

Language Program Evaluation

JOHN M. NORRIS

Georgetown University

Department of Linguistics

421 37th St. NW, Box 571051

Washington, DC 20057-1051

Email: norrisj@georgetown.edu

Language program evaluation is a pragmatic mode of inquiry that illuminates the complex nature of language-related interventions of various kinds, the factors that foster or constrain them, and the consequences that ensue. Program evaluation enables a variety of evidence-based decisions and actions, from designing programs and implementing practices to judging effectiveness and improving outcomes. Though often narrowly construed as an external accountability requirement only, evaluation may provide a heuristic for generating new knowledge; raising awareness; and transforming the educational, social, and economic circumstances of individuals and communities. In this article, I review the origins and interdisciplinary nature of program evaluation, and I trace its development within applied linguistics and language education in particular. I then consider the current contributions of language program evaluation, and I point to likely future trends. In particular, I highlight how changing global circumstances, technological affordances, and contexts and purposes for language learning and language use are both impacting the nature of evaluation and presenting challenges for which evaluation is uniquely suited to respond. I also emphasize the potential—versus as yet realized—contribution of program evaluation as a way of ameliorating gaps in current language research and practice, and I point to implications for training, dissemination, and integration of this mode of inquiry as an important part of the future intellectual core of applied linguistics.

Keywords: program evaluation; accountability testing; outcomes assessment; curricular innovation; formative, participatory

LANGUAGE TEACHING AND LEARNING, AS well as other educational and social endeavors related to language learners and users, play out not within sanitized laboratories where theories are carefully tested, but rather under the realities of geopolitical and economic forces; governmental budgets and policies; institutional affordances and constraints; and the everyday actions of administrators, teachers, learners, and others. Indeed, language programs—intentional interventions designed to address one or more clearly defined societal needs related to language learning and/or use—present a persistently evolving constellation of factors that interact to determine often life-changing outcomes for individuals. Accordingly, teachers, curriculum developers, policy

makers, and the public want to know ‘what works’ in pursuing educational or social good, yet most research on language teaching and learning does not reflect the multifaceted realities of educational or related program contexts, nor does it produce findings in a form readily available for application to decisions that need to be made. Language program evaluation provides a framework, heuristic, and methodologies for making sense of language teaching and learning (and other language-related endeavors) *in situ*, as well as for answering ‘what works’ questions and informing practical efforts at improvement. To date, however, program evaluation largely has been ignored by the mainstream of applied linguistics, and as a result the capacity of evaluation to transform how we inquire, reason, and act in relation to language programs is yet to be realized.

One challenge for understanding the potential contribution of program evaluation within applied linguistics relates to confusion regarding

the term 'evaluation.' While it is a certainty that different discourse communities utilize terms like evaluation, assessment, and measurement interchangeably, it is also clear that evaluation has developed over the course of the 20th century to imply a particular set of values, purposes, and practices that represent and guide the endeavors of a body of scholars and professionals (American Evaluation Association [AEA], 2004). According to Norris (2006), in the *MLJ Perspectives*:

Evaluation is the gathering of information about any of the variety of elements that constitute educational programs, for a variety of purposes that primarily include understanding, demonstrating, improving, and judging program value. Evaluation brings evidence to bear on the problems of programs, but the nature of that evidence is not restricted to one particular methodology. (p. 579)

Evaluation is thus distinct from assessment (gathering of information about individual learning, knowledge, and ability) and measurement (quantifiable indicators of theoretical constructs) primarily in its scope, purpose, and inherently multimethodological epistemology. Evaluation is also distinct from research, in which inquiry is driven fundamentally by the need to generate or test *theoretical* claims (whatever ontological persuasion the theory may happen to present); evaluation, by contrast, pursues empirical inquiry for the pragmatic ends of making decisions and informing actions (Cronbach & Suppes, 1969). This basic distinction is readily apparent in the foremost professional standard of evaluators, where the "utility" of evaluation is "intended to increase the extent to which program stakeholders find evaluation processes and products valuable in meeting their needs" (Yarbrough et al., 2011, p. 3). The value of evaluation, then, is determined in large part by the extent to which it helps specific people understand how and to what ends programs are functioning, and by the degree to which that understanding is put to use (Patton, 2008).

A related challenge is the common misconception that evaluation is monomethodological (i.e., the positivistic measurement of program outcomes in quantifiable terms); bureaucratic (i.e., driven by top-down regulatory policies or mechanisms); and focused only on holding teachers, learners, employees, and others within programs accountable to a financial or 'minimum competency' bottom line. While it is true that program evaluation grew out of a narrowly experimental orientation in the Great Society era (e.g., Campbell, Stanley, & Gage, 1963), evaluation scholars quickly realized that methodological

pluralism was essential if evaluation was to meet the actual decision-making needs of actors within educational and social programs. Thus, for example, an early distinction between formative and summative evaluation (Scriven, 1967) emerged from the observation that evaluation was not particularly useful if it only provided information for making yes/no, fund/no-fund decisions—underperforming programs, like many schools which could not simply be shut down, required richer information that would lead to improvements in practices, not merely the 'finding' that they were underperforming in experimental comparisons with other schools. Along with multiple purposes for evaluation, then, came the demand for multiple methodologies and ways of observing and interpreting; different questions about programs called for different types of data. Thus, while the social sciences (including applied linguistics) have sustained debilitating paradigm wars and associated epistemological elitism to date, evaluators quickly got over the qualitative–quantitative, experimental–interpretive divide (though not without aggressive resistance, e.g., Lincoln & Guba, 1985), opting instead for a resolution in pragmatism. According to Cronbach and colleagues (1980), "The evaluator will be wise not to declare allegiance to either a quantitative–scientific–summative methodology or a qualitative–naturalistic–descriptive methodology" (p. 7).

Contemporary program evaluation practice, then, embraces multiple purposes, methodological pluralism, and a pragmatic focus on usefulness, and evaluators have repudiated policies and practices that threaten these foundations. For example, the American Evaluation Association (AEA) has spoken out against the prioritization of randomized control trials over other scientific methods by the U.S. Department of Education (AEA, 2003), against the exclusive use of standardized testing for educational accountability (AEA, 2006), and against the near-exclusive reliance on outcomes measurement and experimental designs for the evaluation of federally funded programs (AEA, 2013). The bottom line for professional evaluators, then, is that systematic data collection and analysis is undertaken in response to specific questions that reflect a wide range of possible uses (and users) of information in understanding, judging, and improving programs. Indeed, evaluation may be done to identify needs, clarify values, allocate resources, design curricula, develop materials, improve delivery, demonstrate outcomes, raise awareness, and transform the lives of program stakeholders,

among many other purposes (Patton, 2008; Rossi, Lipsey, & Freeman, 2004).

Like mainstream evaluation, language program evaluation emerged during the Great Society era, and it has evolved through several phases of practice and scholarship.¹ Norris and Watanabe (2012) characterize three stages in the history of language program evaluation to date. In an initial, preschoolly stage, beginning in the post-World War II era, language program evaluation took the form of large-scale experimental studies that sought to demonstrate the superiority of one language teaching method over another (e.g., Scherer & Wertheimer, 1964; Smith, 1970), or it involved the external monitoring of overseas language programs by language education experts representing funding organizations from Western nations (e.g., Beretta & Davies, 1985). Reactions to the weaknesses of program-level experimentation and bureaucratic oversight, along with the rapid rise in English language teaching worldwide, spawned the next stage, which might be considered the birth of a scholarship of language program evaluation. Reflected most cogently by Beretta's observations on the realities and possibilities of evaluation practice (e.g., Beretta, 1986a, 1986b, 1986c, 1990a, 1990b, 1992a, 1992b), language program evaluators carved out a new domain of applied linguistics work based on several core insights, as summarized by Norris and Watanabe (2012):

- (1) Truly experimental research on language programs or teaching methods was ill advised, due to lack of control over numerous factors (population samples, teachers, settings, etc.) and the consequent inevitability of misinterpreting findings.
- (2) External/expert reviews provided only one kind of subjective, often arbitrary, information about program functioning and outcomes, and led frequently to negative reactions by program insiders.
- (3) Understanding of language program implementation was essential for interpretations about program effectiveness; methods for observing implementation and contextual setting were thus required.
- (4) Program stakeholders' ideas and actions, in particular those of language teachers, were key to illuminating how language programs function as well as encouraging formative change; involvement of stakeholders in evaluation processes was therefore important for enabling contextual relevance of findings.

- (5) Language program evaluation needed to occur for a variety of purposes, from development to improvement to accountability; intended uses of evaluations should thus lead methodological choices, not vice versa.

Publishing on language program evaluation reached a zenith during this phase, with almost one book per year appearing in the 10 years from 1989 to 1998 (Alderson & Beretta, 1992; Brown, 1995a; Johnson, 1989; Lynch, 1996; Pennington, 1991; Rea-Dickins & Germaine, 1992, 1998; Weir & Roberts, 1994), including two volumes in the authoritative Cambridge Applied Linguistics series. Key contributions of this stage included: (a) the development and dissemination of numerous methods for collecting data—interviews, observation protocols, focus groups, surveys, proficiency tests, and so forth—about the diverse elements that constitute language programs (e.g., Weir & Roberts, 1994), (b) the design of frameworks for describing the logic of language programs and the interconnections among distinct elements, from needs to curriculum and materials to teachers to assessments (e.g., Brown, 1995a, 1995b; Lynch, 1996), and (c) the perhaps naïve if admirable notion that program evaluation had the potential to contribute not only to local improvements in specific language programs, but also to the transformation of stakeholders' roles and prospects within them, through their participation, and to the in-depth understanding of how language education programs work in practice (e.g., Kiely, 1998; Mackay, Wellesley, & Bazerogan, 1995). Notably, with a few exceptions (e.g., Heining-Boynnton, 1990; Horwitz, 1985), nearly all of the examples of language program evaluation generated during this period came from English as a foreign or second language EFL/ESL teaching contexts.

Following a lull in language program evaluation activity from the late 1990s to the mid-2000s, several impetuses worldwide led to renewed interest in the roles played by evaluation in institutions and society, the expansion of evaluation ideas and practices into new sectors of language education, and the exploration of diverse models and methods. A primary harbinger of this next stage in language program evaluation came in the form of the 'accountability movement' and its requirements for numerous domains of endeavor, including: (a) mandated testing of core subject matter competences and related consequences for learners, teachers, and schools in U.S. K–12 public education, (b) the requirement

that academic programs in higher education (primarily North America, the United Kingdom, and the British Commonwealth) incorporate student learning outcomes assessments as a chief component of institutional accreditation, (c) increased monitoring and oversight of the implementation of governmental policies and standards worldwide, and (d) external inspection and impact analysis of federally funded programs in many countries (see Kiely & Rea-Dickins, 2005, 2009). Interestingly, although the top-down, authoritarian, bureaucratic, and high-stakes orientation of the accountability movement was unmistakable—and the consequences for some language programs and learners quite dire (as described in a later section)—the effect on language program evaluation thought and practice was, in many cases, quite the opposite. As evaluators and program stakeholders struggled to respond to these mandates, they developed a newfound awareness of the array of possible roles played by evaluation, they designed thoughtful approaches that simultaneously met external demands and provided internally useful information, they reflected on the value and practice of evaluation, and they initiated empirical investigations into the nature of evaluation use and impact within language programs (see, e.g., Kiely & Rea-Dickins, 2005; Norris, 2009; Norris et al., 2009). Norris and Watanabe (2013) observed several key developments within this next generation of language program evaluation:

- (1) shifting roles and responsibilities, with increased awareness, participation, and ownership over evaluation by program-internal stakeholders and the recasting of external evaluators as facilitators rather than judges;
- (2) cyclical, ongoing, process approaches to evaluation, where questions are prioritized, answered, and acted upon, followed by a new cycle, rather than one-time reviews that seek to address all aspects of a program at once;
- (3) focus on specific program elements (e.g., teachers, assessments) in depth, as they interact with other elements, stakeholders, and factors;
- (4) triangulation of evidence types and sources gathered for evaluation purposes, with an emphasis on mixed-methods data collection;
- (5) identification of intended uses and users of evaluation as the key first step in pursuing evaluations that can provide useful

information and lead to decisions and actions;

- (6) awareness-raising and capacity development for evaluation through the involvement of language program stakeholders in all phases of the evaluation process.

It is also important to note that these recent trends in language program evaluation have emerged not primarily from the typical EFL/ESL settings, but at least as actively from foreign language education contexts in North America and worldwide (e.g., Birckbichler, 2006). Additionally, the new generation of language program evaluators has sought to disseminate not only case studies of evaluation in practice, but also reflections and empirical findings about evaluation in use, all contributing to a still nascent scholarship of language program evaluation, as highlighted in the following section.

CURRENT CONTRIBUTIONS OF LANGUAGE PROGRAM EVALUATION

Contemporary language program evaluation occurs in response to several key mandates, as indicated previously, but it has also begun to develop a set of practices that are suggestive of a newfound value for evaluation from the perspective of language educators and some applied linguists. Thus, language program evaluations can be distinguished on the one hand by the impetuses for engaging in evaluation in the first place, but also by the approaches taken and the consequences that ensue.

Accountability Evaluation

One major impetus for program evaluation has to do with a heightened emphasis on holding public schools, teachers, and students accountable to certain expectations for adequate performance, primarily in terms of learning achievements. Although such accountability schemes exist worldwide (e.g., Harris, 2009; Kiely & Rea-Dickins, 2009), they are perhaps most readily and recently associated with the No Child Left Behind Act of 2002 (NCLB) in the United States, which sought to ensure perpetual improvement in the meeting of educational objectives in K–12 public schooling. The basic premise of NCLB, and other recent test-based accountability movements, is the assumption that regular assessment of all learners in core subject matters, longitudinal tracking of results, and public scrutiny of performance will lead to needed improvements in educational practices

(Koretz, 2003; Linn, Baker, & Betebenner, 2002). An added dimension in the United States has been the requirement that schools demonstrate 'adequate yearly progress' as defined by certain percentages of students (and groups of students, including linguistic minority groups) demonstrating improvements in performance from one year to the next. While the details of accountability testing are beyond the scope of this article, it is important to consider a few of the ways in which it has impacted our thinking about language program evaluation.

One fundamental observation about test-based accountability movements like that inspired by NCLB is that they lack many of the dimensions considered to be essential for ethical program evaluation practice—merely measuring achievements through standardized assessments does little to shed light on myriad factors that conspire to determine test scores, to provide formative information that can be acted upon for improvement, to encourage innovation or experimentation, to enable accurate data-based interpretations, and so on (e.g., AEA, 2006). The realities of such test-based regimes also suggest relatively severe consequences for language learners and language education. Thus, for example, ample evidence has been collected to indicate the detrimental effects of U.S. accountability testing on the academic preparation and achievements of English language learners as well as the depletion of a national resource in the form of bilingual speakers of a variety of languages, where bilingual education or first-language support programs are overwhelmed by the frenzy to meet adequate yearly progress in English proficiency development and academic achievement (e.g., Menken, 2008; Menken & Solorza, 2014). Furthermore, it is clear that the teaching of foreign languages in U.S. public schools has declined dramatically, in particular at elementary and middle school levels, as teachers, instructional hours, and curriculum are shifted toward the core subject areas that are subject to accountability testing (languages are not; e.g., Rhodes & Pufahl, 2010; Rosenbusch, 2005). Another negative consequence is that, in light of the massive media attention paid to accountability assessment, test-based regimes have incorrectly become synonymous with evaluation in the minds of the public as well as academics, including language and humanities educators in leadership positions (e.g., Holquist, 2010).

At the same time, it may be that some (unintended?) positive consequences have ensued, as well-intentioned language educators and evaluators have responded proactively to the pressures

of accountability. One example has been the development of much more insightful English-language proficiency standards and assessments for providing information to K–12 teachers and administrators in a form that can be used to reflect on and improve instruction while also gauging students' proficiency development. For example, the WIDA English language proficiency standards and associated assessments have been developed precisely because of the mandate to measure yearly progress, yet they have also been delivered with a variety of support mechanisms intended to help teachers and administrators interpret and act upon findings (e.g., Gottlieb & Nguyen, 2007; Kenyon et al., 2011), and they themselves are subject to ongoing evaluation (e.g., Bailey & Huang, 2011). Another all-too-infrequent example of a proactive response to accountability testing under NCLB, with respect to linguistic and cultural minority students, has been the development of alternative assessments that can be used to meet accountability demands while simultaneously supporting the rights of indigenous people and the development of languages other than English. For example, the Hawaiian Aligned Portfolio Assessment (HAPA, a play on words for those familiar with the Hawaiian language) was developed as a way of testing students' subject matter learning within the context of immersion schools that conduct the early years of schooling entirely in Hawaiian (Burger & Ryan, 2007). These kinds of responses are indicative of the possibilities for transforming test-based accountability mandates into less pejorative realities for language learners and educators. However, as accountability pressures are sustained into the future, it is likely that additional steps will need to be taken if test-based information is to be usefully integrated into an approach to evaluation that supports meaningful reflection and improvement in language programs while avoiding negative consequences of the sort encountered so far.

Student Learning Outcomes Assessment

Another major impetus for language program evaluation takes the form of accreditation-mandated student learning outcomes assessment in higher education. In the United States, accreditation of universities is conducted by federally authorized regional accrediting agencies, and the major focus of accreditation in the modern era is on 'student learning effectiveness' (Ewell, 2001; for examples outside of the United States, see Brindley, 1998, 2001). The primary mechanism for gauging this dimension of universities

comes in the form of a mandate that all academic programs should regularly assess the key learning outcomes of their students, for example, as they graduate from major degree programs. A key factor that distinguishes outcomes assessment of this sort from accountability testing is that academic programs and institutions are largely left to their own devices to determine what outcomes need to be assessed as well as the preferred methods for doing so. Another key distinguishing factor is that assessment findings are supposed to be used as evidence in ongoing cycles of program monitoring for improvement.

Language educators in higher education have reacted in two ways to the outcomes assessment movement, which, according to a national survey reported in Watanabe, Norris, & González-Lloret (2009), has been the primary locus of evaluation in tertiary language programs in recent years. On the one hand, there has been a predictable rejection of such efforts for a variety of reasons, largely owing to the misinterpretation of what is being requested and ignorance or lack of capacity to engage in assessment and evaluation (see discussion in Norris, 2006). A second reaction, however, has been to embrace outcomes assessment as an opportunity to engage in meaningful program evaluation, and it is in fact in conjunction with the outcomes assessment movement that there has emerged a considerable new scholarship of language program evaluation. Thus, foreign language educators have set about identifying, stipulating, and developing consensus on precisely what kinds of learning (of language, culture, and other targets) should occur as a result of their programs, in the form of learning outcomes statements, something that was heretofore virtually nonexistent (e.g., Gorsuch, 2009; Grau Sempere, Mohn, & Pieroni, 2009; Liskin-Gasparro & Vasseur, 2014; Walther, 2009). They have mapped curriculum, developed courses, and provided new and more diverse learning opportunities in order to help learners achieve targeted outcomes (e.g., Windham, 2008). They have adopted, adapted, and/or designed new assessments for collecting data on student learning, ranging from standardized proficiency assessments to individualized portfolio assessments to alumni surveys (e.g., Pfeiffer & Byrnes, 2009). They have developed other ways of gathering data about diverse factors within their programs that play a role in influencing student learning, such that cause-effect relationships might be determined (e.g., Young, 2008). They have analyzed data, interpreted findings in local relief, and taken actions to improve their programs as a result (e.g., Norris & Davis,

2015). Ultimately, they have learned through the process, both about their programs and the kinds of student learning they encourage, and also about program evaluation and assessment as viable processes that can contribute to the betterment of language and humanities education (e.g., Cachey, 2014; Cachey & Pfeiffer, 2015).

While it would be premature to suggest that outcomes assessment is having anything like a widespread impact along these lines (indeed, misguided and vitriolic reactions persist), it is nevertheless encouraging that at least some foreign language educators have taken the mandate at face value and taken up the responsibility to pursue meaningful evaluation as a result. Recent discussions and empirical observations (Davis, 2015; Davis, Sinicrope, & Watanabe, 2009) suggest that key factors underlying this proactive adoption of outcomes assessment for useful purposes include: the will to take ownership over the process and make it work for the sake of the program and its stakeholders, sometimes in spite of external impetuses; leadership that facilitates assessment and evaluation through encouragement and provision of resources; an iterative and balanced approach that sets achievable targets and proceeds cyclically; and internal efforts at developing capacity to pursue evaluation.

Program Development, Monitoring, and Improvement

Although accountability regimes and outcomes assessment present the most readily apparent impetuses for program evaluation, at least as much energy has been devoted to formative, local, and largely internal language program evaluation efforts. Rather than holding programs to a standardized bottom line or enforcing external criteria for achievement, these evaluations prioritize the gathering of locally useful information for answering priority questions about language programs and helping stakeholders and decision makers resolve immediate challenges.

Needs analysis is one such endeavor, in which various data collection methods and information sources are utilized to determine what the purposes, goals, and targeted outcomes of language programs should be (Long, 2005). Needs analyses may assume a primarily developmental role, that is, prior to program creation, by providing a basis for understanding who the clients and other stakeholders of the program are, what their (and others') expectations for the program may be, and what kinds of program activities may best foster desired outcomes. Thus, in a highly localized example, Chaudron et al. (2005) conducted

surveys to determine the target tasks that interest learners of Korean as a foreign language, followed by the recording and discourse analysis of authentic exemplar tasks, all as a basis for curriculum design and materials development in a college language program (see Iwai et al., 1999, for a similar example in Japanese). At a much broader level of use, Van Avermaet and Gysen (2006) report on the use of interviews and surveys across multiple stakeholder groups in Belgium—including learners of various ages, employers, educators—as a foundation for identifying the types of Dutch language tasks that immigrants to Flanders would need to be able to accomplish in various sectors of society and for diverse communication purposes, including everything from survival to schooling to vocational to higher education uses of the language. Findings from this needs analysis went on to inform the development of published materials, sample syllabi and lesson plans, teacher training, and, importantly, language assessment. Needs analysis may also be pursued as a way of augmenting already existing programs, by determining perceived gaps, identifying changes in learner or client population characteristics, and otherwise examining what is being delivered versus what might be sought from a program. For example, Milleret and Silveira (2009) utilized focus groups, surveys, and document analyses to identify uniquely differing groups within an existing Portuguese language program—including Spanish speakers as well as non-Spanish speakers—and to establish key differences in their learning trajectories; these findings were used as a means for creating new courses and associated materials aligned with the differentiated needs of the two learner groups, and as a basis for justifying expansion of the program.

Program monitoring, typically required as a stipulation of funding or as an approach to management, may also adopt an explicitly formative orientation as opposed to the perfunctory meeting of external demands (Mackay, 1994; Mackay et al., 1998). Monitoring frequently involves the use of tests and other outcomes measures, as a basis for establishing program effectiveness, as well as observation and inspection to determine the extent to which key program elements (e.g., materials, classroom practices) are being realized as intended. However, recent examples suggest that mandatory monitoring may also adopt a formative role when implemented in consultation with, and with participation by, program-internal stakeholders. For example, Llosa and Slayton (2009) discuss the evaluation of a high-cost supplementary reading program in a U.S. public school system,

the measured outcomes of which suggested no benefit, in particular for ESL learners. However, by carefully investigating the processes by which the program materials were implemented, from the perspectives of the teachers and learners involved, the evaluators were able to identify numerous features in need of improvement such that the program might have a chance of success. In another example, Elder (2009) reflects on the evaluation of state-funded heritage language programs in Australia, pointing to the key roles for participation and communication among stakeholders, evaluators, and funders, as a baseline condition for enabling improvement of programs above and beyond monitoring expectations. One key contribution of these and other recent monitoring evaluations is the realization that multiple modes of inquiry are necessary to accomplish diverse purposes of monitoring evaluation (see additional examples in Ekkens & Winke, 2009; Harris, 2009; Kiely & Rea-Dickins, 2005; Koyalan, 2009; Van den Branden, 2006).

Another increasingly frequent focus of language program evaluation in the last decade has centered on explicitly formative purposes, that is, internally oriented evaluations that endeavor from the outset to provide information that can be used by local decision makers for specific aspects of program improvement (see, e.g., collected articles in Norris et al., 2009). Key features of these evaluations, many of which draw heavily on utilization-focused evaluation (Patton, 2008), include a highly participatory orientation, where teachers or other stakeholders take the lead in evaluation; extensive use of qualitative data collection methods (e.g., interviews, focus groups, meetings, journals); and cycles of posing questions, gathering evidence, interpreting, and making decisions, followed by new questions. For example, Yang (2009) reports on a teacher-generated evaluation of new teacher induction practices within a university ESL program, where the challenge of constantly evolving teachers suggested the need for understanding and improving the ways in which they were introduced to the job and otherwise brought into the program. Yang worked with program administrators and experienced as well as new instructional faculty first to identify the broad goals of the program and second to gather perspectives on whether and how the goals were being met through various induction practices; her detailed findings were in turn utilized by program administrators for making substantial expansions in induction practices. In another example, Kiely (2006, 2009) explores factors that influence how teachers respond to and

act upon evaluation data, highlighting in particular the ways in which teachers' identities and awareness affect how they approach evaluation, the rich information requirements (e.g., about how learners respond to materials and pedagogy) for useful teaching-related evaluations, and the importance of teacher ownership as a means for encouraging learning and action through evaluation.

Effectiveness Evaluation

Another very common orientation in recent language program evaluation work is the determination of 'effectiveness' or 'impact' of various interventions on language learning and/or other outcomes (e.g., learner motivation). Distinct from accountability or outcomes assessment evaluation, these evaluations tend to focus quite narrowly on specific, measurable phenomena that are ostensibly treated by innovative program designs, perhaps most commonly in the form of teaching approaches, materials, or curricula. As well, these evaluations typically reflect a particular theoretical premise from the field of second language acquisition, and they often adopt quasi-experimental methodologies to 'test' for program effectiveness. Thus, for example, numerous evaluations have been published in recent years on the effectiveness and/or impact of approaches to technology-mediated language teaching, in terms of learner motivation (e.g., Cai & Zhu, 2012), student perceptions of usefulness (e.g., Pardo-Ballester, 2012), and proficiency outcomes (e.g., Young, 2008). Other effectiveness evaluations have examined intensive summer courses (e.g., Xu et al., 2012), study abroad (e.g., Badstübner & Ecke, 2009), heritage language program design (e.g., Nocus et al., 2012), and models of bilingual education (e.g., Admiraal, Westhoff, & de Bot, 2006). A major challenge for effectiveness evaluations is the determination of what, precisely, within the program led to which kinds of impacts. As Kiely and Rea-Dickins (2009) have pointed out, programs are highly dynamic and complex, such that richly contextualized information is called upon to shed light on putative causal mechanisms in effectiveness evaluations. Indeed, it may be that there has been a tendency in such work to conclude that a particular impact occurred on the basis of a particular program design feature, simply because that was the assumption from the outset of the evaluation (i.e., rather than investigating the variety of plausible factors that might have contributed to observed outcomes of the program experience). Whether effectiveness eval-

uations of this sort legitimately can be considered *program* evaluations, versus more or less controlled experimental research and hypothesis testing, is a question in need of further consideration.

Language Teacher Evaluation

A final area of considerable related attention has been the evaluation of language teachers for various purposes, including certification, the improvement of teacher training programs, and individualized formation. Of course, language teachers do not constitute language programs, and the reasons, approaches, and methods for teacher evaluation are somewhat distinct from those adopted for broader program evaluations. Nevertheless, teachers are important elements of programs (in many cases probably the most important elements), hence their evaluation may play a role in our thinking about program evaluation. For example, Burden (2008) considers the implementation of externally mandated evaluations of English language teachers by students at the end of semester in Japanese schools, identifying numerous problems with the process, including credibility of students' perspectives and understandings regarding teaching, lack of buy-in by teachers, and lack of formative follow-through procedures (see also Sundquist & Neary-Sundquist, 2008). These concerns, of course, go to the heart of thinking about the utility and ethics of any evaluation procedure, suggesting that teacher evaluation might also benefit from consideration of the fundamental principles of program evaluation. Other language teacher evaluation work has focused on teacher certification programs (e.g., Carstens-Wickham, 2008; Peacock, 2009) and the role played by teacher beliefs in program delivery (e.g., Martinez & Sanz, 2008). Finally, Coombe et al. (2007) provide a compendium of recommended practices, case studies, and reflections on the evaluation of English language teachers.

THE FUTURE OF LANGUAGE PROGRAM EVALUATION: INNOVATIONS AND CHALLENGES

While many of the current trends in language program evaluation will persist, much remains to be accomplished in realizing the value of evaluation within applied linguistics. The immediate future will see a sustained focus on responding to the kinds of evaluation mandates that characterize the contemporary era (e.g., accountability,

outcomes assessment) to be sure, yet there is reason to believe that the groundswell of program-internal evaluation ownership, formative orientation, and emphasis on utility will continue to grow (for some examples, see Norris & Mills, 2014). Language program evaluation practice may also evolve in new directions. Some developments in mainstream evaluation models and methods have yet to gain purchase in the language fields, while evolving characteristics of language programs themselves may demand new ways of enacting evaluation. At the same time, key tensions, constraints, and other challenges will determine the extent to which language program evaluation becomes a core contributor within applied linguistics or remains a relatively marginalized endeavor.

Potential Innovations in Language Program Evaluation

Innovations in language program evaluation are likely to be driven by both (a) adoption and adaptation of heuristics and methods from other domains of evaluation practice, and more likely (b) needs for new ways of engaging in evaluation that respond to the changing nature of language programs and our understandings about language learning, learners, and teaching. Here, a handful of likely innovations are considered.

Logic Models and Theory Testing. While there has been sustained interest over many years in the possibility of testing theories of language learning and education (e.g., Prabhu, 1987), in an attempt to support or falsify core ideas, it was also determined early on by language program evaluators (e.g., Beretta, 1992a) that methods-comparisons and related experimental approaches typically revealed very little about whether a particular theory of instructed language learning was actually being delivered, was not unduly influenced by other factors of the educational (or extracurricular) setting, was in fact causally leading to measured outcomes, and so on. Perhaps the major challenge to theory testing within a field like language education is the fundamental lack of a theory of practice or change in the first place.² Ralph Tyler (1948), one of the foremost proponents of rational design and objectives setting in mainstream education, even commented on this state of affairs in the *MLJ* well over a half century ago in reference to early evaluations of foreign language teaching methods in the United States:

The kinds of students to be taught, the ends to be attained, the psychological nature of language learning, the conditions under which effective learning

takes place, the requisites of effective organization of learning experiences were either unstated or were largely expressed in unique terms making comparison impossible. (p. 560)

His comments are telling in their sustained relevance to models of language education today. Thus, while hypotheses and theories related to diverse aspects of instructed language learning abound, as yet none describes the detailed causal, interactive, and moderating relationships among the variety of factors that conspire to bring about language learning in a certain way and with specific resulting outcomes. Program logic models (e.g., Funnell & Rogers, 2011) may be particularly helpful in this regard, if developed by language educational experts and communities of practitioners. A logic model outlines the program's theory of change by spelling out the critical assumptions that define all identifiable components of a program, typically including: (a) the needs to which a program responds, characteristics of the setting, and the priorities targeted, (b) the theory of change or development that informs all aspects of instruction, including scope and sequence, pedagogic activities, assessments, and so on, (c) the inputs provided, including various resources like teachers and materials and physical space, (d) the outputs or activities and participants that constitute the delivery of the program, in other words what is supposed to happen with whom as something like teaching ensues, and (e) the short-, medium-, and long-term outcomes that are intended to result from program delivery, such as learner engagement and uptake, development of knowledge and abilities, changes in individual life prospects and/or social conditions, and so on. Given an adequately specified program logic model, it may then be possible to set about collecting evidence regarding not only (a) whether the program actually achieves targeted outcomes, but more importantly (b) whether all dimensions of the model are realized and implemented as intended, (c) which factors in particular may be moderating outcomes achievement (including possibly factors external to the program per se), and (d) whether or not the model as a complex, interactive whole is viable.

Logic models of this sort are, to date, virtually nonexistent in language education or applied linguistics, and they certainly have not been utilized as a means for fully testing a program's theory of change about language teaching and learning.³ Ross (2009) provides a useful extended description of a hypothetical language program logic model and how it might be implemented to

both develop and then test a theory of language learning, and particularly the factors that may constrain language learning, within a given instructional setting. Perhaps the only publicly available logic model of a functioning language program is that of the Language Instruction for Newcomers to Canada (LINC) program (Government of Canada, 2004). For this federally funded initiative, a series of detailed logic models was developed to outline the core components and expectations, from activities to outcomes, for each of three interconnected programs: language assessment, language training, and delivery assistance. The logic model was then utilized as a framework for informing the evaluation of the LINC program with a particular focus on its need and rationale, delivery, and effectiveness/efficiency. Rather than serving as a 'test' of the program theory, though, the model provided a comprehensive basis for initial monitoring of all of the different dimensions of the program to ensure accountability and highlight directions for improvement.

The potential of program logic models for advancing language education and related endeavors is tremendous, in that—if developed by language educational experts, practitioners, and other insiders (i.e., versus external 'logic model' experts)—they would lay bare the rationales, activities, and implicit theories that constitute language programs and thereby render them amenable to empirical confirmation (or rejection). Of course, doing so may be particularly threatening to language teachers, curriculum developers, materials designers, and program administrators, as it opens up the real possibility that they are simply not functioning with any kind of program logic in mind, never mind the likelihood that expectations for how programs are functioning will not meet observable realities. One domain of theory and practice within language education, task-based language teaching (TBLT), does exhibit many of the necessary conditions for taking the next step to logic model development and evaluation. TBLT, in its fullest realization (Bygate, Norris, & Van den Branden, 2015), is built upon the explicit connections between learners' needs and the identification of valuable target language tasks, the use of pedagogical cycles that are related to those tasks and follow carefully designed instructional sequences, the capacities of teachers to engage in quite specific instructional activities, the use of authentic materials, and the assessment of learners' outcomes in terms of language performance on target tasks. While testing a 'theory' of TBLT may not be an appropriate use of evaluation at this point (though see initial ef-

forts in, e.g., González-Lloret & Nielson, 2014; Van den Branden, 2006), it is quite possible that TBLT program logic modeling could help to advance endeavors within that field, especially by requiring clarity in assumptions regarding the delivery of TBLT in specific educational settings, and it might also enable a more comprehensive evaluation along the lines of that demonstrated with the LINC program.

Assessment Validity Evaluation. Practices in the validation of language assessments have developed rapidly over the past decade, with Kane's (2006) argument-based validity approach gaining particular prominence among language testers (e.g., Chapelle, Enright, & Jamieson, 2008). However, according to Kane (2001) himself, "validation of decision procedures has always depended on the *evaluation* of the consequences of the decisions" (p. 339, my emphasis). Indeed, there is a substantial line of thought within educational assessment that has linked program evaluation with test validation, as stated most directly by Cronbach (1988): "Validation of a test or test use *is* evaluation (. . .) so I propose here to extend to all testing the lessons from program evaluation" (p. 4, his emphasis). In language assessment, much as in mainstream educational assessment, this idea has yet to be fully realized. While greater attention has been paid in recent years to the uses of language assessments, the interpretations implied by distinct uses, and the consequences that ensue from particular approaches to assessment (e.g., Bachman, 2005), to date, argument-based validation and attendant validity evaluation practices remain focused primarily on assessment instruments themselves and on construct interpretations (even those implied in uses) while falling short of other dimensions of language assessment that have more to do with the roles they play and the effects they have in educational and social programs.

However, there may be considerable promise in the further exploration of an explicitly program evaluative approach to validation as a way of understanding, judging, and improving the value of assessments in actual use—and as opposed to the more typical practice of 'validating' an assessment as a trustworthy measure of a given construct (never mind how it is used) or even the much improved practice of investigating validity as a series of inferential links from instrument to interpretation to use. The basic starting point in a program evaluative approach to assessment validity evaluation is the reconceptualization of assessments as programs, not instruments, with all of the

attendant elements of programs (i.e., assessments respond to specific needs, they are designed according to some kind of logic, they require resources and inputs, they are enacted by people, they produce outputs and outcomes, and their use has consequences for individuals and programs). Just as with the evaluation of other kinds of programs, evaluation of assessments would proceed on the basis of identifying the intended uses and users of the evaluation (why are we evaluating validity in the first place?), the prioritization of evaluation questions that are relevant at particular points in time (e.g., for developing assessments, for improving assessments, for determining the impact of assessments, for terminating assessments?), the collection of fitting data, and the resulting use of findings to inform decisions and actions related to the assessment program. To date, the only comprehensive example of programmatic validity evaluation in language assessment has been Norris's (2008) multiyear investigation into the development, implementation, and impact of a German language placement exam. According to Norris (2008), "the shift to validity evaluation seeks to transform validation into a worthwhile and relevant endeavor by making its purposes explicit and by contextualizing its use within a specific community with clearly defined interests in a particular assessment program" (p. 75). It may well be that this use of program evaluation applied to assessments will gain purchase in particular at the local level, that is, as language educators within programs seek a framework for inquiring into the design, use, and impact of assessments within their classes, programs, and institutions. Whether larger-scale, standardized assessments will also benefit from an explicitly program evaluative approach to validation remains to be explored as well.

Technology, Big Data, and Evaluation. The rapid rise of large-scale, distributed, open-access learning in the form of technology-mediated education models and platforms (e.g., Massive Open Online Courses, or MOOCs) is suggestive of certain directions for language program evaluation, though to date it is uncertain to what degree language education will fully participate in this movement (Godwin-Jones, 2014). An associated development has been the advent of 'big data analytics,' or the tools and methods used for making sense of the massive amounts of data generated by social media, including internet-based education. Indeed, it has been speculated that, as education and other social programs are increasingly technology-mediated—which implies the

automated capture and archiving of all associated data—our understandings of the possibilities, frameworks, and methods for evaluation of the same will be fundamentally changed (Few, 2012; Means, 2014). Differences may include: (a) the re-envisioning of programs, in terms of scope, participation, and design (i.e., as they are distributed via the internet to very large numbers of stakeholders), (b) the availability of large amounts of data, persistently collected, and on demand (i.e., as opposed to traditional assumptions regarding the expense and difficulty of collecting data), (c) the rapid turnaround in impact assessment (i.e., the possibility of capturing certain effects of interventions immediately, based on tracking of participants' next actions), and (d) the shift in evaluation focus away from reflection on quality and effectiveness and toward prediction of how participants will respond and what they will do as a result.

While it may be as yet premature to conceive of 'big data' evaluation related to online language education, a few examples point in possible directions. Data mining of large and rich samples of language learning activities and outcomes is one clear possibility. Thus, for example, Bloomfield et al. (2014) describe the analysis of some 800 language learners for whom not only language proficiency test scores in the four skills were collected over substantial periods of time, but also several other sources of information regarding length of study, participation in formal instruction, and so on. Using latent growth and event history analysis techniques, they were able to identify relatively clear relationships among these variables and both the development and attrition of language proficiency. In another example, Ecke and Ganz (2014) show how tracking of student enrollments, preferences, career trajectories, and institutional data over time allows for insights to be gained into program-external factors that may influence a language program's viability. While neither of these examples reflect online language education per se, they both point to the analytic evaluation activities that might benefit from the kinds of 'big data' easily accessible in technology-mediated settings.

A second dimension of technology-mediated language education will imply other expectations for language program evaluation. Beyond big data through large numbers of participants, the fundamental shift toward computer-based, online, and other technological affordances suggests that the nature of language teaching and learning will itself be transformed into distinct delivery models and participation patterns. What

language programs look like in virtual worlds, social media environments, telecollaboration approaches, fully computer-automated courses, and so on, will differ in ways that are as yet not fully comprehended. To be sure, the ways in which language teachers and learners engage with these distinct spaces and practices of language learning will be the subject of program evaluation (e.g., Cai & Zhu, 2012; Gallardo del Puerto & Gamboa, 2009; Ma, 2008), if only to establish a baseline for what is possible and what really does not work (technologically and pedagogically speaking). More comprehensive evaluations of the design and delivery of particular models of education via technological platforms and/or media have been pursued infrequently as yet (Hill & Tschudi, 2008), perhaps because the technological capabilities themselves are developing and changing so rapidly. While a handful of evaluations have been conducted on popular computer-based learning packages, such as Rosetta Stone, the results have differed so dramatically that it is unclear what to make of them (e.g., Rockman et al., 2009, versus Nielson & Doughty, 2008). In the end, what is perhaps most certain is that (a) evaluation is currently lacking in the midst of rapid (rabid?) innovation and deployment of technology-mediated language courses and programs, and (b) evaluation is sorely required in order to sort out basic issues related to the need for such programs, their theories of change and associated pedagogical designs, their delivery and other aspects of their realization (including technological capabilities required of participants and technological demands on computing power of various applications), and their outcomes in terms of language learning and other achievements.

Transformative Evaluation. In response to arguments that program evaluation replicates power inequalities and serves to uphold political or social hegemonies, mainstream evaluation has developed a host of practices collectively referred to as 'transformative evaluation' (Mertens, 2009). While most evaluation models today incorporate participation and collaboration by key stakeholders as a fundamental principle, program evaluation is considered transformative to the extent that it adopts an explicit social justice agenda as an overarching dimension of its conduct. Popular versions that have experienced considerable uptake in various sectors (including in particular evaluations conducted for transnational human rights organizations like UNESCO) include empowerment evaluation (Fetterman, Kaftarian, & Wandersman, 2014), which empha-

sizes capacity development of program actors for social change, and deliberative democratic evaluation (House, 2005), which involves the canvassing of all relevant perspectives, facilitation of dialogue among stakeholders, and extended deliberation about the meaning of evaluation findings. Fundamentally, all transformative evaluations utilize methodologies for eliciting, showcasing, and acting upon the voices of disempowered or otherwise marginalized stakeholders, balancing the weight afforded to distinct groups in decision-making processes, and otherwise encouraging social change via evaluation.

Although a social justice approach aligns well with contemporary orientations in applied linguistics (e.g., critical applied linguistics, the multilingual turn, language identity research), there have been very few examples to date of transformative language program evaluation. Occasionally, mainstream evaluators have undertaken transformative projects in relation to language minority populations. Thus, for example, House (2004) adopted a deliberative democratic approach in the evaluation of court-ordered English language programs for Spanish-speaking students in the Colorado public school system, and Mertens et al. (2010) engaged in transformative evaluation of teacher preparation programs for deaf students. One very recent example of transformative evaluation at work within language education is more suggestive of the possibilities, though. Zannirato (2014) adopted an empowerment evaluation approach in seeking to illuminate the problems and conflicts inherent in two-tiered (i.e., consisting of tenure-track and non-tenure-track faculty) foreign language departments in U.S. universities. In particular, his work gave voice to the perspectives of typically disempowered non-tenure-line language faculty, and it indicated clearly the extent to which these groups experienced greater perceived conflict and less control in relation to their professional lives. Zannirato ended this important work by suggesting an action plan for building toward change at the individual, departmental, and national level, thereby taking the social justice agenda full circle. Whether these kinds of transformative practices will become more prevalent in language program evaluation will depend to considerable extent on whether applied linguists and language educators are exposed to and receive training in such ideas. To be sure, there are many opportunities for good to be done through transformative evaluation in application to the situations of language teachers, language learners, and language minority populations worldwide.

Meta-Evaluation. A final area of likely development in the future of language program evaluation has to do with the practice(s) of meta-evaluation. In its original interpretation (e.g., Scriven, 1969; Stufflebeam, 1973), meta-evaluation was intended to imply the evaluation of evaluation itself—in other words, it involved the use of evaluation methods for judging and improving the quality and effectiveness of specific program evaluations, models, and methods. In this sense, meta-evaluation is a sort of validation procedure that helps to ensure the trustworthiness and utility of evaluation in practice, and the mainstream evaluation community has fully embraced this notion, creating guidelines for meta-evaluation (Stufflebeam, 1999), incorporating recommendations for meta-evaluation into professional program evaluation standards (Yarbrough et al., 2011), and engaging in numerous examples of meta-evaluation across diverse fields of evaluation work (e.g., Sánchez-Meca, 1997). To date, meta-evaluations of language program evaluation have not been undertaken, due very likely to a lack of available models specific to language program evaluation, the minimal accumulation of published reports on language program evaluation, and the still emergent status of the endeavor. Nevertheless, it may be that the meta-evaluation of a few prominent examples of evaluation in relation to language education or social programs would provide a much needed impetus for clarifying and improving what we do.

A second interpretation of meta-evaluation is the application of meta-analytic techniques to the synthesis of findings from evaluation reports. In brief, meta-analysis involves the review of all available accumulated research on a given topic and the use of various statistical and logical techniques to identify the average effects, key moderating variables, and other dimensions of findings as these have emerged across numerous primary studies (see Norris, 2012). In meta-evaluation, these techniques are applied to domains where numerous program evaluation reports have accumulated, such that answers about ‘what works’ in terms of program effectiveness (for example) may be reliably determined. Again, within applied linguistics, there are very few domains of practice where program evaluations are common, never mind published. One exception is the considerable accumulation of evaluations that have focused on the effectiveness of different models of language and academic education for linguistic minority students in the United States. Consequently, a handful of meta-evaluations have emerged, perhaps the most representative of

which is that by Rolstad, Mahoney, and Glass (2005). Their work endeavored to synthesize findings of evaluations that examined the effectiveness of bilingual transition programs versus English-only immersion programs, and they were able to show convincingly that Spanish-speaking students who experienced bilingual education benefited considerably more than their counterparts in English-only programs, both in linguistic and academic development. The application of meta-evaluation in other domains of applied linguistics and language education will depend, most obviously, on the accumulation of published program evaluation reports as well as the development of meta-analytic capacities within the field.

Challenges to Realizing the Potential of Language Program Evaluation

The potential of language program evaluation to contribute along the lines outlined previously will depend at least in part on whether and how several challenges are resolved. Prominent among these are the lack of training in program evaluation available to or pursued by applied linguists and language educators, ambiguity regarding language program evaluation as a scholarly endeavor (and concomitant challenges for dissemination), and several tensions internal to the conduct of language program evaluation that will define how it is perceived and utilized into the future.

Challenges for a Scholarship of Language Program Evaluation. A cursory look at major graduate (a.k.a. postgraduate) training programs in applied linguistics and language education worldwide reveals the plain fact that almost none of them offers the opportunity for students to develop knowledge and abilities in language program evaluation. Free-standing classes devoted to language program evaluation per se can be counted on one hand, and it is rare that space is afforded in most MA or PhD programs for students to seek out program evaluation coursework from other disciplines. Furthermore, although much would be gained by training in evaluation from outside of the applied linguistics field, it is also the case that many of the issues core to the effective practice of language program evaluation probably would not be addressed in such classes (e.g., the nature and complexity of student learning outcomes within language education; the logic and theory of language learning programs; the types of data collection methods appropriate for use

with language learners and other stakeholders, including in particular language assessments for diverse skills, competencies, and dispositions; and so on). What is perhaps of most concern, though, is the reality that program evaluation is simply not on the map from the perspective of applied linguistics. In a recent survey of applied linguists in the United States conducted by the Center for Applied Linguistics, Malone and Donovan (2014) found no indication that language program evaluation was considered an important competence for scholars, nor was it perceived to be a necessary component in graduate training. Without basic training in, or even awareness of, language program evaluation, it is doubtful that applied linguists or language educators are prepared to either respond to mounting demands for evaluation in relation to language programs or to take advantage of the considerable potential that innovations in evaluation may have to offer.

A closely related concern is the ambiguous scholarly status afforded language program evaluation within applied linguistics and language education. Perhaps most indicative here are the limited opportunities for dissemination of evaluation work, either in the form of conference presentations or publications. In terms of presenting at refereed conferences, on the one hand, it is typical for 'evaluation' to be listed as part of a research strand or emphasis at most applied linguistics conferences. For example, the International Association of Applied Linguistics, AILA, included the research strand 'Language Evaluation, Assessment, and Testing' at its 2014 triennial congress, and the U.S.-based American Association for Applied Linguistics (AAAL) conference regularly includes a strand for 'Assessment and Evaluation.' Similarly, the annual convention of the American Council on the Teaching of Foreign Languages (ACTFL) features an 'Assessment' strand, which includes 'evaluation' as a possible presentation focus. Typically, however, presentations on language program evaluation are rare or nonexistent at any of these conferences. Indeed, when the term evaluation appears in a paper title or abstract, it much more frequently implies assessment (as in teacher 'evaluation' of student writing), not *program evaluation*. This common conflation in terminology, moreover, is likely reflective of the orientations and interpretations of reviewers selected for such strands; they are no doubt baffled by receiving proposals for program evaluation presentations when they are focused on reviewing the much more typical assessment and testing proposals. Regardless, the bottom line is that presentations on language program evaluation at major confer-

ences are scant; for example, the AAAL 2014 conference featured a single presentation addressing a program evaluation topic, and at the ACTFL 2014 conference, 8 out of 802 presentations could be considered to have addressed some aspect of program evaluation to some degree.

Scholarly publishing on language program evaluation faces similar challenges.⁴ Publications about the topic appear rarely in applied linguistics and language education journals, and total book-length monographs and edited collections published since the 1980s number in the low teens. A quick review of the Foreign Language Program Evaluation Project⁵ annotated bibliography shows that, in that database, some 143 publications related specifically to language program evaluation appeared between 1978 and 2012, for around four total publications per year, many of which were methodological guides or commentaries rather than substantive reports. Of course, there are multiple reasons for such low publication rates. First, some program evaluations may be associated with sensitive findings and high-stakes implications that are unavoidably connected with a particular institution and group of individuals (so, the possibility of anonymity is ruled out, and university Institutional Review Boards may not know what to make of this situation). Frequently, such evaluation reports take the form of confidential technical reports provided to intended users and implicated stakeholders only, so public dissemination is ruled out.

Second, and more troubling, program evaluations per se (never mind program evaluation endeavors themselves serving as a focus of research) do not fit easily into either the explicitly theoretical research orientations or the typical research report formats of refereed journals in applied linguistics. Where evaluations are largely emergent and evolving over time, mixed-methodological, and relatively agnostic to theory (beyond local curricular theory or program logic), most applied linguistics journals explicitly prioritize the publication of theory-driven empirical research that adopts the constrained features and methodological expectations typical of established epistemologies and associated reporting genres. Indeed, some journals, like the *Modern Language Journal*, actively dissuade contributors from submitting work of the very sort that program evaluations typically embrace. Thus, the author guidelines of the *MLJ* emphasize on the one hand that articles should be driven by theory: "Irrespective of methodological orientation, authors should position their study within a theoretical framework." They also cast a highly dubious light on some

of the empirical methodologies most typical of evaluations: “Purely Questionnaire and Interview-based studies tend not to advance theoretical and empirical knowledge in any appreciable way. In general, such submissions are discouraged” (*Modern Language Journal*, 2013). Last, there are no applied linguistics journals that explicitly encourage or regularly publish work on language program evaluation. In fact, no single applied linguistics journal even mentions program evaluation within the scope of their publishing purviews, though perhaps the one that comes closest is *Language Teaching Research*, where some 10 articles (including one special issue) have been published on language program evaluation since 2009.⁶

Tensions Inherent in Language Program Evaluation.

A final set of challenges for the future of language program evaluation has to do with tensions inherent in the purpose and conduct of evaluations, a few of which are considered here. One major tension involves who is responsible for evaluation and what happens as a result. On the one hand, when evaluation is pursued internally by program stakeholders themselves, considerable awareness-raising and development may ensue as the realities of programs are illuminated by empirical findings and as individuals and groups are encouraged/required to take joint actions in response (e.g., Cachey, 2014; Grau Sempere et al., 2009; Yang, 2009). At the same time, these internal evaluations will inevitably be called into question in terms of supposed subjectivity, with external evaluation and standardized measurement of outcomes proposed in their stead. Key to resolving this tension is the determination of intended uses for evaluation and the ultimate consequences or impact sought. In language programs, where the goal is improvement of practice *and* accountability, a solid case is beginning to emerge in favor of primarily internal evaluation (Davis et al., 2009), in that it seems to enable those responsible for programs to pursue needed change and innovation rather than repressing their abilities to do so. Nevertheless, the need to communicate beyond the boundaries of a program to other interested constituents (e.g., funders, the public) is common, and doing so will require a balance to be struck in terms of the information demands of various program-external versus -internal users.

A second tension has to do with the possibility that language program evaluation is a culturally biased undertaking, reflecting in particular the values of modern Western societies (Kawakami et al., 2008; Mertens & Wilson, 2012). Indeed, the

practices of questioning how and how well programs and individuals are functioning, gathering data in transparent and public ways, and suggesting improvements or making decisions about the future of programs, are all activities that may be at best uncommon and at worst antithetical to what is considered appropriate conduct in certain cultures. For language program evaluators, these concerns are of particular importance, given the inextricably cultural dimensions of language learning and use. Interestingly, the mainstream program evaluation community has actively addressed the issue of cultural bias in its models and methods as well as the need for cultural competence of practicing evaluators. The “Cultural Reading” of the Program Evaluation Standards (AEA, 2011) provides thoughtful recommendations for culturally sensitive practices, where culture is defined according to individual or group distinctions by “race, ethnicity, gender, age, sexual orientation, social class, disability, language, and educational level or disciplinary background” (n.p.).

It is also important to note that the International Organization for Cooperation in Evaluation lists current membership from national evaluation associations in 105 countries worldwide, suggesting that program evaluation is a common enough practice across many regions and cultures to support the creation of professional organizations. One such organization, for example, is the Kaupapa Māori, responsible for evaluation in relation to public programs for the Māori people in New Zealand. Their approach to program evaluation is instructive: On the one hand, they assert the fundamental value of their cultural (and linguistic) practices in defining and sustaining a Māori place in the world, yet they also embrace fundamental evaluation values like systematic data collection, evidence-based deliberation, and rational decision making for improvement. Key to enacting appropriate program evaluation is participation and collaboration by Māori stakeholders, the use of indigenous methods for data collection, the persistent viewing and interpreting of issues and findings from a Māori point of view (and in comparison with other points of view), and the realization that interacting with other cultural communities and external agencies is a reality that can be shaped by participation in (and ownership of) evaluation (Kawakami et al., 2008; Pipi et al., 2003). Their example is indicative of the possibilities for engaging in useful evaluation that is at once pursuant of the core values of evaluation and simultaneously

imbued with culture-specific meaning, ethics, and acceptance.

CONCLUSION

In closing, a final tension may serve to summarize the status quo and future potential of language program evaluation. Evaluation is considered by some to be a means for determining what works, for understanding program effectiveness and impact, and for making decisions about the distribution of resources. That is, evaluation is fundamentally a science that examines and interprets the delivery of theoretical (e.g., policy) ideas as they are realized in holistic, programmatic practices. The job of evaluation in this sense is to uncover the truth about programs. A distinct perspective on evaluation is that it entails a process for helping individuals and groups understand and improve their efforts at addressing some kind of a societal need through the provision of intentionally designed services in the form of programs. Here, evaluation emphasizes participation by program stakeholders, the articulation of data to their information needs, awareness-raising and learning about program practices, and enabling change through the empirical lens of evaluation findings. In this sense, evaluation is much less about discovering truths regarding which or whether programs work than it is focused on the pragmatic endeavor of crafting and refining a particular, sought-after program reality.

Language program evaluations may pursue either or both of these ends. What is crucial is not that one or the other is decided to be the most appropriate way of doing evaluation; rather, it is realizing that program evaluation may be undertaken for a variety of legitimate purposes, that those purposes must be identified and made transparent to all involved, and that fitting methodologies be articulated to meeting the corresponding information needs. The real contribution of program evaluation in applied linguistics—what we can learn from mainstream evaluation practice—may be that it helps us to both understand our theories and ideas as they are applied in action, and to facilitate their application by real individuals and groups in ways that are meaningful, practical, and useful in the first place. Furthermore, to the extent that applied linguistics really is an *applied* (versus ‘basic’) science focused primarily on “dealing with practical problems of language and communication” (AILA mission statement), it may be that there is a unique and important niche to be filled by

language program evaluation, above and beyond the adoption of helpful practices from the evaluation mainstream. Language program evaluation may also provide a much needed and heretofore missing epistemological space to applied linguists for deliberating, understanding, and (critically) acting upon the specific problems of language learning and use, as we go about crafting intentional and ethical responses to them in the form of educational and social programs.

ACKNOWLEDGMENTS

My thanks go to Heidi Byrnes for her belief in the value of this contribution, and to two anonymous reviewers for their encouraging and helpful comments on an earlier version of the article. I would also like to acknowledge the considerable influence that many students and colleagues over the years have had on my thinking about language program evaluation, and I wish to highlight in particular my mentor at the University of Hawai‘i, J. D. Brown, who first introduced me to program evaluation and who never failed to ask of my work: “So what?”

NOTES

¹ A first, major contribution of evaluation-like practice in language education might be attributed to the Modern Foreign Languages Study of the American and Canadian Committees on Foreign Languages (Coleman, 1929), which sponsored a variety of investigations into language teaching and learning, with the goal of informing best practices. However, systematic principles of evaluation were not pursued in this project, nor was language education considered from a programmatic (i.e., designed, comprehensive) perspective.

² Indeed, one reviewer pointed out that theories of second language acquisition have yet to provide adequate accounts of the full scope of second language development across diverse skills and components, without which it remains difficult to know what the expected outcomes of comprehensive language education programs ought to look like over reasonable sequences of instruction.

³ At least in part, this state of affairs can be attributed to relative agnosticism on the part of theoreticians of second language acquisition, and associated prevailing research paradigms, vis-à-vis the design of language educational experiences—SLA theories are linguistic, cognitive, sociocultural, or even hybridized ‘sociocognitive’ but they are never educational (e.g., Hulstijn et al., 2014).

⁴ Ironically, as pointed out by one reviewer, the public media seems much more interested in publishing on evaluation, for example, in response to accountability evaluation schemes such as those associated with No

Child Left Behind in the U.S. or the National Assessment Program—Literacy and Numeracy in Australia.

⁵ http://nflrc-archives.ill.hawaii.edu/evaluation/biblio/biblio_all.cfm

⁶ To his credit, Rod Ellis, editor of *Language Teaching Research* through 2014, exhibited unique and sustained interest in publishing articles on language program evaluation.

REFERENCES

- Admiraal, W., Westhoff, G., & de Bot, K. (2006). Evaluation of bilingual secondary education in the Netherlands: Students' language proficiency in English. *Educational Research & Evaluation*, 12, 75–93.
- Alderson, J. C., & Beretta, A. (Eds.). (1992). *Evaluating second language education*. Cambridge: Cambridge University Press.
- American Evaluation Association (AEA). (2003). *American Evaluation Association response to US Department of Education*. Accessed 3 June 2015 at <http://www.eval.org/p/cm/ld/fid=95>
- American Evaluation Association (AEA). (2004). *American Evaluation Association guiding principles for evaluators*. Accessed 3 June 2015 at <http://www.eval.org/p/cm/ld/fid=51>
- American Evaluation Association (AEA). (2006). *Public statement: Educational accountability*. Accessed 3 June 2015 at <http://www.eval.org/p/cm/ld/fid=94>
- American Evaluation Association (AEA). (2011). *American Evaluation Association statement on cultural competence in evaluation*. Accessed 3 June 2015 at <http://www.eval.org/p/cm/ld/fid=92>
- American Evaluation Association (AEA). (2013). *An evaluation roadmap for a more effective government*. Accessed 3 June 2015 at <http://www.eval.org/d/do/472>
- Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2, 1–34.
- Badstübner, T., & Ecke, P. (2009). Student expectations, motivations, target language use, and perceived learning progress in a summer study abroad program in Germany. *Die Unterrichtspraxis [Teaching German]*, 42, 41–49.
- Bailey, A. L., & Huang, B. H. (2011). Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing*, 28, 343–365.
- Beretta, A. (1986a). A case for field-experimentation in program evaluation. *Language Learning*, 36, 295–309.
- Beretta, A. (1986b). Program-fair language teaching evaluation. *TESOL Quarterly*, 20, 431–445.
- Beretta, A. (1986c). Toward a methodology of ESL program evaluation. *TESOL Quarterly*, 20, 144–155.
- Beretta, A. (1990a). Implementation of the Bangalore Project. *Applied Linguistics*, 11, 321–340.
- Beretta, A. (1990b). The program evaluator: The ESL researcher without portfolio. *Applied Linguistics*, 11, 144–155.
- Beretta, A. (1992a). Evaluation of language education: An overview. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 15–24). Cambridge: Cambridge University Press.
- Beretta, A. (1992b). What can be learned from the Bangalore Evaluation. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 250–271). Cambridge: Cambridge University Press.
- Beretta, A., & Davies, A. (1985). Evaluation of the Bangalore project. *ELT Journal*, 39, 121–127.
- Birckbichler, D. W. (Ed.). (2006). *Evaluating foreign language programs: Content, context, change*. Columbus, OH: Foreign Language Center, the Ohio State University.
- Bloomfield, A. N., Ross, S. J., Masters, M. C., Gynther, K., & O'Connell, S. P. (2014). How does foreign language proficiency change over time? Results of data-mining official test records. In J. Connor-Linton & L. Amoroso (Eds.), *Measured language: Quantitative studies of acquisition, assessment, and variation* (pp. 199–212). Washington, DC: Georgetown University Press.
- Brindley, G. (1998). Outcomes-based assessment and reporting in language programs: A review of the issues. *Language Testing*, 15, 45–85.
- Brindley, G. (2001). Outcomes-based assessment in practice: Some examples and emerging insights. *Language Testing*, 18, 393–407.
- Brown, J. D. (1995a). *The elements of language curriculum: A systematic approach to program development*. New York: Heinle & Heinle Publishers.
- Brown, J. D. (1995b). Language program evaluation: Problems and solutions. *Annual Review of Applied Linguistics*, 15, 227–248.
- Burden, P. (2008). Does the use of end of semester evaluation forms represent teachers' views of teaching in a tertiary education context in Japan? *Teaching and Teacher Education*, 24, 1463–1475.
- Burger, D., & Ryan, J. (2007). Content assessment in the native language: Development of the Hawaiian Aligned Portfolio Assessment. *Pacific Educator*, Fall, 6–7.
- Bygate, M., Norris, J. M., & Van den Branden, K. (2015). Task-based language teaching. In C. Chapelle (Ed.), *Encyclopedia of applied linguistics*. Wiley–Blackwell.
- Cachey, T. (2014). Reframing assessment: Innovation and accountability between the global and the local. In J. M. Norris & N. Mills (Eds.), *Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 230–244). Boston: Cengage Learning.
- Cachey, T., & Pfeiffer, P. (2015). The uses of use-focused assessment: Two chairs' perspectives. In J. M. Norris & J. M. Davis (Eds.), *Student learning outcomes assessment in college foreign language programs*

- (pp. 37–56). Honolulu, HI: National Foreign Language Resource Center.
- Cai, S., & Zhu, W. (2012). The impact of an online learning community project on university Chinese as a foreign language students' motivation. *Foreign Language Annals*, 45, 307–329.
- Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research* (pp. 171–246). Boston: Houghton Mifflin.
- Carstens-Wickham, B. (2008). Assessment and foreign languages: A chair's perspective. *ADFL Bulletin*, 39, 36–43.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the test of English as a foreign language*. New York: Routledge.
- Chaudron, C., Doughty, C., Kim, Y., Kong, D., Lee, J., Lee, Y., Long, M. H., Rivers, R., & Urano, K. (2005). A task-based needs analysis of a tertiary Korean as a foreign language program. In M. H. Long (Ed.), *Second language needs analysis* (pp. 105–124). Cambridge: Cambridge University Press.
- Coleman, A. (1929). *The teaching of modern foreign languages in the United States: A report prepared for the modern foreign language study* (Vol. 12). New York: The Macmillan Company.
- Coombe, C., Al-Hamly, M., Davidson, P., & Troudi, S. (Eds.). (2007). *Evaluating teacher effectiveness in ESL/EFL contexts*. Ann Arbor, MI: The University of Michigan Press.
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S. (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass.
- Cronbach, L. J., & Suppes, P. (1969). *Research for tomorrow's schools: Disciplined inquiry for education*. London: Macmillan.
- Davis, J. M. (2015). The usefulness of accreditation-mandated outcomes assessment: Trends in university foreign language programs. In J. M. Norris & J. M. Davis (Eds.), *Student learning outcomes assessment in college foreign language programs* (pp. 1–35). Honolulu, HI: National Foreign Language Resource Center.
- Davis, J. M., Sinicrope, C., & Watanabe, Y. (2009). College foreign language program evaluation: Current practice, future directions. In J. M. Norris, J. M. Davis, C. Sinicrope, & Y. Watanabe (Eds.), *Toward useful program evaluation in college foreign language education* (pp. 209–226). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.
- Ecke, P., & Ganz, A. (2014). Student analytics and the longitudinal evaluation of language programs. In J. M. Norris & N. Mills (Eds.), *Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 62–82). Boston: Cengage Learning.
- Ekkens, K., & Winke, P. (2009). Evaluating workplace English language programs. *Language Assessment Quarterly*, 6, 265–287.
- Elder, C. (2009). Reconciling accountability and development needs in heritage language education: A communication challenge for the evaluation consultant. *Language Teaching Research*, 13, 15–33.
- Ewell, P. T. (2001). *Accreditation and student learning outcomes: A proposed point of departure*. CHEA Occasional Paper.
- Fetterman, D. M., Kaftarian, S., & Wandersman, A. (2014). *Empowerment evaluation: Knowledge and tools for self-assessment, evaluation capacity building, and accountability*. Thousand Oaks, CA: SAGE.
- Few, S. (2012). Big data, big ruse. *Visual Business Intelligence Newsletter* (July/August/September). Accessed 3 June 2015 at http://www.perceptualedge.com/articles/visual_business_intelligence/big_data_big_ruse.pdf
- Funnell, S. C., & Rogers, P. J. (2011). *Purposeful program theory: Effective use of theories of change and logic models*. San Francisco, CA: Jossey-Bass/Wiley.
- Gallardo del Puerto, F., & Gamboa, E. (2009). The evaluation of computer-mediated technology by second language teachers: Collaboration and interaction in CALL. *Educational Media International*, 46, 137–152.
- Godwin-Jones, R. (2014). Games in language learning: Opportunities and challenges. *Language Learning & Technology*, 18, 9–19.
- González-Lloret, M., & Nielson, K. B. (2014). Evaluating TBLT: The case of a task-based Spanish program. *Language Teaching Research*, 6.
- Gorsuch, G. (2009). Investigating second language learner self-efficacy and future expectancy of second language use for high-stakes program evaluation. *Foreign Language Annals*, 42, 505–540.
- Gottlieb, M., & Nguyen, D. (2007). *Assessment and accountability in language education programs: A guide for administrators and teachers*. Philadelphia: Caslon Publishing.
- Government of Canada. (2004). *Evaluation of the Language Instruction for Newcomers to Canada (LINC) program*. Accessed 3 June 2015 at <http://www.cic.gc.ca/english/resources/evaluation/linc/2004/index.asp>
- Grau Sempere, A., Mohn, M. C., & Pieroni, R. (2009). Improving educational effectiveness and promoting internal and external information-sharing through student learning outcomes assessment. In J. M. Norris, J. M. Davis, C. Sinicrope, & Y. Watanabe (Eds.), *Toward useful program evaluation in college foreign language education* (pp. 139–162). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.
- Harris, J. (2009). Late-stage refocusing of Irish-language programme evaluation: Maximizing the potential for productive debate and remediation. *Language Teaching Research*, 13, 55–76.

- Heining-Boynton, A. L. (1990). The development and testing of the FLES program evaluation inventory. *Modern Language Journal*, 74, 432–439.
- Hill, Y. Z., & Tschudi, S. (2008). A utilization-focused approach to the evaluation of a web-based hybrid conversational Mandarin program in a North American university. *Teaching English in China: CELEA Journal*, 31, 37–54.
- Holquist, M. (2010). Measuring the humanities: The slippery slope from assessment to standardization. In D. Heiland & L. J. Rosenthal (Eds.), *Literary study, measurement, and the sublime: Disciplinary assessment* (pp. 69–94). New York: Teagle Foundation.
- Horwitz, E. K. (1985). Formative evaluation of an experimental foreign-language class. *The Canadian Modern Language Review*, 42, 83–90.
- House, E. R. (2004). Intellectual history in evaluation. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 218–224). Thousand Oaks, CA: SAGE.
- House, E. R. (2005). Deliberative democratic evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 104–108). Thousand Oaks, CA: SAGE.
- Hulstijn, J. H., Young, R. F., Ortega, L., Bigelow, M., DeKeyser, R., & Ellis, N. C. (2014). Bridging the gap: Cognitive and social approaches to research in second language learning and teaching. *Studies in Second Language Acquisition*, 36, 361–421.
- Iwai, T., Kondo, K., Lim, D. S. J., Ray, G., Shimizu, H., & Brown, J. D. (1999). *Japanese language needs analysis 1998–1999 (NFLRC NetWork #13)*. Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.
- Johnson, R. K. (Ed.). (1989). *The second language curriculum*. Cambridge: Cambridge University Press.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–342.
- Kane, M. T. (2006). Validation. In Brennan R. (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.
- Kawakami, A., Aton, K., Cram, F., Lai, M., & Pormal, L. (2008). Improving the practice of evaluation through indigenous values and methods: Decolonizing evaluation practice—returning the gaze from Hawai'i and Aotearoa. In N. Smith & P. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 219–242). New York: Guilford Press.
- Kenyon, D. M., MacGregor, D., Li, D., & Cook, H. G. (2011). Issues in vertical scaling of a K–12 English language proficiency test. *Language Testing*, 28, 383–400.
- Kiely, R. (1998). Programme evaluation by teachers: Issues of policy and practice. In P. Rea-Dickins & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 78–104). London: Longman.
- Kiely, R. (2006). Evaluation, innovation, and ownership in language programs. *Modern Language Journal*, 90, 597–601.
- Kiely, R. (2009). Small answers to the big question: Learning from language programme evaluation. *Language Teaching Research*, 13, 99–116.
- Kiely, R., & Rea-Dickins, P. (2005). *Program evaluation in language education*. New York: Palgrave Macmillan.
- Kiely, R., & Rea-Dickins, P. (2009). Evaluation and learning in language programmes. In K. Knapp & B. Seidlhofer (Eds.), *Handbook of foreign language communication and learning* (pp. 663–694). New York: Mouton de Gruyter.
- Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice*, 22, 18–26.
- Koyalan, A. (2009). The evaluation of a self-access centre: A useful addition to class-based teaching? *System*, 37, 731–740.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: SAGE.
- Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the 'No Child Left Behind Act of 2001.' *Educational Researcher*, 31, 3–26.
- Liskin-Gasparro, J., & Vasseur, R. (2014). Designing an embedded outcomes assessment for Spanish majors: Literary interpretation and analysis. In J. M. Norris & N. Mills (Eds.), *Innovation and accountability in language program evaluation* (pp. 83–110). Boston: Cengage Learning.
- Llosa, L., & Slayton, J. (2009). Using program evaluation to inform and improve the education of young English language learners in US schools. *Language Teaching Research*, 13, 35–54.
- Long, M. H. (2005). Methodological issues in learner needs analysis. In M. H. Long (Ed.), *Second language needs analysis* (pp. 19–76). Cambridge: Cambridge University Press.
- Lynch, B. K. (1996). *Language program evaluation: Theory and practice*. Cambridge: Cambridge University Press.
- Ma, Q. (2008). Empirical CALL evaluation: The relationship between learning process and learning outcome. *CALICO Journal*, 26, 108–122.
- Mackay, R. (1994). Undertaking ESL/EFL programme review for accountability and improvement. *ELT Journal*, 48, 142–149.
- Mackay, R., Wellesley, S., & Bazergan, E. (1995). Participatory evaluation. *ELT Journal*, 49, 308–317.
- Mackay, R., Wellesley, S., Tasman, D., & Bazergan, E. (1998). Using institutional self-evaluation to promote the quality of language and communication training programmes. In P. Rea-Dickins & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 111–131). London: Longman.
- Malone, M., & Donovan, A. (2014, August). *The state of applied linguistics in the United States: Results of a nationwide survey*. Presentation at the annual conference of the International Association of Applied Linguistics (AILA), Brisbane, Australia.

- Martinez, A., & Sanz, C. (2008). Instructors' and administrators' beliefs within a Spanish LSP program. In H. J. Siskin (Ed.), *From thought to action: Exploring beliefs and outcomes in the foreign language program* (pp. 67–91). Boston: Heinle Cengage.
- Means, A. (2014). *The death of evaluation*. Accessed 3 June 2015 at <http://www.marketsforgood.org/the-death-of-evaluation/>
- Menken, K. (2008). *English learners left behind: Standardized testing as language policy*. Clevedon, UK: Multilingual Matters.
- Menken, K., & Solorza, C. (2014). No child left bilingual: Accountability and the elimination of bilingual education programs in New York City schools. *Educational Policy*, 28, 96–125.
- Mertens, D. M. (2009). *Transformative research and evaluation*. New York: Guilford Press.
- Mertens, D. M., Bledsoe, K., Sullivan, M., & Wilson, A. (2010). Utilization of mixed methods for transformative purposes. In C. Teddlie & A. Tashakkori (Eds.), *Handbook of mixed methods research* (2nd ed., pp. 193–214). Thousand Oaks, CA: SAGE.
- Mertens, D. M., & Wilson, A. T. (2012). *Program evaluation: Theory to practice*. New York: Guilford.
- Milleret, M., & Silveira, A. (2009). The role of evaluation in curriculum development and growth of the UNM Portuguese program. In J. M. Norris, J. M. Davis, C. Sinicrope, & Y. Watanabe (Eds.), *Toward useful program evaluation in college foreign language education* (pp. 57–82). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.
- Modern Language Journal. (2013). *Author Guidelines*. Accessed 3 June 2015 at [http://onlinelibrary.wiley.com/journal/10.1111/\(ISSN\)1540-4781/homepage/ForAuthors.html](http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1540-4781/homepage/ForAuthors.html)
- Nielson, K., & Doughty, C. (2008). *Rosetta Stone™ evaluation*. College Park, MD: Center for Advanced Study of Languages.
- No Child Left Behind Act of 2001, P.L. 107–110, 20 U.S.C. § 6319 (2002).
- Nocus, I., Guimard, P., Vernaudon, J., Paia, M., Cosnefroy, O., & Florin, A. (2012). Effectiveness of a heritage educational program for the acquisition of oral and written French and Tahitian in French Polynesia. *Teaching and Teacher Education*, 28, 21–31.
- Norris, J. M. (2006). The why (and how) of assessing student foreign language programs. *Modern Language Journal*, 90, 590–597.
- Norris, J. M. (2008). *Validity evaluation in language assessment*. New York: Peter Lang.
- Norris, J. M. (Ed.). (2009). Understanding and improving language education through program evaluation [Special issue]. *Language Teaching Research*, 13, 1–13.
- Norris, J. M. (2012). Meta-analysis. In C. Chapelle (Ed.), *Encyclopedia of applied linguistics*. Wiley–Blackwell.
- Norris, J. M., & Davis, J. M. (Eds.). (2015). *Student learning outcomes assessment in college foreign language programs*. Honolulu, HI: National Foreign Language Resource Center.
- Norris, J. M., Davis, J. M., Sinicrope, C., & Watanabe, Y. (Eds.). (2009). *Toward useful program evaluation in college foreign language education*. Honolulu, HI: National Foreign Language Resource Center.
- Norris, J. M., & Mills, N. (2014). *Innovation and accountability in foreign language program evaluation*. Boston: Cengage Learning.
- Norris, J. M., & Watanabe, Y. (2013). Program evaluation. In C. Chapelle (Ed.), *Encyclopedia of applied linguistics*. Wiley–Blackwell.
- Pardo-Ballester, C. (2012). CALL evaluation: Students' perception and use of LoMasTv. *CALICO Journal*, 29, 532–547.
- Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: SAGE.
- Peacock, M. (2009). The evaluation of foreign-language-teacher education programmes. *Language Teaching Research*, 13, 259–278.
- Pennington, M. C. (1991). *Building better English language programs: Perspectives on evaluation in ESL*. Washington, DC: NAFSA.
- Pfeiffer, P., & Byrnes, H. (2009). Curriculum, learning, and the identity of majors: A case study of program outcomes evaluation. In J. M. Norris, J. M. Davis, C. Sinicrope, & Y. Watanabe (Eds.), *Toward useful program evaluation in college foreign language education* (pp. 183–208). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.
- Pipi, K., Cram, F., Hawke, R., Hawke, S., Huriwai, T. M., Keefe, V., et al. (2003). *Māori and Iwi provider success: A research report of interviews with successful Iwi and Māori providers and government agencies*. Wellington: Te Puni Kokiri.
- Prabhu, N. S. (1987). *Second language pedagogy*. Oxford: Oxford University Press.
- Rea-Dickins, P., & Germaine, K. P. (1992). *Evaluation*. Oxford: Oxford University Press.
- Rea-Dickins, P., & Germaine, K. P. (Eds.). (1998). *Managing evaluation and innovation in language teaching: Building bridges*. London: Longman.
- Rhodes, N. C., & Pufahl, I. (2010). *Foreign language teaching in U.S. schools: Results of a national survey*. Washington, DC: Center for Applied Linguistics.
- Rockman et al. (2009). *Rosetta Stone Evaluation Report*. Accessed 11 June 2015 at <http://resources.rosettastone.com/CDN/us/pdfs/Rockman-Evaluation-Report.pdf>
- Rolstad, K., Mahoney, K., & Glass, G. V. (2005). The big picture: A meta-analysis of program effectiveness research on English language learners. *Educational Policy*, 19, 572–594.
- Rosenbusch, M. H. (2005). The No Child Left behind Act and teaching and learning languages in U.S. schools. *Modern Language Journal*, 89, 250–261.
- Ross, S. J. (2009). Language program evaluation. In M. Long & C. Doughty (Eds.), *The handbook of language teaching* (pp. 756–778). Malden, MA: Wiley–Blackwell.

- Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: SAGE.
- Sánchez-Meca, J. (1997). Methodological issues in the meta-evaluation of correctional treatment. In S. Redondo, V. Garrido, J. Pérez, & R. Barberet (Eds.), *Advances in psychology and law: International contributions* (pp. 486–498). Berlin: Walter de Gruyter.
- Scherer, G. A., & Wertheimer, M. (1964). *A psycholinguistic experiment in foreign-language teaching*. Columbus, OH: McGraw-Hill.
- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Chicago: Rand McNally.
- Scriven, M. (1969). An introduction to meta-evaluation. *Educational Products Report*, 2, 36–38.
- Smith, P. D. (1970). *A comparison of the cognitive and audiolingual approaches to foreign language instruction: The Pennsylvania foreign language project*. Center of Curriculum Development Incorporated.
- Stufflebeam, D. L. (1973). *Toward a science of educational evaluation*. Englewood Cliffs, NJ: Educational Technology Publications.
- Stufflebeam, D. L. (1999). *Evaluation plans and operations checklist*. Kalamazoo, MI: Evaluation Checklist Project, The Evaluation Center, Western Michigan University.
- Sundquist, J., & Neary-Sundquist, C. (2008). Student evaluations and teacher assessment: How much is the course a reflection of the teacher? In H. J. Siskin (Ed.), *From thought to action: Exploring beliefs and outcomes in the foreign language program* (pp. 245–260). Boston: Heinle Cengage.
- Tyler, R. (1948). The need for a more comprehensive formulation of theory of learning a second language. *Modern Language Journal*, 32, 559–567.
- Van Avermaet, P., & Gysen, S. (2006). *From needs to tasks: Language learning needs in a task-based approach*. In K. Van den Branden (Ed.), *Task-based language teaching in practice* (pp. 17–46). Cambridge: Cambridge University Press.
- Van den Branden, K. (2006). Training teachers: Task-based as well? In K. Van den Branden (Ed.), *Task-based language teaching in practice* (pp. 217–273). Cambridge: Cambridge University Press.
- Walther, I. (2009). Developing and implementing an evaluation of the foreign language requirement at Duke University. In J. M. Norris, J. M. Davis, C. Sinicrope, & Y. Watanabe (Eds.), *Toward useful program evaluation in college foreign language education* (pp. 119–140). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.
- Watanabe, Y., Norris, J. M., & González-Lloret, M. (2009). Identifying and responding to evaluation needs in college foreign language programs. In J. M. Norris, J. M. Davis, C. Sinicrope, & Y. Watanabe (Eds.), *Toward useful program evaluation in college foreign language education* (pp. 5–56). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.
- Weir, C., & Roberts, J. (1994). *Evaluation in ELT*. Oxford, UK: Blackwell.
- Windham, S. (2008). Redesigning lower-level curricula for learning outcomes: A case study. *ADFL Bulletin*, 39, 31–35.
- Xu, X., Padilla, A. M., Silva, D., & Masuda, N. (2012). A high school intensive summer Mandarin course: Program model and learner outcomes. *Foreign Language Annals*, 45, 622–638.
- Yang, W. (2009). Evaluation of teacher induction practices in a US university English language program: Towards useful evaluation. *Language Teaching Research*, 13, 77–98.
- Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: SAGE.
- Young, D. J. (2008). An empirical investigation of the effects of blended learning on student outcomes in a redesigned intensive Spanish course. *CALICO Journal*, 26, 160–181.
- Zannmirato, A. (2014). Promoting change in “two-tiered” departments: An exploratory evaluation of conflict and empowerment among language and literature faculty. In J. M. Norris & N. Mills (Eds.), *Innovation and accountability in language program evaluation* (pp. 154–182). Boston: Cengage Learning.