# Introduction to Special Issue

Catherine Elder (Guest Editor)
*University of Melbourne*

Although recent frameworks for test validation have emphasized utilization as a key component (e.g. Bachman & Palmer, 2010; Chapelle, 2012; Kane, 1992), what happens in the sometimes messy contexts where assessments are carried out is often not documented. This is perhaps because test developers are more focused on defining and operationalizing test constructs than on issues of implementation. It may also be the case that those who actually use assessment tools either lack the relevant expertise to evaluate them or are simply too busy with day-to-day practical matters to reflect and report on how these assessment tasks or systems are functioning in the context of concern.

This special issue therefore focuses attention on issues surrounding the use of language assessments. A collection of six articles offers critical reflections on how different parties interact with or make use of tools designed for a particular purpose and context, and describes what happens to a test, examination or set of qualifications when system requirements or external circumstances change. At the end of the issue, a review by Gruba of *Talking about Assessment* (Kunnan, 2015) reminds us of the importance attached to issues of test use, test consequences and professional accountability by key scholars in our field.

The term *evaluation* in the title is used advisedly in preference to *validation* because, as Norris (2016) proposes, an evaluation perspective forces us to consider assessments not just as instruments functioning in isolation from their context but as programs enacted by a range of users to fulfill certain needs and to produce particular outcomes and consequences. A program evaluation focus also draws attention to the evaluator as potential agent of change, a role that is not usually emphasized in test validation research.

**Objects and contexts of evaluation**

The object of the evaluation studies reported in this issue varies widely as do the contexts in which they occur: from a nation-wide end-of-school foreign language examination system in Austria (Spöttl, Kremmel, Holzknecht & Alderson) to a suite of interconnected English qualifications in New Zealand (Read); from low-stakes self-evaluation tools for practising language teachers in Auckland (Erlam) to a high-stakes standardized English test used for admission to a private university in Beirut (Pill); from the rater-training component of a testing system developed by the British Council for use in multiple locations

inside and outside Europe (Knoch, Fairbairn & Huisman) to the course assessment components of a pathway program designed by a private provider to prepare students for entry to an Australian university (Macqueen, O'Hagan & Hughes). In their reports the evaluators provide practical discussions of the backgrounds, contexts, stakeholders and methods adopted, thereby revealing the diverse forms which evaluations of test use may take and the complex issues and decisions involved.

**Level of evaluation**

Gruba, Cárdenas-Claros**,** Suvorov **&** Rick **(2**016) have usefully distinguished three levels at which an evaluation can be carried out: the macro- societal or organizational level where larger policy decisions are made (particularly relevant to Spöttl *et al.*'s reform initiative, to Read's examination of nation-wide certification standards and to the university entrance requirements which drive Macqueen *et al.*'s investigation); the meso-level of the program where particular work cultures and divisions of labour come into play in implementing an assessment initiative (somewhat evident in the accounts of all authors) and the micro-level where assessments are enacted by teachers and assessors (central to the data gathered by Erlam, Knoch *et al*. and Macqueen *et al.).* While the boundaries between these levels are not always clear-cut, it is helpful to consider how they might affect the roles of evaluators and the evaluative stances they adopt.

**Roles of evaluators**

Those conducting the evaluations (the authors of the papers in this issue) are at various degrees of remove from the object of their evaluation. Macqueen *et al.* and Knoch *et al.* are external evaluators, although they work in close collaboration with program insiders, some of whom have joined forces with the evaluators as co-authors. Read's status is that of an assessment expert with long-term involvement with the English language certification systems he is reviewing. His reflections are not however part of a formal evaluation and may therefore be less constrained by practical and political considerations than might otherwise be the case. Erlam is a little closer to the object of evaluation than Read. As Academic Director of the Teacher Professional Development Languages program in Auckland she has an official advisory role, but her evaluative report is, like that of Read, self-instigated rather than commissioned by the educational authority that has designed and implemented the program. Pill straddles both insider and outsider roles. He works at the university where his investigation is conducted, but as a newcomer charged with the task of reviewing and revising an existing admissions testing system he is positioned somewhat differently from his colleagues who have been operating within that

system for some time. Spöttl *et al.* embark on their review of the traditional examination system as outsiders, but then come to assume an insider role as designers of the new foreign language examinations. However as language testers they inevitably have a limited view of the larger reform picture and are restricted in their power over the entire implementation process. Collectively these studies reveal the complex interactions between evaluators and other stakeholders and point to the importance of managing relationships in any assessment reform or evaluation initiative (Elder, 2009).

**Evaluative stance**

The different relationship of each evaluator or team of evaluators with the evaluand is evident in the way each study is framed. Knoch *et al.,* as externally commissioned evaluators, clearly set out the terms of reference for their evaluation. In collaboration with the Aptis team they draw up a set of criteria against which the efficacy of the new online rater-training program can be determined. Macqueen *et al.* outline the aims of their project using evaluative terms such as "adequacy" (of standards) and "suitability" (of final course assessments), which clearly signal their brief to judge the value of the pathway program in relation to the required exit standard for such courses. Pill's account references a range of studies on high-stakes admissions testing and describes the shifting conceptions of language and language assessment that have occurred over recent decades. These serve as the basis for critiquing the existing test and proposing new directions. His recommendations, like those of Knoch *et al.* and Macqueen *et al*., are intended as the basis for improving future assessment policy and practice at his university.

Read's commentary on the process of benchmarking New Zealand Certificates of English Language (NZCEL) levels by reference to national and international frameworks is more descriptive, offering a language tester's take on the complex issues of equivalence that have emerged during this process in the interests of informing an imminent review of the new certification scheme rather than evaluating the scheme directly. Erlam frames her study as a validity investigation, invoking Weir's (2005) framework and drawing on multiple sources of evidence to consider the theoretical, contextual and consequential validity of the classroom language teacher evaluation tools described in her paper. Her chief aim however is to evaluate the capacity of these tools and the training program more generally, to bring about positive change in language teacher practice. Spöttl *et al.* reflect on their experience of introducing a national examination reform with the benefit of hindsight, describing the challenges encountered in the process in order to "raise awareness of issues that language testers are often inexperienced in dealing with" (p. 3).

**Evaluation constraints**

Most authors comment on the constraints encountered in the evaluation process, including contextual issues such as relationships with stakeholders and the problem of securing participation in evaluation research as well as time and funding limitations, which can influence both the methods adopted and the uptake of findings or recommendations.

Pill mentions the problem of taken-for-granted assumptions made by various stakeholders, which must be unpacked and interrogated by a researcher who is new to the site for evaluation. Limited assessment literacy is presented as an obstacle by Spöttl *et al.*, who faced the task of educating the Austrian authorities about why the construct of communicative competence requires different testing solutions from those proposed in other subject areas. They also mention adverse reactions from teachers to the proposed reform and highlight the political awareness needed by language testers if they are to successfully 'sell' their proposals to the relevant authorities.

Erlam notes the problems of recruiting participants for her investigation. Only one teacher volunteered for her study, perhaps due to a concern that the purpose of Erlam's enquiry was to evaluate the teachers themselves rather than to consider the value of the self-evaluation tools that had been devised as part of the professional development program. As a result her evaluation is somewhat reliant on anecdotal evidence from parties who may have a vested interest in claiming success for the program.

Timing was an issue for Knoch *et al.*'s evaluation, undertaken at various points during the Aptis group's process of designing and implementing a new online rater training program. Part of Knoch *et al.*'s study was conducted prior to a change in the delivery system for this training module with the result that some of their suggestions for refinement of the program may not be entirely relevant to the new online format.

Macqueen *et al.* mention the practical considerations leading them to settle for the locally developed Diagnostic English Language Assessment (DELA) (rather than IELTS) as the criterion for evaluating the extent to which scores derived from the in-house assessments used in the pathway program were in line with the required standard for university entry.

Managing such constraints are, however, part and parcel of the evaluation process. As Macqueen *et al.* put it:

… evaluators are constrained by multiple forces, including but not limited to practical concerns such as time, staffing and project financing, on the one hand, and, on the other, the demands of their own discipline including the need to provide warrants for their claims about the program in the form of accepted research tools, processes and methods (p. 121).

**Utilization**

The need to balance these different forces and make judicious compromises may lead to a perception that evaluation studies lack rigour. However this is to misconstrue the fundamentally utilitarian (rather than theory-driven) role of evaluation, the value of which depends largely on how well it aids understandings of the goals and workings of a program and the extent to which these understandings are put to use (Patton, 2008).

This raises the question of who might be the users, audiences or beneficiaries of the evaluation findings or outcomes reported here. We can list a range of potential users: educational administrators, policy makers and/or receiving institutions; the testing agency; university lecturers and language teachers; course designers, language assessors and item writers and, in all cases, language learners who stand to benefit one way or another from enhanced assessment systems. Although not all may profit directly or immediately from what emerges from the studies collected here (this will depend partly on the communication efforts of the evaluators and on numerous other contextual factors), we would hope that the studies' outcomes will serve to expand the language assessment literacy of key players in the relevant contexts. A better understanding of language assessment among users would contribute to improving the quality of locally enacted assessment policies and practices and, ultimately, to enhancing the validity of assessment-related decisions.

Norris (2009) also mentions fellow evaluators as a potential audience for evaluation studies and emphasizes the important educative purpose of evaluation research. This aligns with the intent of this special issue: on the one hand to highlight the opportunities afforded by evaluation pursuits and, on the other, to lay bare for future investigators some of the complexities and challenges of evaluating and/or reforming language assessment programs and systems in use. We hope that this collection of papers will encourage other language assessment researchers and practitioners to bring new information to light about their experiences of conducting evaluations in different assessment contexts, including the difficulties encountered and lessons learned in the process.

## References

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Chapelle, C. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 19-30). Abingdon: Routledge.

Elder, C. (2009) Reconciling accountability and development needs in heritage language education: A communication challenge for the evaluation consultant. *Language Teaching Research, 13*(1), 15-33.

Gruba, P., Cárdenas-Claros, M.S., Suvorov, R., Rick, K. (2016). *Blended language program evaluation*. London: Palgrave Macmillan.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527-53.

Kiely, R. & Rea-Dickins, P. (2005). *Program evaluation in language education.* New York: Palgrave Macmillan.

Kunnan, A. (Ed.) (2015). *Talking about language assessment: The LAQ interviews*. New York: Routledge.

Norris, J. (2009). Understanding and improving language education through program evaluation: Introduction to the special issue. *Language Teaching Research, 13*(1), 7-13.

Norris, J. (2016). Language program evaluation. *Modern Language Journal*, *100*(Issue SI), 169-189.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage Publications, Inc.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* New York, NY: Palgrave Macmillan.