programs and limited language gains in many cases.

Professional language testing activity, I would suggest, can operate at different levels of this policy network. We may apply our expertise formatively, offering advice which feeds into policy intentions or indeed, as Shohamy and others have pointed out, our tests may themselves, for better or for worse, become a *de facto* policy standing in place of any formal statement of intent. Our testing expertise can also create new discourses at Lo Bianco's middle *rhetoric* level of policy by interpreting, distilling, and offering instruments to operationalise officially stated policy intentions, alongside other solutions offered by different authorities charged with policy implementation. We may also position ourselves within the *experience* layer of the policy web, working with teachers and learners and other stakeholders on the receiving end of policy edicts to evaluate how a policy is unfolding on the ground, by observing test preparation activities or measuring language learning gains, for example.

Given the complex ecology of policy, estimating the impact of any single intervention is obviously a fraught exercise. Multiple influences at various levels of the policy network may function to diffuse expert knowledge and evidence and to foster or thwart policy implementation. The efficacy of our contributions will of course on how power is distributed between different arbiters (Johnson & Johnson 2014). Policy change, as Thomas (2007) points out, is itself dynamic, emerging from of a web of decisions that may reflect competing values and result in political compromises. Such decisions take time, meaning that impact is not always immediate. Whether impact is direct or less direct, immediate or delayed, will also depend on whether an intervention is oriented to conceptual (consciousness-raising) or instrumental (action-oriented) change (Nutley et al., 2007). Attempting to gauge policy impact is nevertheless worthwhile, as the outcomes of any initiative (even when uneven or non-existent) can influence future policy discussion, leading potentially to a process of learning, adaptation and improvement (Dunlop & Radaelli, 2018).

The current study: overview of policy work at the Language Testing Research Centre

The focus of this paper will be on the language tester's role and degree of agency and impact within the discursive web of policy as it plays out in the work of a particular testing organisation in Australia, the Language Testing Research Centre (LTRC) at the University of Melbourne. The account is written from an insider perspective, drawing on my personal experiences first as a junior researcher and some years later as Centre Director. The LTRC, founded late in 1989, was one of several applied linguistics centres created as subsidiaries of the National Language Institute of Australia (NLIA). The NLIA was set up in response to Australia's National Language Policy (NLP) (Lo Bianco, 1987, a landmark document which laid out the plurilingual principles and goals for the nation. The LTRC's mission encompassed various research and development activities in language assessment that would work for the furtherance of NLP goals. The Centre relied initially on seed funding from the Commonwealth and then, as time progressed, continued as a self-funding body, seeking grants from federal, state, and institutional sources to sustain its activities. These activities reflect the shifting language policy environment in Australia while also shaping policy in various ways, as will be seen.

To prepare for this paper I consulted Centre records as well as current and former Centre staff to draw up an inventory of assessment-related projects in languages other than English (hereafter LOTE) that have been conducted at the Centre over the thirty-year period since its inception in 1990. The LOTE acronym was once widely used in Australia rather than foreign language, in recognition of the fact that languages offered within the primary or secondary school system are not in fact foreign to those who study them given that they may be used by their parents at home. The term has since fallen out of favour amongst those arguing their case within the rhetoric layer of policy due to the implication that it "others" and thereby demotes the status of non-English languages. It has been replaced simply by 'languages' or 'additional language'. I use LOTE in this article both for historical reasons – this was the term previously in use - and because it makes it clear that projects involving English, whether as first or additional language, are not included in the inventory. My decision to focus on LOTE projects was to draw attention to areas of the Centre's work that have been central to its role both locally and nationally but may be less well-known than more widely publicised work relating to high stakes English assessment.

Table 1 lists the various projects chronologically, specifying for each one the source of funding (Column 4), aim (Column 5) and broad policy function (Column 6). It can be seen that many of these projects are funded by the Australian federal government, using resources allocated to official language policies. Others are commissioned by educational authorities at the state or regional level, whose jurisdictions and policies overlap with federal ones in some cases. Two have been directly funded by private schools. One project is funded by a regional health service in the state of New South Wales and another by a not-for-profit research organisation (the Australian Council for Educational Research) based in Victoria. The remainder have been resourced by grants from the Centre's host institution, the University of Melbourne. The language or range of languages addressed by each project also varies. Indigenous and immigrant languages are represented, reflecting various policies oriented to meeting the needs of Australia's multicultural and multilingual constituencies that have resulted in ongoing funding provision for multilingual services and for language education, including mainstream school language programs catering for heritage and non-heritage students (i.e., those with and without a home background in the target language). There is also a notable emphasis on Asian languages in response to successive national policies geared to Asia literacy and the study of Asian languages for utilitarian purposes, to serve Australia's political and economic interests in the region (Lo Bianco & Slaughter, 2009).

The aims of each project vary widely as does the type of testing activity undertaken. I have broadly classified the projects into three categories (Column 6) characterising their function within the policy network as follows:

- G: those directed to *guiding* policy formation
- I: those involving the development of a test or other assessment tool needed to *implement* a policy that had already been formulated
- E: those geared to *evaluating* the outcomes of a particular policy after it had been enacted.

It is acknowledged that these categories are not always mutually exclusive, but they serve to identify what those working on the project at the time perceived to be its prime function in the policy environment. Overall, it can be seen that the Centre's LOTE activities over the years have been fairly evenly spread across these categories: G (n = 9); I (n=8); E (n = 8).

Table 1. LOTE projects conducted at the LTRC over a 30-year period (1990-2020)

Dates	No	Name of project	Funding body	Aim of project	Policy function
1990-91	**1**	**Hebrew language immersion program evaluation**	Mt Scopus Secondary College	To gauge attitudes and learning achievements among participants in a late Hebrew immersion at a Jewish day school	**E**
1991 - 1995	**2**	**Australian Language Certificates** (test development and analysis). Various languages.	Australian Council of Educational Research	To recognise school-based language learning achievements and provide incentives for continuing language study	**I**

1993-4	**3**	**Proficiency tests for language teachers** (Italian, Japanese & Indonesian)	Commonwealth Department of Employment Education & Training	To assess readiness to teach languages in the school context. To contribute to the process of defining competency levels in relation to LOTE teaching and provide a national benchmark for language teacher educators	**I**
1993-5	**4**	**Tests of Japanese & Korean for tour guides**	Commonwealth Department of Employment Education & Training	To certify LOTE competence for the workplace and enhance employment opportunities for Japanese and Korean speakers	I
1993-4	**5**	**Placement tests for incoming university language students** (French, Japanese, German and Italian).	University of Melbourne	To assess entry proficiency levels of students to aid in placement decisions	I
1994-5	**6**	**Proficiency test for exiting Japanese university students**	University of Melbourne	To monitor outcomes of university language learning and provide certification of Japanese language competence	E
1995-7	**7**	**Common Assessment Tasks 2 & 3** (Japanese)	Victorian Board of Studies	To develop oral and reading assessments for end-of-school Japanese examinations	I
1995 - 1998	**8**	**Categorisation of LOTE learners by language background** (various languages)	Victorian Tertiary Admissions Centre	To assist the state assessment authority in implementing a compensatory scheme to rectify perceived bias in the end-of-school language examinations	G
1996	**9**	**The National Asian Languages in Schools project**	Australian Government Department of Employment, Education, Training & Youth Affairs	To survey student outcomes in Asian languages	G
1996-7	**10**	**Investigating the relationship between metalinguistic knowledge and success at university language study**	University of Melbourne	To inform curriculum design by exploring the contribution of metalinguistic knowledge to success in university language study	G

		(French, Italian, Chinese)			
1998-1999	**11**	**Evaluating bilingual programs in Victorian schools** (Chinese, Japanese, Arabic)	Victorian Department of Education and 6 Victorian government schools	To evaluate the implementation and language learning outcomes of state-funded bilingual programs	E
1998-9	**12**	**A description and exploratory evaluation of program types in indigenous and community languages** (Arabic, Khmer, Italian, Chinese, Noongar, Yindjibarndi)	Commonwealth Department of Education and Training & Youth Affairs	To consider the value of different approaches to language learning via exemplary case studies in Australian schools	E
1998-9	**13**	**A comparison of beginning and continuing students of French in Years 7, 8 & 9**	Presbyterian Ladies' College and Association for Independent Schools	To estimate the value of an early start in foreign language learning in a private school context	E
1998-9	**14**	**Longitudinal and Comparative Study of the Attainment of Language Proficiency** (French Italian, Indonesian and Japanese)	Commonwealth Department of Employment, Education, Training & Youth Affairs	To compare achievements across a range of language taught in school-based language programs	E
1999	**15**	**Bilingual health workers language assessment project**	South-Western Sydney Area Health Service	To investigate the feasibility of developing a test of oral skills in three community languages (Arabic, Spanish and Vietnamese) to measure linguistic competence of health workers interacting with NESB patients in the healthcare setting	G
1999-2000	**16**	**Development of annotated student work samples to accompany Curriculum & Standards Framework**	Victorian Department of Education	To assist teachers with interpreting, assessing and reporting student performance against the CSF	I

		(CSF) (Indonesian, French, Chinese and Japanese)			
2000	**17**	**Monitoring standards in education in languages other than English** (Japanese)	Western Australian Department of Education	To assess student performance in Japanese at two levels of schooling	E
2003	**18**	**Student outcomes in Asian languages** (Japanese and Indonesian)	Australian Department of Education Science and Training	To raise awareness of what is achievable in school-based Asian languages education	G
2009-2011	**19**	**Student Achievement in Asian Languages Education (SAALE), (**Chinese, Japanese, Indonesian, Korean)	Australian Department of Education Employment & Workplace Relations	To establish baseline for describing student achievement and to consider the contribution of language background and time-on-task to school achievement in Asian languages	G
2011-2013	**20**	**Online placement testing for university language programs** (French, Arabic, German, Italian, Spanish, Russian, Indonesian, Chinese and Japanese)	University of Melbourne	To assess proficiency of incoming undergraduate students for placement in particular course levels	I
2011-2012	**21**	**Language Assessment at the Australian Defence Force School of Languages (DFSL)**	Defence Force School of Languages	To review current assessment practices within the DFSL in relation to a competency-based curriculum reform	E
2015-2016	**22**	**LOTE proficiency screening of candidates seeking accreditation as interpreters and translators**	National Authority for Accreditation of Translators and Interpreters (NAATI)	To review feasibility of available options for preliminary testing of LOTE proficiency prior to sitting the NAATI translation and Interpreting exams	G
2016	**23**	**Determining and implementing language proficiency standards for the Australian interpreter and translator professions**	National Authority for Accreditation of Translators and Interpreters (NAATI)	To set minimum language proficiency standards in a range of languages for interpreting and translating purposes and to review test options for determining attainment of these minimum standards	G

2016	**24**	**LOTE Public Service Language Aide test**	National Authority for Accreditation of Translators and Interpreters (NAATI)	To determine eligibility for language allowances for those needing to communicate at a basic level in LOTE in the workplace	**I**
2019 - 2020	**25**	**Language rating scales mapping**	Defence Force School of Languages, Australian Department of Defence	To align levels on the Australian Defence Force Rating Scale (ADLPRS) with those of other formal language qualifications	G

From each category one project in which I was personally involved has been chosen to reflect on how the Centre's policy contribution might be framed and evaluated. These three cases will be presented in chronological order.

Case One

The first project I will discuss, the Italian Proficiency Test for teachers (the first component of Project 3 in Table 1 above) has been categorised as I (*Implementing* policy), since it involved the development of an assessment tool to be used in the service of overlapping national and state language policies of the time (early 1990s) geared to fostering an early start for studying a second or additional language by expanding program offerings in Australian primary schools. Implementation of these policies was dependent on the availability of well-trained teachers proficient in the languages to be taught (Nicholas et al. 1993). Accordingly, the NLIA secured funding on the Centre's behalf from the federal government's Innovative Languages Other than English in Schools (ILOTES) program to design a prototype test for measuring teacher proficiency in Italian, which at the time was the most widely taught language in primary school[3].

Project aims and implementation

The immediate aim was to assist the implementation of the language learning policy by offering a tool for assessing the communicative competence of trained primary school teachers who had not completed a major study in Italian at the university, but who might, by dint of qualifications gained elsewhere or home exposure to the language as children of Italian immigrants, be sufficiently proficient to qualify as specialist LOTE teachers. It was also hoped that, by modelling the communicative demands of the language teaching situation in the test, we might generate positive washback, signaling to teachers and

[3] Italian was attractive both as a prestige "world language" as well as a language widely used by members of the local Italian community.

teacher trainers the importance of providing an input-rich environment in which the target language was used to perform multiple functions in and outside the classroom. We opted for a task based Languages for Specific Purposes (LSP) approach to test design informed by a scan of the Second Language Acquisition (SLA) literature, classroom observation and insights from classroom teachers about target language use in the professional context (Elder 1994). The speaking sub-test, for example, required test candidates to simulate the teacher role in performing various classroom tasks in Italian such as, reading aloud, setting up a roleplay activity and giving feedback on a piece of written work as if to a class of children.

Discussions around the use of the test for its intended purpose were disappointing. The Italian Department in the School of Languages where the Centre is located had simultaneously secured state funding to run intensive language upgrade programs for Italian teachers in the school break and agreed to offer these teachers as trial participants for the test. However, department lecturers baulked at endorsing our test for official use as a screening tool, believing its content and task types were in no sense equivalent to what was offered in their degree programs. They were also, understandably, anxious about any gatekeeping role for entry to teaching being allocated to an outside agency. Anxious to assuage such sensitivities, the Department of Education ultimately decided to by-pass our testing model and give autonomy to university language departments in determining the communicative readiness to teach of those without a local university language qualification. These departments were asked to undertake their own assessments of applicants' language competence and to issue "Statements of Equivalence" affirming that the applicant's competence was on a par with the exit standard of a 3-year major in the target language. The assessment guidelines offered to them were general in nature[4] and did not relate specifically to classroom language use or to the particular communicative demands of the teaching profession.

Perhaps we were naïve in expecting uptake of our specific purpose assessment model given the push from various quarters to maintain standards of a more traditional kind. It should nevertheless be acknowledged that turf wars between different stakeholders are a normal part of the *rhetoric* layer of policy, which is an 'agitational space' (Lo Bianco & Aliani, 2013) filled with vested interests, both professional and political, competing for influence on policy implementation.

Impact of the project

[4] For a current copy of the guidelines see https://www.education.vic.gov.au/Documents/school/principals/curriculum/Updated_SoE_Guidelines_and_assessment_record.pdf)

Our success in achieving the project's immediate goal to assess and assist in the selection of linguistically proficient classroom teachers the test clearly fell short of expectations - perhaps a function of too little attention paid to politically strategic alignments before embarking on the project. The story of the project did not end there, however. If we take a longer view of the policy impact of the test, it becomes clear that the approach adopted and the publications ensuing from the project generated significant interest among different teaching faculties and other agencies around Australia. They too were grappling with questions of the readiness of teachers to service the LOTE programs funded under federal and state policies. For example, a teacher educator in the state of Victoria was commissioned by a local forum of teacher educators and administrators in the LOTE sector to draw up a set of specifications indicating the communicative functions and linguistic repertoire teachers would need to command for effective performance in the classroom and associated professional domains (Nicholas, 1997). To do so he drew on the findings of the published needs analysis undertaken for our project (Elder, 1994). These proficiency requirements were then communicated to university language departments to assist them in planning and delivering a language curriculum that would equip students with the language skills needed in teaching contexts. We know of at least one department which subsequently offered an elective subject along these lines focussing specifically on classroom language use.

In the meantime, the LTRC sought and received NLIA funding for parallel teacher test projects in Japanese and Indonesian. Although space constraints preclude our mapping the fortunes of these tests, the conceptual work behind their design disseminated through various publications (e.g., Elder, 1993a & b; Elder, 1994; Iwashita & Brown, 2005; Elder 2001; Elder & Kim 2014; Hill, 1996) has informed similar LSP assessments both within and outside Australia. Moreover, the reputation the Centre has built from this work has led to requests for advice by other agencies. One recent request has yielded a new project for the LTRC, commissioned by a body responsible for the professional licensure of teachers. Its brief is to review the suitability of English proficiency tests (such as IELTS, TOEFL and PTE) to determine whether overseas-trained teachers applying to work in Australia have the skills needed to communicate effectively in the school context. Funding comes again from the federal government, now concerned with raising teacher standards across the board as part of a push to arrest Australia's declining PISA outcomes.

In sum, although the direct impact of the Italian teacher proficiency test on the language teachers and teacher educators for whom it was intended (i.e., those forming part of Lo Bianco's *experience* layer of policy mentioned above) may have been slight, the project does seem to have generated significant further thinking, planning and action on teacher language proficiency over the longer term.

Case Two

The second example of an LTRC policy-related testing activity (see Project 11 in Table 1 above), is classified as E (*evaluative*), and dates to the late 1990s when the Department of Education in Victoria, one of the most progressive in Australia with respect to its language policies, allocated funding to initiate or extend the provision of bilingual education programs in several primary and secondary schools. A condition of receiving such funding was that the relevant schools seek assistance from a qualified external consultant in evaluating the effectiveness of their program during the three-year funding period. Pre- and post-testing of proficiency in the target language and in English was stipulated as a central component of the evaluation, along with other forms of investigation such as classroom observation, interviews with teachers and surveys of children and parents.

I will focus here on just one of the four schools which hired the LTRC for this purpose, a school offering a late (Years 7-10) secondary school Arabic-English partial immersion programme for a volunteer cohort of second-generation immigrants (n=65) of Lebanese origin. These students were exposed to a dialectal variety of Arabic at home but had undergone all their prior schooling in English in Australia. The majority had some experience of Modern Standard Arabic (MSA) language instruction in the primary school, although this amounted to only one hour per week on average. The bilingual program they had opted to join offered Arabic-medium instruction in three school subjects (initially maths, social studies, and science) as well as an Arabic language-as-object course of instruction. Taken together the Arabic-medium input amounted to 12 hours per week or around a third of the total teaching time.

Project implementation and outcomes

The testing component of the evaluation involved pre- and post-testing geared to comparing the achievement (both academic and linguistic) of those in the bilingual program with that of a control group of heritage language students in the same high school who were studying Arabic for only three hours per week. Arabic-speaking research assistants studying program evaluation as a subject in the University's Master of Applied Linguistics program at the time of the study were engaged by the LTRC to assist with test development and other aspects of the evaluation.

Outcomes of pre- and post-testing using custom made tests in MSA, the target language, and age-appropriate standardised tests in English and mathematics over a three-year period revealed only a marginally favourable effect for the bilingual program in all areas except for spoken Arabic, where there was, disturbingly, no observable score gain in either the bilingual or the control group (Elder, 2005). We attributed the lack of measurable progress in spoken Arabic to the fact that, while teaching materials developed for the program were in MSA, there was extremely limited MSA input on the part of the bilingual teachers who, contrary to the conditions stipulated for the bilingual experiment, often resorted to speaking dialects (mostly Egyptian) that were barely intelligible to the students as well as to translating Arabic-medium texts into English to aid student comprehension. Classroom observations also revealed that there were limited opportunities for classroom discussion in Arabic, with the prime focus being reading and writing in what was a predominantly teacher-centred style of instruction. Students therefore persisted with their use of a mix of English and their home dialect, which was not measured by the tests in question. We made various recommendations based on these findings, including that only students with some knowledge of MSA be admitted to the program, that teachers desist from using dialects unfamiliar to the students, that more opportunities be offered for MSA spoken input and output and also that students' dialect-influenced deviations from MSA should not be unduly penalised, but instead be treated as a bridge to acquiring the standard form (Elder & Mayer-Attenborough, 2000).

Impact of the project

The impact of the evaluation findings and recommendations was limited by the school's reluctance to take on board the implications of the test results which, in our view, indicated rather limited benefits from the bilingual program. The school had invested enormous effort in mounting the program and was keen to cast the evaluation findings in the best possible light with the parents of the students involved (who had high hopes for the program as a bridge between home and school cultures). They were also anxious to fend off opposition to the program from other teachers who felt the bilingual program was drawing resources away from mainstream curriculum. It was clear that our voice was only one of many at the *experience* level of the policy network and that different participants associated with the school had different views and degrees of investment in the program's outcomes. Also, the school's attitude to our evaluative feedback was no doubt coloured by the LTRC's conflicted role both as shepherd, monitoring and guiding the implementation process, and as inspector, holding the school accountable to the Department of Education which held the purse strings for continuation of the program beyond the immediate 3-year term (Elder, 2009). In the end it was agreed that the final report to the Department of Education would present our own muted conclusions from the test results along with a section indicating the school's more favourable take on the program's achievements and how it intended to respond to our various recommendations. Whether any subsequent adjustments made to the design and delivery of the program were a direct result of our intervention we cannot be sure, given that our association with the school finished when our contract expired.

It is also difficult to discern the longer-term impact of this and other evaluations we conducted of school bilingual programs in Victoria. The LTRC was only one of a group of consultants evaluating such programs throughout the state, each with different kinds of expertise and varying one from another in how they interpreted the evaluation brief and designed their interventions. Although funding continues for many such programs in Victoria, the Department of Education does not release information about individual school outcomes, and consultants hired for the evaluation by the relevant state schools are legally bound to keep their findings confidential. What we do know is that, some years later, the results of our evaluation report (and those of the other evaluation consultants) were used in a Department instigated meta-evaluation of the state-wide bilingual experiment (Jane et al., 2005). While the findings of this meta-evaluation were also confidential, its authors produced a set of publicly available guidelines for best practice in LOTE education based on the lessons learned from the different programs' experience of implementing the policy. These guidelines can be viewed as a refinement of the Department's initial policy intentions with the potential to influence future LOTE learning initiatives. Also worth noting is the fact that the LTRC evaluators, including the postgraduate students hired to assist with the project, gained valuable expertise from their participation in the project. A number of them have gone on to conduct further free-lance consultancies in language and literacy education, contributing in their own right to the shaping of language policy.

Case Three

The third illustrative example of how testing can interface with policy is the Student Achievement in Asian Languages Education Project (SAALE) (Project 19 in Table 1, above) headed by the Research Centre for Languages and Cultures in South Australia in partnership with the LTRC and funded by the Department of Education Employment and Workplace Relations (DEEWR) under the Labour government's National Asian Languages and Studies in Schools Program (NALSSP). The NALSSP, one of a series of initiatives by the Australian government geared to boosting the acquisition of Asian languages by Australian students, had set overly ambitious targets, namely that by the end of their secondary schooling 12% of students nationwide would achieve sufficient fluency in one of 4 priority Asian languages (Chinese, Japanese, Indonesian, and Korean) to engage in trade or commerce in the Asian region. This project (coded G**,** *Guiding* policy formation) was the culmination of a series of testing initiatives over the years (including Projects 9, 14 & 18 in Table 1, above) geared to tempering such targets with evidence, by assessing and describing *actual* school achievement among the highly diverse population of students of Asian languages at 3 different levels of schooling: end of primary school [Year 6/7], mid secondary school [Year 10] and end of secondary school [Year 12]).

Project aim and outcomes

The project set out to investigate the effect of two potentially influential variables on test performance, namely: time-on-task (i.e., frequency, duration, and intensity of prior target language instruction) and language background (e.g., home exposure to the target language, prior study of other languages). Tests of achievement in each of the four Asian languages were developed and administered to a nation-wide sample of students at each year level, who also supplied data on their language background and prior learning experience (Elder et al., 2012).

Quantitative analysis of test results yielded information about salient learner groupings in each language and about the effect of years of study on learning outcomes, revealing, for example, that language background was a much more powerful and consistent predictor of achievement in all languages than language study in the local school context, particularly in the case of Chinese, where a significant proportion of learners were recent immigrants and had substantial experience of mother tongue (Chinese-medium) schooling in their home country (Scarino & Elder, 2012). Test performances were analysed qualitatively by teams of expert teachers from around the country who drew up profiles of typical achievement reflecting the particularities of the languages studied, including their linguistic and pragmatic features, as well as the diverse learning trajectories of enrolled students[5]. Illustrative samples of performance and the associated commentary drew attention to the variability within each learner sub-group and to the importance of taking individual histories into account when planning teaching and monitoring learning. We argued in our project report (Scarino et al., 2011) that policy ambitions about learning outcomes should be informed by what learners of particular languages bring to the task of school language study. There followed a set of published papers emphasising how the project findings differed for each language (Iwashita, 2012; Kim, 2012; Scrimgeour, 2012; Kohler, 2012).

Impact of the project

Ten years on it is difficult to estimate what aspects of this message were received by policy makers in Australia's capital and how or whether the study's findings about the likely outcomes of school-based language programs will inform future language education policy at national or state level. Australia's history of policy and planning in relation to language learning is full of fits and starts with lessons of the past more often ignored than heeded (Lo Bianco & Slaughter, 2009; Lo Bianco & Aliani, 2013). What we can attest to, however, is the direct and significant impact of this project on the LOTE

[5] These descriptions were formulated for 'High' and 'Average' level learners in each of the language background groupings that emerged as distinct in our statistical analysis of test performance.

component of the Australian Curriculum (n.d.) first approved by the Council of Commonwealth State and Territory education ministers in 2009 and rolled out in Australian schools from 2015. The winding up of the SAALE project coincided with the writing of this national curriculum, with the SAALE project leader and several of the expert teachers in the project team directly involved in its creation. This meant that the descriptive profiles of test performance developed for the SAALE project directly informed the writing of the Achievement Standards for these languages in the new curriculum[6]. The structure adopted for these achievement standards moreover became the model for writing other language specific standards in additional languages for which there was no direct evidence base (at least 13 language-specific frameworks as well as a framework for Aboriginal languages, for Auslan[7], and for classical languages) (Scarino, personal communication, 2020).

While it is too soon to gauge the impact of the Australian Curriculum (currently under review) on the processes and outcomes of teaching and learning languages in Australian schools, it is a definite improvement that this curriculum formulates standards in language- and context-specific terms, which acknowledge the diverse nature and range of achievements to be expected of learners with different language backgrounds under different program conditions. Previous state-based outcome statements have tended to treat additional languages taught at school as foreign, on the mistaken assumption that English, the official language of Australian schooling, was the common point of departure for all (Elder, 2014).

Discussion and conclusion

The three case studies outlined above have been somewhat simplified due to space constraints but nevertheless give a sense of the complex ecology of policy and planning with its different layers of intent, rhetoric and experience and the multiple parties involved interpreting, responding to, acting on and in some cases revisiting or resisting policy intentions. Language testing expertise, I have tried to show, can have a key role within the policy contexts described, serving different purposes, including to inform (Case Three), to evaluate (Case Two) and to implement (Case One) policy intentions whether this be at the level of the nation, the state, or a particular institution. But language testing activities are just tiny cogs in the larger machinery of policy and planning as Deygers (2021) has emphasised. The impact of our incursions is inevitably diluted by other inputs and constraints and therefore difficult to gauge. Among the complicating factors are the competing agendas, vested interests, and sensitivities of those involved in

[6] https://www.australiancurriculum.edu.au/f-10-curriculum/languages/

[7] Auslan is the majority sign language for the Australian deaf community.

policy implementation (especially in Case One), the confused communication channels and power imbalances between different administering bodies and those on the receiving end of policy initiatives (with respect to Case Two).

The case studies have highlighted, more generally, the inchoate nature of policy impact which, as we have seen, may be direct or indirect, can emerge over different time frames, and take unpredictable forms. The most immediate and powerful instance of instrumental impact was observable for Case Three, where serendipitous timing, careful planning, and the proximity of the chief investigator to national networks of authority meant that the project outcomes could provide an evidence base and a structure for a powerful *de facto* policy document, the Australian curriculum (n.d.), the first in Australia's history. This new curriculum can be seen as a distillation of the earlier NASSLP policy, a new statement of intent with the potential to guide language teaching and learning expectations and behaviours throughout the country for the foreseeable future. This influence may well be enhanced by input from the network of expert teachers engaged for the SAALE project, who are now well placed to work for its enactment in schools across the nation and who may well be involved in subsequent cycles of curriculum review. Case Two, as we noted, may well have influenced instrumental decisions about program development at the school level, or indeed about continuing funding for the program itself, but the terms of our consultancy constrained our role with respect to both the school and the government department to whom we were also accountable. Estimating the impact of the LTRC's involvement from the perspective of both these parties would be valuable, but too much time has elapsed for this to be feasible. From our perspective the impact of this intervention was primarily educational, building expertise in program evaluation for those involved and, in the longer term, contributing (indirectly) to a set of principles for effective LOTE teaching devised by subsequent consultants which in turn fed into future educational policy. The impact of Case One was likewise more conceptual than instrumental and took some years to become visible via publications on the subject. Citations from these publications suggest they have influenced the constructs of additional tests of teacher proficiency both at home and abroad and have helped to build the Centre's reputation in the teacher language proficiency domain.

If it is hard to estimate the impact of individual projects, it is even harder to evaluate the overall impact of the Centre's work in the LOTE arena, as represented by the projects listed in Table 1 above. Not all of these have resulted in publications, and many of the reports they have generated are stored in old filing cabinets and unlikely to ever see the light of day. The cumulative experience of conducting these projects has nevertheless built a notable body of expertise and the mere fact of the LTRC's survival as a self-funding centre shows that this expertise continues to be valued. More importantly, there have been lessons learned from these projects that can inform future relationships with policy makers.

As stated in the literature review above, discussions of test impact have tended to focus how stakeholders, such as teachers and learners or institutional administrators, are affected by a test in use rather than on the broader policy implications of testing activity for policy formation, implementation, or review. Seeing testing activity as part of a policy web (Goldberg, 2006) should make us more conscious of our duty as *policy responsible* professionals to assist decision-makers in navigating the complex challenges they face in bringing about change, while also highlighting the limits of our power. The insights gleaned from the three examples and other LTRC projects over the years are listed as a set of desiderata below.

1. ***Take stock of the policy context***
   Taking stock of the policy context should surely be a starting point in any testing activity. This was certainly the case with Case Three, where generous funding allowed us to clarify the intentions, values, and expectations of key policy players from different states and school sectors who were flown from across the country to serve on the project steering committee. By contrast the failure in Project One to fully acquaint ourselves with policy settings in different jurisdictions around Australia may have limited the uptake of our teacher proficiency test in the immediate aftermath of the project.

2. ***Anticipate policy affordances and consequences***
   It would also seem important to anticipate the policy affordances and consequences of a testing activity even where these are not specified in the project brief. Chalhoub-Deville (2009) has argued along these lines in her discussion of the role of Social Impact Assessment (SIA) in implementing large-scale educational reform in the United States. SIA, she claims, can improve understandings of past developments and predictions of future change. For Case Two, in hindsight, it would have been useful to seek information from the Department of Education which funded the evaluation component of each bilingual program on the primary goal of the evaluation and the use they intended to make of the evaluation reports – whether these were intended for monitoring of individual school performance for accountability purposes, or primarily formative, in the sense of assisting the Department to refine its overall bilingual policy. Advance understanding of these policy intentions would have guided the evaluation process, helping us to refine the foci of our investigation and to better manage our relationship with the evaluand. For Case Three, we conceived of the potential links between the test outcomes (i.e., the descriptive profiles of performance generated for diverse

learners at various levels of schooling) and the structure and content of the new national curriculum from early on, and this certainly informed the project design and enhanced its impact.

3. ***Plan to maximise policy impact***
   Maximising policy impact should be front and centre in project planning. As happened with Case One, impact may be enhanced via dissemination of project findings in reports and academic papers. However, the readership of academic outlets will always be limited. Identifying key influencers of policy and drawing up communication plans and tactics tailored to educating policy makers and influencing their agendas should be a central part of our project design. The program evaluation literature, sadly neglected in our field, offers insights on ways to foster the utilization of project findings (e.g., Patton, 2003, Owen 2006). So too does the growing literature on science communication oriented to building stakeholder understanding (Pill & Harding, 2019) and to bridging the divide between evidence and policy (e.g., Cairney & Kwiatkowski, 2017; Boswell & Smith, 2017), a challenge which universities are now taking seriously (e.g., Smith & Stewart 2017).

4. ***Build evaluation of policy impact into project planning***
   While, as has been demonstrated, policy impact is seldom immediate, visible, or predictable, we should nevertheless make conscious efforts to monitor evidence of policy uptake, whether instrumental or conceptual (Nutley et al., 2007), at any level of the policy web during the life of a project and beyond. The metrics we might use for such an exercise are various, ranging from lists of deliverables, to documented exchanges with clients and other stakeholders, to counts of citations in policy documents and academic outlets, and to other tangible evidence of response to policy recommendations. Defining what would constitute impact for a given project and including mechanisms for evaluating it in our contract with the client might go some way to alleviating the difficulties faced in relation to Case Two, where commissioned reports and their recommendations on the bilingual education programmes have remained hidden from the public eye and decisions on the bilingual education initiative are made behind closed doors.

In conclusion, while we can seldom claim direct causal relationships between our testing interventions and policy outcomes, given our limited influence as just one of the many players within the multiple strands of the policy web, documenting the policy aims and diffuse uptake of our projects remains important. It allows us, as language testing practitioners, to reflect on past practices, temper our expectations and refine our future interventions to render them as effective as possible. This kind of self-monitoring will surely help to build a 'professional milieu' (Davies, 1997) and to render ourselves publicly accountable. The experience of the Language Testing Research Centre (LTRC) accrued over the 30 years of its operation, however, invites a broader conceptualisation of professional accountability beyond that defined by Bachman and Palmer (2010) as "being able to demonstrate to stakeholders that the intended uses or our assessments are justified" (p. 92). Such a formulation positions language testers merely as defending the validity of their instruments, rather than as experts contributing to knowledge exchange in the larger policy arena. Accountability for language testing, if it is to consider itself a mature field, should surely be more outward-looking. It should not be just about justification of intentions but also about demonstrating responsiveness, within the limits of our expertise and influence, to the needs of stakeholders at all levels of the policy web in a manner conducive to informed formulation, enactment, and evaluation of policy agendas.

References

Australian Curriculum (n.d.). *Languages.* https://www.australiancurriculum.edu.au/f-10-curriculum/languages/

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

Ball, S. J. (1993). What is policy? Texts, trajectories and toolboxes. *Discourse 13*(2), 10–17. https://doi.org/10.1080/0159630930130203

Boswell, C., & Smith, K. (2017). Rethinking policy "impact": Four models of research-policy relations. *Palgrave Communications 3,* Article 44. https://doi.org/10.1057/s41599-017-0042-z

Bunch, M. B. (2011). Testing English language learners under No Child Left Behind. *Language Testing 28*(3), 323–341. https://doi.org/10.1177/0265532211404186

Cairney, P., & Kwaitowski, R. (2017). How to communicate effectively with policy makers: Combine insights from psychology and policy studies. *Palgrave Communications 3*, Article 37. https://doi.org/10.1057/s41599-017-0046-8

Chalhoub-Deville, M. (2009). The intersection of test impact, validation, and educational reform policy. *Annual Review of Applied Linguistics, 29,* 118–131. https://doi.org/10.1017/S0267190509090102

Davies, A. (1997). Demands of being professional in language testing. *Language Testing, 14*(30), 328–329. https://doi.org/10.1177/026553229701400309

Deygers, B. (2021, April). *Language testing ethics, and our place in the ecosystem* [Plenary presentation]. The 1st ALTE International Digital Conference.

Deygers, B., & Malone, M. E. (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*, *36*(3), 347–368. https://doi.org/10.1177/0265532219826390

Dunlop, C. A., & Radaelli, C. M. (2018). The lessons of policy learning: Types, triggers, hindrances and pathologies. *Policy & Politics*, *46*(2), 255–272. https://doi.org/10.1332/030557318X15230059735521

Dwyer, M. (2002). On New York's assessment policy: A perspective from the field. Teachers College, *Studies in Applied Linguistics & TESOL*, *2*(1), 1–8. https://doi.org/10.7916/salt.v2i1.1648

Elder, C. (1993a). *The proficiency test for language teachers: Italian. Volume 1: Final report on the test development process* [National project report for Department of Employment, Education and Training]. Language Testing Research Centre, University of Melbourne.

Elder, C. (1993b). *The proficiency test for language teachers: Italian. Volume 2: Appendices* [National project report for Department of Employment, Education and Training]. Language Testing Research Centre, University of Melbourne.

Elder, C. (1994). Performance testing as benchmark for foreign language teacher education. *Babel: Journal of the Federation of Modern Language Teachers Associations*, *29* (2), 9–19. (Also in *Melbourne Papers in Language Testing*, *3*[1], 1–26).

Elder C. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing*, *18*(1), 149–170. https://doi.org/10.1177/026553220101800203

Elder, C. (2005). Evaluating the effectiveness of heritage language education. What role for testing? *International Journal of Bilingualism and Bilingual Education*, *8*(2&3), 198–212. https://doi.org/10.1080/13670050508668607

Elder, C. (2009). Reconciling accountability and development needs in heritage language education: A challenge for the evaluation consultant. *Language Teaching Research*, *13*(1), 15–34. https://doi.org/10.1177/1362168808095521

Elder, C. (2014). Reflecting the diversity of learner achievements in first and additional languages: Towards context-specific language standards. In B. Spolsky, O. Inbar-Lourie, & M. Tannenbaum (Eds.), *Challenges for language education and policy: Making space for people* (pp. 52–64). Routledge.

Elder, C., & Kim, S-H. (2014). *Assessing teachers' language proficiency.* In A. Kunnan. (Ed.), *The companion to language assessment* (Vol. 1, pp. 1–17). John Wiley & Sons. https://doi.org/10.1002/9781118411360.wbcla138

Elder, C., Kim, H., & Knoch, U. (2012). Documenting the diversity of learner achievements using common measures. *Australian Review of Applied Linguistics*, *35*(3), 251–270. https://doi.org/10.1075/aral.35.3.02eld

Elder, C., & Mayer Attenborough, C. (2000). *'Seaview' secondary college Arabic-English bilingual program* [External evaluator's report]. Language Testing Research Centre, University of Melbourne.

Gale, T. (1999). Policy trajectories: Treading the discursive path of policy analysis. *Policy Trajectories, 20*(3), 393–407. https://doi.org/10.1080/0159630990200304

Ginther, A., & Yan, X. (2017). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing, 35*(2), 271–295. https://doi.org/10.1177/0265532217704010

Goldberg, M. P. (2006). Discursive policy webs in a globalisation era: A discussion of access to professions and trades for immigrant professionals in Ontario, Canada. *Globalisation, Societies and Education, 4*(1), 77–102. https://doi.org/10.1080/14767720600555103

Hill, K. (1996). *The Asian languages proficiency project: Indonesian* [Final report prepared for the Department of Employment, Education, Training and Youth Affairs]. Language Testing Research Centre, University of Melbourne.

Iwashita, N. (2012) Cross-linguistic influence as a factor in the written and oral production of school-age learners of Japanese in Australia. *Australian Review of Applied Linguistics, 35*(3), 290–311. https://doi.org/10.1075/aral.35.3.04iwa

Iwashita, N., & Brown, A. (2005). *LOTE proficiency test for teachers: Japanese* [Final report]. Language Testing Research Centre, University of Melbourne.

Jane, G., Griffiths, B., Russell, J., & Nicholas, H. (2005). *Review of the implementation and outcomes of the designated bilingual programs in primary and secondary schools* [Confidential final report submitted to the Victorian Department of Education and Training].

Johnson, D. C., & Johnson, E. J. (2014). Power and agency in language policy appropriation. *Language Policy, 14*(3), 221–243. https://doi.org/10.1007/s10993-014-9333-z

Kim, S-H. (2012). Learner background and the acquisition of discourse features of Korean in the Australian secondary school context. *Australian Review of Applied Linguistics, 35*(3), 339–358. https://doi.org/10.1075/aral.35.3.06kim

Knoch, U., & Macqueen, S. (2020). *Assessing English for professional purposes.* Routledge. https://doi.org/10.4324/9780429340383

Kohler, M. (2012). How does *time-on-task* affect the achievement of early and late starters in Indonesian schools? *Australian Review of Applied Linguistics 35*(3), 271–289. https://doi.org/10.1075/aral.35.3.03koh

Lo Bianco, J. (1987). *National policy on languages.* Australian Government Publishing Service.

Lo Bianco, J. (2001). Policy literacy. *Language and Education*, 15(2&3), 212–227. https://doi.org/10.1080/09500780108666811

Lo Bianco, J. (2019). Talking to the Pollies: Academic researchers and public officials. In C. Roever & G. Wigglesworth (Eds.), *Social perspectives on language testing: Papers in honour of Tim McNamara* (pp. 89–114). Peter Lang.

Lo Bianco, J., & Aliani, R. (2013). *Language planning and student experiences: Intention, rhetoric and implementation.* Multilingual Matters. https://doi.org/10.21832/9781783090051

Lo Bianco, J., & Slaughter, Y. (2009). *Second languages and Australian schooling* (Australian Education Review No. 54). Australian Council of Educational Research.

McNamara, T. (2009). Language tests and social policy: A commentary. In G. Hogan-Brun, C. Mar-Molinero, & P. Stevenson (Eds.), *Discourses on language and integration: Critical perspectives on language testing regimes in Europe* (pp. 153–164). John Benjamins. https://doi.org/10.1075/dapsac.33.12nam

McNamara, T. (2011). Managing learning: Authority and language assessment. *Language Teaching, 44*(4), 500–515. https://doi.org/10.1017/S0261444811000073

Menken, K. (2008). High-stakes tests as de facto language education policies. In E. Shohamy & N.H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Volume 7: Language testing and assessment, pp. 401–413). Springer. https://doi.org/10.1007/978-3-319-02261-1_25

Nicholas, H., Moore, H., Clyne, M., & Pauwels, A. (Eds.) (1993). *Languages at the crossroads: The report of the national enquiry into the employment and supply of teachers of languages other than English.* National Language and Literacy Institute of Australia.

Nicholas, H. (1997). Language proficiency outcomes for intending teachers [Report prepared for the Victorian LOTE proficiency forum].

Nutley, S. M., Walter, I., & Davies, H. T. (2007). *Using evidence: How research can inform public services*. Policy Press.

Owen, J. M. (2006). *Program evaluation: Forms and approaches* (3rd ed). Routledge. https://doi.org/10.4324/9781003116875

Patton, M. (2003). Utilization-focussed evaluation. In K.T. & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (Vol. 9, pp. 223–244). Springer. https://doi.org/10.1007/978-94-010-0309-4_16

Pill, J., & Harding, L. W. (2019). A most engaging scholar: Tim McNamara and the role of language testing expertise. In C. Roever, & G. Wigglesworth (Eds.), *Social perspectives on language testing: Papers in honour of Tim McNamara* (pp. 217–228). Peter Lang.

Scarino, A., & Elder, C. (Eds.). (2012). Describing school achievement in Asian languages for diverse learner groups [Special issue]. *Australian Review of Applied Linguistics, 35(3*).

Scarino, A., Elder, C., Iwashita, N., Kim, S. H. O., Kohler, M., & Scrimgeour, A. (2011). *Student achievement in Asian languages education* [Part 1: Project report]. Department of Education, Employment and Workplace Relations.

Scrimgeour, A. (2012). Understanding the nature of performance: The influence of learner background on school-age learner achievement in Chinese. *Australian Review of Applied Linguistics, 35*(3), 312–338. https://doi.org/10.1075/aral.35.3.05scr

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Routledge.

Shohamy, E. (2003). Implications of language education policies for language study in schools and universities. *The Modern Language Journal, 88*(2), 277–286. https://www.jstor.org/stable/1193038

Shohamy, E. (2008). Language policy and language assessment: The relationship. *Current Issues in Language Planning, 9*(3), 363–373. https://doi.org/10.1080/14664200802139604

Slaughter, Y., Lo Bianco, J., Aliani, R., & Hajek, J. (2019). Language programming in rural and regional Victoria: Making space for local viewpoints in policy development. *Australian Review of Applied Linguistics* 42(3), 274–300. https://doi.org/10.1075/aral.18030.sla

Smith, K. E,. & Stewart, E. (2017). We need to talk about impact: Why social policy academics need to engage with the UK's research impact agenda. *Journal of Social Policy, 46*(1), 109–127. https://doi.org/10.1017/S0047279416000283

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford University Press.

Spolsky, B. (2001). Cheating language tests can be dangerous. In C. Elder, A. Brown, E. Grove, K. Hill, T. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 212–221). Cambridge University Press.

Stansfield, C., Rivera, C., Scialdone, L., & Sharkey, M. (2000). *An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during 1998-1999* [Final report]. The Center for Equity and Excellence in Education of George Washington University.

Thomas, P. (2007). The challenges of governance, leadership and accountability in the public services. In M. Wallace, M. Fertig, & E. Schneller (Eds), *Managing change in the public services* (pp. 116–135). Wiley-Blackwell.