

# References on Evaluation in Language Education Programs

ANNOTATED BIBLIOGRAPHY · 2024

Assessment and Evaluation  
Language Resource Center

## About This Bibliography

This annotated bibliography can be a useful resource for anyone in the language education field, from educators to researchers, seeking familiarity with the major concepts of language program evaluation. Our goal is to provide broad coverage of the literature to inform researchers and practitioners, providing a knowledge base that can be further applied to actual learning and teaching of language. In this bibliography, we define program evaluation as follows:

Evaluation is the gathering of information about any of the variety of elements that constitute educational programs, for a variety of purposes that include primarily understanding, demonstrating, improving, and judging program value...

(Norris, 2006, p. 579)

As noted by Norris (2016) and Davis and McKay (2018), the evaluation process involves collecting and utilizing information with the purpose of understanding and improving language programs, rather than merely passing judgment on the program. Unlike research driven by the purpose of testing or generating theories, evaluation aims to answer empirical questions in the stakeholders' decision-making process (Norris, 2016). Therefore, the value of program evaluation is determined by how it aids stakeholders' understanding of the program and the extent to which this understanding can be applied to the actual use of the program (Patton, 2008).

This annotated bibliography was initially published in 2016 and is now updated to encompass studies published from 2016 to 2022. Included sources for the review consist of published peer-reviewed journal articles and book chapters. Dissertations, theses, conference presentations, and unpublished research were excluded from the review.

We defined language programs as “intentional interventions” (Norris, 2016, p. 169) that are designed to satisfy the language learning needs of individuals. Based on this definition, in this update, we have included various types of programs including second/foreign language programs, world language programs, bilingual education, immersion programs, heritage language programs, computer-assisted language learning programs, study abroad programs, task-based language teaching programs, English as a second/foreign language programs, and teacher education programs. This review specifically excluded works related to language assessment, assessment tools, pedagogy evaluation, review of studies, policy evaluation, general educational programs, curriculum evaluation, teacher classroom practice, critical reviews of language program development, and lesson plan evaluation.

Our search for the update included the following databases: Academic Search Premier (EBSCO), Google Scholar, Educational Resources Information Center (ERIC), Web of Science, and Linguistics and Language Behavior Abstracts (LLBA). We conducted searches using a combination of the terms “program evaluation” with “second language/ foreign language acquisition program,” “world language program,” “bilingual education,” “immersion program,” “heritage language program,” “computer-assisted language learning program,” “study abroad

program,” “task-based language teaching program,” “English as a second/foreign language program,” and “teacher education program.”

The work to update the bibliography was led by Yunjung (Yunie) Ku. We would like to express our sincere gratitude for the time and efforts that our research assistants, Elli Ahn, Quaylin Dang, Chanhee Park, Tianyi Ye, and Ian Timoteo Franza, dedicated to supporting this work.

Please cite the bibliography as follows:

AELRC. (2024). *References on evaluation in language education programs*. Assessment and Evaluation Language Resource Center. Retrieved from <https://aelrc.georgetown.edu/evaluation-bibliography>

# Table of Contents

*References on evaluation in language education programs*

*Assessment and Evaluation Language Resource Center*

## **Contents:**

About This Bibliography	2
Table of Contents	4
Reviews	5
References	146

## Reviews

**ADFL. (2009).** ADE guide for external reviewers. *ADFL Bulletin*, 40(2-3), 138–143.

*Keywords: Spanish HLLs; HLL assessment; placement test*

This essay provides guidance for external reviewers who are called on to facilitate program evaluations for foreign language departments. The guide shares insights and considerations regarding communication with the department, site visits, determining the purpose and use of the evaluation, data collection and analysis techniques, and reporting.

**ADFL. (2009).** Checklist for self-study for departments of foreign languages and literatures. *ADFL Bulletin*, 40(2-3), 144–153.

*Keywords: college, FL, program evaluation, checklist*

This comprehensive program self-assessment tool provides a checklist of questions for consideration in evaluating the following areas typical of the program review process: (a) the department and local community needs; (b) the department and the institution; (c) strategic goals; (d) priority setting; (e) departmental image; (f) curriculum planning; (g) course syllabi; (h) course articulation; (i) course placement; (j) course enrollment; (k) non-degree programs; (l) assessment; (m) faculty characteristics; (n) allocation of faculty workloads; (o) recognition of faculty achievement; (p) faculty recruitment; (q) faculty development; (r) faculty salaries and incentives; (s) faculty language proficiency; (t) instructional methods and approaches; (u) instructional quality; (v) student profiles and characteristics; (w) student recruitment, enrollment, retention; (23) student participation in the department; (24) extracurricular activities; (25) student advising; (26) study abroad; (27) departmental governance; (28) collective bargaining; (29) public relations; (30) support services; and (31) equipment and facilities.

**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(1), 75–93. <https://doi.org/10.1080/13803610500392160>

*Keywords: quasi-experimental, longitudinal, English, Dutch, bilingual programs, Netherlands, secondary education*

Since 1989, the number of Dutch-English bilingual secondary schools in the Netherlands has been growing substantially. Funded by the Dutch Ministry of Education, Admiraal, Westhoff, and de Bot conducted a six-year longitudinal comparative study of lower secondary learners' (12-15 year olds) English proficiency (vocabulary, pronunciation, reading and oral

ability), subject knowledge (history and geography), and Dutch ability in bilingual and non-bilingual schools. Researchers gathered data utilizing national standardized exit tests (Cito, MAVO, etc.) for intermediate secondary education, in addition to the EFL Vocabulary Test created by Meara (1992). Also included were learner background and attitudinal surveys. Results indicated that students in the bilingual schools outperformed the control group in reading, oral, and pronunciation tests. Vocabulary and subject knowledge were comparable across groups. Admiraal, Westhoff, and de Bot caution against overinterpretation of the study's results due to three factors: (a) the limited comprehensive dataset for Dutch and subjects tests, (b) the motivation factor of the students in the pioneer bilingual programs, and (c) the societal perception of English as a prominent language in Dutch society.

**Agustina, N. Q., & Mukhtaruddin, F. (2019).** The CIPP model-based evaluation on Integrated English Learning (IEL) program at language center. *English Language Teaching Educational Journal*, 2(1), 22–31. <https://doi.org/10.12928/eltej.v2i1.1043>

*Keywords: CIPP model, integrated English learning program, Indonesia*

The study evaluates an English language program called Integrated English Learning (IEL) developed by the Language Training Center (LTC) at Muhammadiyah University of Yogyakarta (UMY). The program covers all four skills of English (reading, listening, speaking, and writing) and is divided into 12 levels. The study investigates (a) students' and teachers' perceptions of the program and (b) how the program was implemented in the classroom. For evaluation, the study used the Context, Input, Process, and Product (CIPP) model developed by Stufflebeam (1971).

All enrolled students were asked to take a TOEIC-based test before starting the program. Based on the results, students were then assigned to one of the 12 proficiency levels. During and after the course, both formative and summative evaluations were conducted to provide necessary data to the curriculum developers.

To evaluate the program from multiple perspectives—including that of students, instructors, and staff members—the study based its evaluation on interviews with all the stakeholders. The participants included 25 third-year students, 15 English instructors, and the heads of the departments. The data was analyzed based on qualitative methods. The results are discussed in terms of (a) Context, which assessed the needs of the participants; (b) Input, which assessed alternative approaches to program development and teacher quality control; (c) Process, which evaluated the implementation of materials, teaching methods, and assessment; and (d) Product, which evaluated the outcomes of the program. The study concludes with the importance of program evaluation as a useful resource to improve the IEL program.

**Alderson, J. C. (1992).** Guidelines for the evaluation of language education. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 274–304). Cambridge University Press.

*Keywords: ESP, university, Brazil, perception data, survey, interview, journal, report, participatory evaluation*

Alderson reviews the issues of “who, what, when, how, how long, to evaluate and to point the way forward to further developments...in the methodology and practice of language education evaluation” (p. 274), as gleaned from the case studies in other chapters of the book. He cautions would-be evaluators that evaluation is reflexive, and it depends on the purpose, the nature of the program, the individuals involved (personalities and politics), time constraints, and available resources. Searching for perfect objectiveness in evaluation is not possible, since all stakeholders have different perspectives and values, and this reality must influence the design, implementation, and interpretation of any evaluation. Alderson presents a set of guiding questions for help in planning an evaluation study: (a) Why is evaluation required (Consider official and hidden agendas)? (b) Who is the evaluation for (Identify stakeholders’ purposes)? (c) Who is participating in the evaluation process? (d) What expertise is required as an evaluator? (e) What is the focus of the evaluation based on discussions and negotiations with the stakeholders? (f) How is the program evaluated (Adapt various methodologies and triangulate methods)? (g) When is evaluation to take place (Purposes of evaluation determine when to evaluate)? (h) How long should the evaluation last? And (I) What happens to an evaluation report (Agree on what is to be delivered by the evaluation study with stakeholders to ensure utilization)? During implementation, Critical Path Analysis (recognizing and stating key points and periods of time) may assist in deciding how much adaptability of plans is possible. To respond to various needs for interpretation and reporting, it is “important for the evaluator to devise ways in which the different interpretations of data that are both theoretically inevitable and practically and politically important can be gathered as part of the evaluation” (p. 296). In order to increase the possibility of utilization of the evaluation report, the evaluation should be relevant for the stakeholders, a result of negotiation with the stakeholders, based on adequate resources and feasible implementation, kept to a timeframe, adequately interpreted in terms of educational policy, and adequately reported.

**Alderson, J. C., & Beretta, A. (Eds.). (1992).** *Evaluating second language education*. Cambridge University Press.

*Keywords: bilingual education, EFL, ESP, case study, university, elementary, secondary, overview, framework, guideline, outsider, insider, participatory evaluation, political*

Alderson and Beretta’s edited collection of 10 chapters covers theoretical and methodological issues in language program evaluation, and presents case studies from a variety of contexts.

Chapters reflect the realities of program evaluation, highlighting conflicts and compromises called for at various decision points in the process. The book is comprised of three sections: (a) an overview article by Beretta, examining 25 years of previous evaluation studies in language teaching (since the 1960s); (b) eight case studies of current practice (by Alderson & Scott; Lynch; Mitchell; Palmer; Ross; Slimani; Coleman; Beretta, in chapter order), including summaries of evaluation findings and useful templates, instruments, etc., followed by post-script comments from the editors; and (c) guidelines by Alderson on the design of evaluation projects. The book is very practical for helping would-be evaluators learn and reflect on what previous evaluators' have experienced in the field. The range of language contexts featured in the book include English as a second/foreign language (primary, secondary, and college level), college level English for specific purposes, Gaelic-English bilingual education (primary level), and German (college level).

**Alderson, J. C., & Scott, M. (1992).** Insiders, outsiders and participatory evaluation. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 25-57). Cambridge University Press.

*Keywords: ESP, university, Brazil, perception data, survey, interview, journal, report, participatory evaluation*

Alderson and Scott report on the participatory nation-wide evaluation of English for Specific Purposes (ESP) courses in 45 universities in Brazil, focusing in particular on reading skills. The evaluation was called for after 7 years of development and implementation of the ESP project, for the purpose of continuation of funding from the Overseas Development Administration. With external and internal funding, this unique large-scale evaluation involved considerable time, cooperation, and manpower. The design of the evaluation, construction and piloting of instruments, collection of data, and drafting of the report were all done in a democratic manner involving project coordinators, teachers, research assistants, and a consultant (British Council consultant). The following factors were evaluated: context, methodology, implementation of methodology, project achievement, teacher-training implementation, and exchange of ideas and experience. This study gathered information from multiple source with multiple instruments to describe program outcomes. "Perception data" were collected from current ESP students, graduates of ESP, subject specialists/teachers, and ESP teachers. In addition, ESP student-reports of class discussions, ESP teacher-reports on the same discussions, ESP teachers' post-questionnaire interviews, and statistics on the use of the language center were collected for triangulation. However, the authors acknowledge problems in the sampling, questionnaire, analysis of qualitative data, absence of classroom observation, and lack of testing student outcomes. Despite these problems, the authors demonstrate how evaluation can empower and build the capacity of local stakeholders, including capacity to conduct future internal evaluations. As the authors mention, "Many, if not most, of the teachers involved also seemed to learn a great



deal about evaluation: how it might be planned, how data might be collected, and how results might be interpreted” (p. 52). Provided in the book’s appendix is an outline of this evaluation project’s proposal, instruments, and results, which may be useful as a reference for designing future language program evaluations.

**Allen, L. Q. (2010).** The impact of study abroad on the professional lives of world language teachers. *Foreign Language Annals*, 43, 93–104.  
<https://doi.org/10.1111/j.1944-9720.2010.01062.x>

*Keywords: French, FL, professional development, culture, study abroad, qualitative, essay*

This project brought 30 L2 French teachers to France for a three-week study abroad experience designed to build the participants’ cultural understandings, L2 language skills, and instructional skills through professional development for maximum impact. Activities included lectures, presentations, projects, papers, and field trips. Four months later, the participants were asked to provide written statements describing the impact of the program on their classroom practice and professional lives. The researcher coded the qualitative data for themes related to changes in L2 proficiency, cultural knowledge, classroom practice, and teachers’ professional lives. Findings showed that participation in the program resulted in increased confidence in speaking French, implementation of new cultural knowledge and artifacts in their teaching, and an expanded professional network for sharing what they learned with other teachers.

**Altstaedter, L. L., & Jones, B. (2009).** Motivating students’ foreign language and culture acquisition through web-based inquiry. *Foreign Language Annals*, 42, 640–657.  
<https://doi.org/10.1111/j.1944-9720.2009.01047.x>

*Keywords: Spanish, FL, college, quantitative, qualitative, WebQuest, perceptions, questionnaire, essay, project evaluation*

This article reports on the implementation and evaluation of a one-week-long WebQuest project completed by 14 L2 Spanish undergraduate students with Spanish proficiency levels rated between ACTFL novice-low and novice-mid. Two main outcomes were evaluated: (a) the impact of the web-based inquiry project on students’ ability perceptions regarding their L2 Spanish skills and knowledge of Hispanic culture (based on the Cultures and Comparisons curricular goals of the national Standards for Foreign Language Learning in the 21st Century), and (b) the appropriateness of the WebQuest task design with this population of L2 learners. The participants completed two Likert-scale questionnaires (one on their ability perceptions and motivation levels as a result of the project and one on their perceptions of the WebQuest task itself) and an essay synthesizing, reflecting on, and comparing cultural traditions. The results of

the questionnaires were analyzed quantitatively, while the essays were coded qualitatively for features of the Cultures and Comparisons curricular goals. Findings showed a statistical increase in students' L2 ability perceptions as a result of the project, and that the WebQuest task was appropriate and effective for learners at this ACTFL proficiency level.

**Anderson, J. (1998).** Managing and evaluating change: The case of teacher appraisal. In P. Rea-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 159–186). Longman.

*Keywords: Turkey, university, EFL, teacher appraisal, teacher training, management, meeting, discussion, questionnaire, observation, improvement*

Anderson, from the perspective of a management team member, documents the introduction, implementation, and evaluation of a teacher appraisal scheme (TAS) at Bilkent University School of English Language in Turkey. The purposes of appraisal were teacher accountability and professional development. There were three stages to the appraisal cycle: initial meeting for identification of teachers' interests, skills, and needs (Teacher Profile form); progress review meeting; and an end of the year meeting for reflection. The four classroom observations were linked with the targets identified in the Teacher Profile form. Anderson states some of the factors that affect the quality of teacher performance: teaching experience; qualifications; familiarity/understanding of the curriculum; colleagues; available resources; ability to evaluate, reflect, and change; motivation towards the job. During the implementation of the appraisal system, various tensions arose (the purpose of appraisal, ownership, cultural diversity of staff members, motivation, opportunity cost, reward for participation, operational issues, the continuity of ownership, and monitoring and quality control by the management team). An evaluation of the TAS was conducted after two years of implementation to verify and improve its quality. Based on the existing appraisal documents and data obtained from a half-day workshop for teachers (discussion, open-ended written response, and a questionnaire), the evaluation revealed that 50% of the teachers benefited from the TAS. It also identified the problem areas where TAS needed improvement. In the end, the appraisal system was refined and reworked based on the evaluation, showing the evolving process of the innovation.

**Arias, C. I., Maturana, L. M., & Restrepo, M. I. (2012).** Evaluation in foreign language learning: Towards fair and democratic practices. *Lenguaje*, 40(1), 99–126.

*Keywords: English as a second language instruction, English as a second language tests, action research*

The lack of congruence found between evaluation and pass/fail decisions in foreign languages was addressed in this interinstitutional action research project through the consolidation of a collaboratively designed evaluation system. The implementation of the system in English beginner courses of three adult extension programs revealed that a variety of evaluation types and forms; rigor and systematicity; and the meticulous design of evaluation instruments, forms, and tasks made evaluation practices fair and democratic and benefited students, teachers, and institutions.

**Arnold, N. (2009).** Online extensive reading for advanced foreign language learners: An evaluation study. *Foreign Language Annals*, 42(2), 340–366.  
<https://doi.org/10.1111/j.1944-9720.2009.01024.x>

*Keywords: US, university, German, reading, formative, process, product, qualitative, questionnaire, case study*

Arnold reports on an evaluation of an online extensive reading program implemented in an advanced German as a foreign language course at a US university. This pilot program incorporated certain modifications which distinguished it from traditional extensive reading programs. The purpose of the evaluation was to determine whether the program met its goals and investigate the effects of the modifications. In an effort to gain a deep understanding of the program and its effects, the evaluators focused on process as well as product and utilized qualitative data that illuminated the experiences of the students. Data collection instruments included student questionnaires, reading reports, and reflection journals. Based on the findings, the evaluators determined that the program could be implemented on a large scale, but also made suggestions for improvement. The evaluators note that the evaluation was limited to student perceptions and suggest adding tests to measure linguistic gain and longitudinal research to future evaluations.

**Atar, C., Kır-Cullen, E., & Denkci-Akkaş, F. (2020).** An evaluation of the English preparatory program at İstanbul Medeniyet University. *Language Teaching and Educational Research (LATER)*, 3(1), 94–115. <https://doi.org/10.35207/later.714457>

*Keywords: English preparatory program, mixed method approach, Turkey*

The study examines an English preparatory program implemented for college students before they entered the School of Foreign Languages at İstanbul Medeniyet University. Based on the mixed method approach, the study collected data from 131 students and four instructors through a questionnaire and semi-structured interviews. The quantitative data was analyzed using descriptive statistics, and the interview data was analyzed with descriptive analysis. The findings

revealed that generally, both students and instructors were satisfied with the program in terms of course materials, testing processes, assignments, and instructors. However, students showed dissatisfaction with certain courses and materials; students preferred materials developed by the Material and Curriculum Development Office over one-size-fits-all commercial materials. Other emerging concerns were related to online assignments due to technical issues. Additionally, the study also found that female students and students who voluntarily joined the program were more satisfied with the program than male students and students who attended the program mandatorily.

Based on the results, the study makes some recommendations including restructuring courses to better accommodate student interest and needs, as well as reconsidering several course materials. The study concludes with a suggestion for further research into motivation and compulsory language learning programs.

**Ataş, U., & Balıkçı, G. (2021).** Linking theory and practice in ELT program: An evaluation of materials, testing and practicum courses. *Journal of Narrative and Language Studies*, 9(17), 211–226. <https://nalans.com/index.php/nalans/article/view/345>

*Keywords: English language teaching program, teacher education, practicum course, Turkey*

Though there have been a myriad of studies on the effectiveness of teacher education programs in Turkey, few have focused on how theoretical and practicum courses are linked with one another. Mindful of this gap, the study investigates undergraduate English Language Teaching (ELT) programs, focusing on the connection between theoretical and practicum courses. A range of stakeholders including 17 pre-service teachers, six teacher educators, two mentor teachers from high schools, and three graduates participated in the study. Based on Peacock's (2009) foreign language teacher education model and Weir and Roberts' (1994) checklist of parameters, the study investigates the connection between two theoretical courses (Materials course, Testing course) and a practicum course (School Experience course). To do so, data were collected via an online questionnaire, semi-structured interviews, and document analysis.

The results were twofold. While the link between the Materials course and the Practicum course was strong enough, there was a lack of connection between the Testing course and actual testing practices in real school contexts. This, according to the authors, hinders pre-service teachers from applying assessment theories to real classrooms. To address this issue, the authors suggest collaboration between schools and faculty members in order for pre-service teachers to engage in actual testing and assessment practices.

**Bachman, L. F. (1989).** The development and use of criterion-referenced tests of language ability in language program evaluation. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 242–258). Cambridge University Press.

*Keywords: criterion-referenced testing, communicative language ability, proficiency scale*

Bachman is concerned with what learner outcomes to measure and how to use them in the evaluation of language programs. He points out the inadequacies of norm-referenced testing in addressing the needs of program evaluation, and he notes inadequacies in definitions of language proficiency. In formative evaluation, Bachman suggests that identifying specific instructional objectives and gathering students' information on achievement are necessary. In summative evaluation, Bachman argues, information is needed on both stated and unexpected outcomes that are consistent with the broader goals of educational systems and society. For program evaluation in general, he advocates testing that “involves the combination of the criterion-referenced approach to test development with a current specification of the domain of language proficiency” (p. 251), which he call ‘communicative language ability’ (language competence, strategic competence, and psychophysiological skills). To satisfy comparability across programs, abstract proficiency scales independent of contextual features of language use should be defined. Bachman also stresses the need to empirically test this framework.

**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.  
<https://doi.org/10.1111/j.1756-1221.2009.00034.x>

*Keywords: German, FL, college, study abroad, questionnaire, quantitative, Language Contact Profile*

This article describes the impact of a month-long study abroad experience for 23 German L2 university students in terms of goals and motives for enrollment, expected gains in L2 proficiency and cultural knowledge, and L2 use. Data were collected using a pre- and post-study abroad questionnaire based on the Language Contact Profile. The questionnaire used Likert-scale items for participants' expectations for progress in various language skills, L2 use, and proficiency self-ratings. In comparing students' pre-study abroad expectations with their post-study abroad experiences, quantitative findings demonstrated that expectations were higher than their perceived progress in all skill areas except for cultural learning. Moderate correlations were found between self-reported L2 use and L2 proficiency.

**Barr, D., Leakey, J., & Ranchoux, A. (2005).** Told like it is! An evaluation of an integrated oral development pilot project. *Language Learning & Technology*, 9(3), 55–78.  
<https://www.learntechlib.org/p/74478/>

*Keywords: Canada, CMC, classroom, computer, oral, French, testing, questionnaire, journal*

Barr, Leakey and Ranchoux conducted a methods comparison study between computer-mediated communication (CMC) and face-to-face instruction of a conversation class. They were interested in: (a) whether computer technology enhances progress in students' oral language development; (b) the factors that may affect students' oral language development when using computers; and (c) staff and students' reactions to using computer technology in conversation classes. Four groups of 5 to 11 university students (29 total, Arts students as control and Applied Languages students as treatment) participated in French conversation classes one hour per week for 12 weeks. Data collection included a language experience questionnaire, information-and-communications-technology-use questionnaire (use of email and the web), student journal logs (self-assessing their linguistic development and the class), and pre-and post-test (composed of a pronunciation task, personal questions, a listening comprehension exercise, and an oral résumé of a television documentary). Findings indicated that students in the traditional classroom setting did better than the students in the CMC environment. Some class time was spent getting used to the computer-based environment and software, thus the amount of content covered in the treatment/control classes differed. Based on qualitative findings, CMC students did appreciate the individualized opportunities for practicing pronunciation, but rated discussions and debates as the best aspect of the oral classes (i.e., the parts requiring the least technology). Tutors in CMC felt that technology had a dehumanizing effect on oral classes. Results were inconclusive regarding the role of technology in CMC oral communication classes, since the use of technology was limited to oral drills and not applied to meaningful communication.

**Barreneche, G. I. (2011).** Language learners as teachers: Integrating service-learning and the advanced language course. *Hispania*, 94, 103–120.  
<https://doi.org/10.1353/hpn.2011.0028>

*Keywords: Spanish, FL, college, community service learning, case study, qualitative*

This essay describes a service-learning project in a local Hispanic community for L2 Spanish undergraduates that was conducted in partnership with a Junior Achievement program. The paper provides an overview of service learning as an innovative pedagogy in L2 instruction, with direct connections made to course and curricular learning goals in L2 higher education. It then describes the development and implementation of the service-learning project associated with an advanced Spanish L2 course offered at the university, which placed participants in elementary schools as Spanish tutors. The data that contributed to this case study included participants'

reaction papers, pre-service reflection, and post-service essay/analysis on the experience, as well as evaluation feedback from the cooperating teachers who supervised the participants in the elementary schools. All of the participants' data was collected in their L2 Spanish, which provided additional information about the impact of the project on the participants' linguistic abilities. The case study results conclude with a comprehensive discussion of the positive effects of service-learning on L2 Spanish learners, as well as the logistical and program implementation challenges associated with the project, and the limitations inherent in attempting to quantify learning outcomes as a result of the project. The reflection prompts for the participants are included in the appendix.

**Bayram, I., & Canaran, Ö. (2019).** Evaluation of an English preparatory program at a Turkish foundation university. *Journal of Language and Linguistic Studies*, 15(1), 48–69. <https://www.researchgate.net/publication/332136972>

*Keywords: English preparatory program, program evaluation, focus group interviews, Turkey*

The study explores an English preparatory program offered by a university in Turkey. The evaluation was focused on (a) the strengths and weaknesses of the program, (b) the differences between students' and teachers' perceptions of the program, and (c) their suggestions for improving the program. 241 students and 26 teachers answered the survey that was developed and pilot tested by the researchers. Additionally, focus group interviews were conducted to triangulate the quantitative data.

The results revealed that from the student's perspective, the strengths of the program were homework and exams, while from the teachers' perspective, the strongest point was performance assessment. Both viewed course books and supplementary booklets as up-to-date but at the same time, they commented that reading and listening exercises needed some improvements. In terms of weaknesses, both the students and teachers viewed extracurricular activities and online programs as components that should be improved more. More in-depth interviews also showed that students found the pacing of the program too fast as they were having trouble understanding the lesson and catching up on homework and speaking practices. Based on the results, the authors make suggestions such as reducing the workload of the program and increasing the opportunities to do reading and listening exercises.

**Beretta, A. (1986a).** A case for field-experimentation in program evaluation. *Language Learning*, 36(3), 295–309. <https://doi.org/10.1111/j.1467-1770.1986.tb00557.x>

*Keywords: experimental, field research*

Beretta discusses the limitations of adapting a laboratory research methodology to language teaching program evaluation, arguing instead for what he calls “field-experimentation”. Field-research is a “long-term, classroom-based inquiry into the effect of complete programs, the degree of control being partly dependent on whether correlational or experimental information is sought” (p. 296). Since field research is concerned with the generalizability of the findings to classroom contexts (external validity), he suggests that field research will provide findings that are relevant to immediate pedagogic concerns. Nevertheless, the evaluator has to keep in mind that the choice of methodology depends on the purpose of the evaluation study, the kind of questions stakeholders pose, the feasibility of the methodology, and availability of resources for evaluation.

**Beretta, A. (1986b).** Program-fair language teaching evaluation. *TESOL Quarterly*, 20, 431–445.

*Keywords: norm-referenced testing, criterion-referenced testing, program fair testing, bias, validity*

Beretta gives examples of how non-program-fair tests can favor one teaching methodology over another. Previous program comparison studies on the effectiveness of teaching methodologies have been critiqued due to the use of content-biased tests; the use of standardized testing inappropriate for actual levels of students; and the use of instruments to insensitive actual features of effectiveness. Beretta suggests the use of criterion-referenced (program specific) tests in order to raise awareness and sensitivity to the features of the program, and to strengthen the content validity of the assessment. He also cautions against test bias in judging the outcomes of educational effectiveness, though it is unclear on what basis the bias might be identified.

**Beretta, A. (1986c).** Toward a methodology of ESL program evaluation. *TESOL Quarterly*, 20(1), 144–155. <https://doi.org/10.2307/3586398>

*Keywords: method, experimental, contextual factors, causality*

Beretta points out how rigorous experimental methods do not fit language teaching program evaluation studies from the perspective of external validity. Rather than pursuing causality, he advocates an applied inquiry in which outcomes will be more directly relevant for pedagogy. He suggests that “(a) we conduct our investigations in the field rather than in artificially controlled “laboratory” settings, (b) we consider the effect of total programs rather than isolated components of them, (c) the duration of the studies be long-term rather than short-term, (d) randomization is not always practicable or crucial” (p. 145). Note that Beretta’s view may conflict with the recent expectation for rigorous research methodology announced by the U.S.



Department of Education Institute of Educational Sciences, which prioritizes randomized controlled trial research as the gold standard .

**Beretta, A. (1990a).** Implementation of the Bangalore Project. *Applied Linguistics*, 11(4), 321–340.

*Keywords: India, EFL, implementation, retrospective, questionnaire, specialist, teachers, Communicational Teaching Project*

Beretta reports on evaluation of pedagogical implementations in the Bangalore Project (also known as the Communicational Teaching Project or CTP), through the use of retrospective protocols (asking teachers to describe how they taught). The protocols of 15 teachers' (4 regular public school teachers (RT) and 11 non-regular highly-qualified teachers (NRT)) were coded into three pedagogic implementation levels: (a) orientation (not fully aware of the CTP), (b) routine (operating comfortably with the CTP), and (c) renewal (seeking ways to improve the CTP). Beretta concludes that a different sense of 'ownership' of the CTP was found between the RTs and the NRTs, which was reflected in the different implementation levels. The RTs acted as if they were new to CTP, while the NRTs seemed more comfortable with CTP as routine practice. Because the study was conducted following implementation, no classroom observations were included; it is, thus, hard to say what the frame of reference was when the teachers answered the survey. No other stakeholders of the project were included in the survey, so only a partial view of implementation was possible. The study's main contribution may be to raise language program evaluator's awareness of how important it is to evaluate not only the product but also the implementation process of a program.

**Beretta, A. (1990b).** The program evaluator: The ESL researcher without portfolio. *Applied Linguistics*, 11, 144–155.

*Keywords: dialogue, negotiation, user-oriented, utilization, policy*

This article reflects the acknowledgement in ESL program evaluation of major issues discussed in the broader field of educational evaluation: the importance of negotiation (dialogue) between evaluator and the stakeholders, utilization of evaluation, localization (contextualization) of evaluation, and the examination of policies attached to evaluation. Beretta highlights conflicts the evaluator may face between the policy-shaping community, which makes pragmatic decisions, and scholars, who judge standards of research. In order to shape a user-oriented evaluation, he suggests the following to be fully transparent and falsifiable: (a) obtain a hearing, (b) identify the policy-shaping community, (c) negotiate reasonable research questions, (d) design and collect data, and (e) communicate the findings.

**Beretta, A. (1992a).** Evaluation of language education: An overview. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 15–24). Cambridge University Press.

*Keywords: overview; model; quantitative; qualitative; history*

This chapter provides a comprehensive overview of the trends and models in educational program evaluation since the early “behavioral objectives approach”, which “compare[s] intended outcomes with actual outcomes” (p. 13). Large-scale evaluation studies in the 1960s and 1970s tended to compare one program to another using an experimental approach, but they were inconclusive, due to methodological design flaws (internal consistency, comparison groups, randomization, etc.). Moving from an inadequate dichotomy of quantitative vs. qualitative, a more eclectic/pragmatic philosophy emerged in response to various program evaluation purposes. Professional evaluation standards developed in the 1980s recognized “the heterogeneity of evaluation needs and approaches,” which led to “the four attributes for evaluation: utility, feasibility, propriety and accuracy” (p. 18).

Beretta summarizes what has been learned and developed in the field of education, indicating future directions for evaluation in language education: (a) the most appropriate evaluation methods should be chosen according to what the audience (policy-shaping community) wants to know; (b) in program evaluation, user-relevant information should precede the advance of language learning theory; (c) evaluation should be considered from the outset in the design of the program; (d) step one should involve the negotiation of aims for evaluation, prioritization of questions in terms of the capacity of the program evaluation (time, cost, learnability, impact), and translation of policy questions into evaluation questions; (e) based on information needs and deadlines of clients, appropriate methodologies should be adopted; (f) findings should be translated back into language of policy, and different forms of reporting should accommodate different audiences.

**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 University Press.

*Keywords: India; EFL; Bangalore Project; Communicational teaching project; method comparison; conflict*

Evaluation of the Bangalore Project was reported by Beretta in a variety of publications (see Beretta, 1990 and Beretta & Davies, 1985, for detailed explanation of the Bangalore project, its aim, and evaluation findings). Here, Beretta offers his retrospective “if I had known then” account of the evaluation activities he conducted. He highlights the needs to negotiate and clarify the purpose, methods, and specific information to be collected, during the planning stage. He

also articulates a rationale for why intended use of evaluation outcomes should be identified prior to evaluation design and data collection. However, he points to the fact that external evaluators often do not have adequate time to negotiate plans with program stakeholders and develop the instruments needed. Since the evaluation took place towards the end of the project, information was scattered or unavailable, and it was not possible for Beretta to obtain a rich/thick description of the project. Furthermore, the values of a few select stakeholders led to much of the decision making about evaluation, before it commenced. Evaluation should ideally be integrated within curriculum and with actual program context, but in the real-world, Beretta cautions that the evaluator may have to deal with a situation where there are severe limitations to planning, resources, data, and the like.

**Beretta, A., & Davies, A. (1985).** Evaluation of the Bangalore Project. *ELT Journal*, 39(2), 121–127. <https://doi.org/10.1093/elt/39.2.121>

*Keywords: India; EFL; testing; proficiency; achievement; experimental; Communicational Teaching Project; ODA*

Beretta and Davies report on the evaluation of the Bangalore Project (also known as Communicational Teaching Project (CTP)), an accountability driven evaluation led by the Overseas Development Administration. The evaluation took place towards the end of the project by an external evaluator. The purpose of the evaluation was to compare the CTP methodology (task-based) with the traditional (structural) methodology to determine its effectiveness. Due to constraints imposed by the principal investigator of the CTP project, the evaluation utilized a quasi-experimental approach to evaluation. However, the authors recognized the difficulty of adopting a rigorous design in educational contexts. Students from four schools (each with one experimental and one control class) took both achievement tests (structure test and task-based) and proficiency tests (contextualized grammar, dictation, and listening/reading comprehension). The proficiency tests were administered as neutral measures to overcome any test content bias. The validity of this experimentation was questioned because of the instability of the educational context, the lack of reference points for comparing the two groups, and the absence of the description and observation of classroom practices to justify the differences between the two groups.

**Bernhardt, E. B. (2006).** Student learning outcomes as professional development and public relations. *Modern Language Journal*, 90(4), 588-590.  
[https://doi.org/10.1111/j.1540-4781.2006.00466\\_5.x](https://doi.org/10.1111/j.1540-4781.2006.00466_5.x)

*Keywords: assessment; multiple languages; OPI; SOPI; course evaluation; university*

Bernhardt reports on the successful expansion of the Stanford Language Center as a direct result of the increased visibility of student learning outcomes assessment and student course evaluations. In 1996, the Stanford Language Center began implementing SOPI assessments at the end of the language requirement to evaluate the current program outcomes for the language learners. In addition to the assessment data, course evaluations were gathered to reveal students' satisfaction with their language courses. Two factors were associated with efforts in implementing an assessment system: (a) a cultural change in understanding learner performance assessment, and (b) the need for teacher capacity building via OPI certification training. Learner outcomes and positive course evaluations supported the language center's assessment efforts. They also provided justification for the university to allocate additional funds for teacher development and assessment programs as well as increased staffing and the addition of a professional reward system for instructors. Systematic documentation and publishing of the results provided a strong justification for the efficacy of the language programs offered via Stanford Language Center.

**Bernhardt, E. B. (2008).** Assessment as a keystone for language and literature programs. *ADFL Bulletin*, 40(1), 14–19.

*Keywords: US; university; foreign language; assessment; systemic; systematic; SOPI; oral proficiency*

Bernhardt's essay advocates for "systemic and systematic student assessment" (p.14) as a means of strengthening university foreign language programs and helping them integrate language and literature. She describes how the Stanford Language Center uses the SOPI (Simulated Oral Proficiency Interview) to assess all students at the end of their first year of language study. The SOPI provides the center with data for internal program evaluations, as well as clearly defined outcomes useful for presenting to those outside the program and meeting demands for accountability. Moreover, the ability to demonstrate student improvement using nationally recognized standards has strengthened the position of the language center within the university and facilitated its funding requests. Bernhardt asserts that the development of a systemic assessment procedure can also help programs struggling to unify literature and language because the process of articulating outcomes requires departments to analyze expectations for both language and literature skills.

**Bernhardt, E., & Brillantes, M. (2014).** The development, management, and costs of a large-scale foreign language assessment program. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 41–61). Cengage Learning.

*Keywords: monetary and nonmonetary investment; assessment program, program evaluation via proficiency assessment*

Elizabeth Bernhardt and Monica Brillantes outline the monetary and nonmonetary costs of the Stanford Language Center’s assessment program across two years of instruction in 15 languages. Their chapter emphasizes the relationship between the monetary investment in the assessment program and student performance outcomes. From the perspectives of both the director and the program manager of the language center, the authors discuss the scope of the assessment program, the assessment instruments employed, and the impact that program evaluation via proficiency assessment has had on the language center’s status within the university.

**Birckbichler, D. W. (Ed.). (2006).** *Evaluating foreign language programs: Content, context, change*. Foreign Language Center, the Ohio State University.

*Keywords: guidelines; framework; ethnography; communication; stakeholders; focus group; observation; interview; proficiency testing; reporting; multiple languages; university*

The edited book provides practical guidelines for conducting foreign language program evaluation at the postsecondary level. The chapters are divided into three parts: framing the evaluation (Chapters 1-3), asking the right questions (Chapters 4-6), and reporting for change (Chapters 7-8). Each chapter offers guidelines and example tools for different steps within the program evaluation process.

Part One of the book offers suggestions for setting the stage for program evaluation. In chapter one, Costner gives a bird’s eye views of different program evaluation approaches that are applicable to foreign language education. He proposes a content-specific approach to evaluation, where evaluators obtain specific knowledge of foreign language education (e.g., knowledge of the language taught) and reflect program culture in evaluation design. In Chapter Two, Kawamura advocates an ethnographic approach to evaluation, which requires thick descriptions of various program elements, a balance between emic (program internal) and etic (program external) perspectives, and data collection from multiple sources and levels. To unveil a program’s culture, Kawamura describes two ethnographic data collection methodologies: participant observation and ethnographic interview. In Chapter Three, “Communication: An essential tool in program evaluation,” Lang reminds evaluators that communication is essential to obtain buy-in, support, and cooperation from program stakeholders. To make the evaluation process transparent and to maintain professional distance with the stakeholders, evaluators need

to use various communication tools and strategies, such as an email list, a website with plans and updates, etc.

Part Two covers planning data collection methodologies for program evaluation. Kawamura, Dassier, and Costner approach data collection with stakeholder participation and collaboration in mind. In the three methodology chapters, Kawamura, Dassier, and Costner outline reflective questions in defining the framework and scope of data collection methodology. Focus groups and proficiency testing are covered in detail in Chapters Five and Six, respectively. In planning for a focus group session, carefully sequenced questions, moderator skills, and logistics (e.g., site, time, recording) are important elements to consider. For data analysis, thematic concept mapping and key word identification can be conducted to find common and contradicting ideas that emerge from the transcript (see examples in the chapter). In Chapter Six, Dassier suggests making use of the Proficiency Guidelines and the Standards articulated by the American Council on the Teaching of Foreign Languages for framing what to test. She warns test developers to contemplate validity, reliability, and practicality of the test and offers a practical check-list for choosing and developing appropriate tests.

The last two chapters focus on what to do after data interpretation with the stakeholders. In Chapter Seven, Lang explains six rules-of-thumb for reporting: (a) include a short and concise executive summary; (b) state a clear rationale for evaluation design, instruments, and interpretation; (c) contextualize the program under evaluation; (d) format the report effectively; (e) consider a variety of different forms to report findings (e.g., website); and (f) be tactful and fair. Chapter Seven also contains a detailed list of what to include in an evaluation report. In the last chapter, Birckbichler emphasizes the importance of using evaluation findings for taking programmatic action as well as the ongoing cyclical nature of program evaluation.

**Brantmeier, C., Vanderplank, R., & Strube, M. (2012).** What about me?: Individual self-assessment by skill and level of language instruction. *System*, 40(1), 144–160.  
<https://doi.org/10.1016/j.system.2012.01.003>

*Keywords: Advanced, intermediate, and beginning language learners; Self-assessment; Language program assessment*

In an investigation with advanced language learners, Brantmeier (2006) reports that self-assessment (SA) of second language (L2) reading ability, when measured with self-rated scales, is not an accurate predictor of subsequent reading performance as measured via multiple choice items. In another experiment with advanced learners that utilizes criterion-referenced SA items, Brantmeier and Vanderplank (2008) reveal that learners accurately estimate their reading comprehension when it is measured via multiple choice items. For the present study, an SA instrument of language learning achievement was designed according to specific course content to take into consideration the direct experience learners have had in practicing reading, listening,

speaking, and writing. With 276 participants, the study examines skill-based SA across beginning, intermediate and advanced levels of language instruction, and it offers evidence to validate the relationship between the SA instrument and achievement on an online abilities test with advanced learners. Findings hold important implications for language learner assessment, especially in terms of advanced students' ability to rate themselves when given specific criteria. A discussion about the value of SA as a complement to other traditional approaches for language program assessment is offered.

**Brindley, G. (1998).** Outcomes-based assessment and reporting in language programs: A review of the issues. *Language Testing*, 15(1), 45–85.  
<https://doi.org/10.1177/026553229801500103>

*Keywords:* Australia; England; outcomes-based assessment; summative

Outcomes-based assessment relates summative classroom-external assessment with classroom-based learning assessment as a way of responding to different stakeholders who need to understand what students achieve in terms of well-defined learning outcomes. This article outlines problems the past assessment practices have encountered and suggests strategies for developing and implementing outcomes-based reporting. Brindley suggests: (a) collect a comprehensive range of information; (b) hold a dialogue with stakeholders to clarify the purpose of the assessment reform, identify actual information needs, and increase ownership of the reform; (c) research the relationship between outcome statements and assessment tasks; (d) link assessment tasks to levels of achievement by training teachers and creating banks of assessment tasks; (e) consult a variety of reporting methods for use among key stakeholders so that the educational value of reported information will not be reduced; (f) use multiple raters, scoring, and sources to overcome variability in judgments of performance; and (g) provide ongoing support for teachers, with continual open review and critique of the assessment process. Outcomes based on benchmarks (exemplary student performances) may evolve hand-in-hand with what students are actually doing in their coursework. However, once standardized testing is introduced as a criterion for determining student achievement, there is a danger of standardizing instruction as well.

**Brindley, G. (2001).** Outcomes-based assessment in practice: Some examples and emerging insights. *Language Testing*, 18(4), 393–407.  
<https://doi.org/10.1177/026553220101800405>

*Keywords:* Australia; adult immigrant education; outcomes-based assessment; teachers

Reporting on adult immigrant education and schooling in Australia, Brindley illustrates challenges that arise in outcomes-based assessment. The two major problem areas are: (a) the collision between political and educational perspectives, and (b) the quality (validity and reliability) of teacher-constructed assessment tasks. The first problem derives from typical governmental emphasis on system accountability rather than on learning, leading education authorities to narrow the curriculum (focusing on students' minimum competencies rather than on a more challenging curriculum). The latter problem derives from the lack of adequate teacher training in outcomes-based assessments and principles of good assessment practice. However, on-going projects, such as creating a well-researched task-bank that consistently reflects different levels of achievement, developing professional task development guidelines, and appropriate professional development will assist teachers in using performance-based assessments.

**Brown, A. (2014).** Foreign language course grades as pre-requisites and programmatic gatekeepers. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 183–207). Cengage Learning.

*Keywords: role of course grades*

Alan Brown questions the ambiguous role of course grades as a means of producing valuable information on student learning outcomes. Brown, who serves as the director of undergraduate studies, first presents descriptive statistics on the use of grade-based metrics as prerequisites for undergraduate courses at 73 research-intensive public institutions. He then reports the results of a correlational study of the relationship between student grades from two pre-major Spanish courses at one institution and students' respective reading, speaking, and writing proficiency assessments. He concludes the chapter with a discussion of the role of course grades in the evaluation of FL students' capabilities.

**Brown, J. D. (1989).** Language program evaluation: A synthesis of existing possibilities. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 222–241). Cambridge University Press.

*Keywords: model; history; approaches; dimensions; product; process; static characteristic; decision facilitation; curriculum process model*

Brown defines evaluation as “the systematic collection and analysis of all relevant information necessary to promote the improvement of a curriculum, and assess its effectiveness and efficiency, as well as the participants' attitudes within the context of the particular institutions involved” (p. 223). He reviews the approaches and dimensions of evaluation from the past and



proposes a systematic approach to evaluation integrated in curriculum design. During the forty years of development in educational psychology, four approaches emerged, each one building on the previous: (a) product oriented approach, which measures whether program goals and instructional objectives are met; (b) static characteristic approach, conducted by outside experts and describing the nature of programs; (c) process oriented approach, in which evaluation is used for improvement of curriculum (formative purpose) and also for judging the program (summative purposes); (d) decision facilitation approach, which feeds information into decision making. The three dimensions of evaluation (formative vs. summative, product vs. process, quantitative vs. qualitative) and the available approaches formulate initial decisions about evaluation procedures. A ‘systematic approach,’ Brown proposes, has six components for designing and maintaining language curriculum, from conducting needs analysis, to setting goals and objectives, testing, developing materials, and teaching, all interconnected with ongoing (formative) evaluation from the beginning to the end (summative) of curriculum development.

**Brown, J. D. (1995a).** *The elements of language curriculum: A systematic approach to program development.* Heinle & Heinle.

*Keywords: ESL; EFL; model; curriculum development; needs analysis; goals and objectives; tests; materials; teaching effectiveness; efficiency; attitude; China; US*

In the final chapter of this book, Brown discusses language program evaluation as part of ongoing curriculum development. His systematic approach to designing and maintaining language curriculum (also see Brown, 1989; Pennington & Brown, 1991) posits evaluation as a component that can “utilize all the information gathered in the processes of (a) developing objectives; (b) writing and using the tests; (c) adopting, developing, or adapting materials; and (d) teaching” (p. 24) for improving curriculum. The quality of each program component can be analyzed in terms of effectiveness, efficiency, and attitude. Brown also offers an overview of evaluation approaches (goal-attainment, static-characteristic, process-oriented, and decision-facilitation) and three dimensions that shape the perspective taken on evaluation (formative vs. summative; process vs. product; and quantitative vs. qualitative). In adopting the viewpoint framework, appropriate data sources and evaluative questions are outlined. Brown concludes with insights from program evaluation projects at the Guangzhou English Language Center in China, illustrating how his framework was fine-tuned to the context.

**Brown, J. D. (1995b).** Language program evaluation: Decisions, problems and solutions. *Annual Review of Applied Linguistics*, 15, 227–248.

*Keywords: second language; foreign language; evaluation decisions; positivistic; interpretive; quantitative; qualitative; overview*

Brown defines program evaluation as a systematic collection and analysis of information necessary to improve a curriculum, assess its effectiveness and its efficiency, and determine participants' attitudes within the context of a particular institution (p. 227). He reviews the field of program evaluation, specifically the work on second and foreign language programs between 1986 and 1994, extending Beretta's (1992) survey of the methodology of evaluation implemented between 1967 and 1985. Brown also outlines some decisions and problems previous evaluators considered or encountered when planning evaluation. There are six types of decisions evaluators have to make: (a) What is the purpose of the evaluation to judge the success (summative evaluation at the end) and/or to improve (formative evaluation, during) the program; (b) What amount and type of expertise are necessary to bring outside experts for credibility, to involve all stakeholders in the program, or to have insiders and outsiders work together (participatory model); (c) What form will the evaluation take field research (long-term, classroom-based) and/or laboratory research (short-term, test-based); (d) When is the evaluation performed during or after the program, or both, and for how long; (e) What type of data is collected and for what purpose quantitative data and/or qualitative data (interview, observation, journal, correspondence); (f) Is there a need to know the process involved in the curriculum and/or resulting products? When conducting evaluation, Brown cautions evaluators to consider (a) sampling, (b) teacher effects, (c) test practice effect, (d) Hawthorne effect, (e) reliability of the instruments, (f) finding valid program-fair instruments, (g) politics involved in evaluation, and (h) other potential problems.

**Brown, J. D. (2008).** Testing-context analysis: Assessment is just another part of language curriculum development. *Language Assessment Quarterly*, 5(4), 275–312.  
<https://doi.org/10.1080/15434300802457455>

*Keywords: validity; testing; stakeholder; context; utility; relevance; consequence*

Brown advocates conducting testing-context analysis in order to create a stakeholder-friendly and ultimately a defensible test. Resonating with Messick's idea of unitary validity, the author advocates that defensible testing will require an empirical study of not only the traditional notions of construct, content, and criterion-related test validity, but also the context in which the test is nested. The testing-context analysis involves 12 steps, starting from defining the purpose of the test (e.g., proficiency, diagnostic) to deciding on directions for ongoing research to "[understand] the validity issues of relevance, utility, values implications, and social consequences" (p. 302). Brown views testing as part of integrated curriculum development, and emphasizes the importance of stakeholder (including the test users) perspectives in test validity analysis, because test constructs are socially-constructed and situated in the context where tests are used. The testing-context analysis treats testing as a program, and the steps described align with program evaluation procedures.

**Brown, J. D., & Pennington, M. C. (1991).** Developing effective evaluation systems for language programs. In M. C. Pennington (Ed.), *Building better English language Programs: Perspectives on evaluation in ESL* (pp. 3–18). NAFSA.

*Keywords: systematic; participatory model; stakeholders*

Brown and Pennington view evaluation as “a process of determining the value of the individual aspects of an organization as a basis for ongoing change and development within that organization” (p. 4). They argue for “program evaluation to be a team effort involving many different personalities and varied input into the review process from others, both within and outside the program” (p. 13) rather than a top-down process. A participatory model enhances the possibility that evaluation will be more responsive to local factors, and it contributes to the growth of “self-determination and professionalism” (p. 15) of teachers as well as the positive evolution of the organization. The authors list the required conditions for fair and effective evaluation to occur: (a) gather information from multiple resources; (b) use different types of instruments; (c) all stakeholders understand evaluation as an ongoing process; (d) all stakeholders understand evaluation criteria and processes and the their link with the philosophy and goal of the program; (e) administrators believe in the productivity of interaction with the instructors and the leadership of education; and (f) all stakeholders see evaluation as a means for achieving balance between administrative control and individual autonomy.

**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. <https://doi.org/10.1016/j.tate.2007.11.012>

*Keywords: EFL; Japan; university; course evaluation; teacher perception; formative; accountability*

A recent higher education reform in Japan required Japanese universities to implement mandatory student evaluation-of-teaching surveys (SETs; i.e., end-of-semester course evaluation surveys) as a way to respond to accountability demands and improve educational quality. The study explores university English language teachers’ (ELTs) perceptions towards SET. Sixteen local and expatriate ELTs holding either tenured, fixed-contract, or part-time positions at private Japanese universities participated in interviews. Interviewees viewed SET as perfunctory, questioned its credibility and utility, articulated the limited views of teaching in SETs, and reported a lack of both buy-in and follow-through. Burden suggests that teaching evaluation and improvement should be participatory, reflective, dialogic, and constructive.

**Burden, R., & Williams, M. (1996).** Evaluation as an aid to innovation in foreign language teaching: The “SPARE” wheel model. *Language Learning Journal*, 13, 51–54.

*Keywords: French-English; bilingual program; elementary education; illuminative; monitoring; innovation*

Using an evaluation study of a French-English bilingual program as a backdrop, Burden and Williams recommend incorporating evaluation from the onset of a program’s innovation to better inform program design and strengthen program implementation. Building on Parlett’s (1981) illuminative evaluation approach, the authors describe an evaluator as an “interpreter of complex systems” (p. 51), who illuminates program contexts, perceptions of key program players, and processes of a program’s implementation. To evaluate innovation, a five-step SPARE wheel model is proposed—examine program Setting, formulate realistic program Plans, monitor and record Actions in place, assist the participants to Review actions and outcomes, and Evaluate the gathered information to make program decisions.

**Burke, B. M. (2012).** Experiential professional development: Improving world language pedagogy inside Spanish classrooms. *Hispania*, 95, 714–733.

*Keywords: Spanish; FL; professional development; communicative language teaching; teachers; questionnaires; observations; qualitative; secondary*

This article describes the implementation and evaluation of an experiential professional development (EPD) course provided to four high school Spanish teachers interested in increasing the number of communicative language teaching (CLT) activities in their instruction. The EPD lasted 10 weeks and included implementation of new communicative activities, observations, professional discussions, and reflective activities. Data collection included teacher pre-EPD and post-EPD questionnaires, written reflections, field notes, and observation notes. The qualitative analysis of these data included the constant comparison method and member checks, with short narratives depicting teacher change in implementing the new communicative activities, in order to address how the EPD affected the teachers’ understandings of CLT and their instructional decisions. Appendices include the teacher questionnaires.

**Byrd, P., & Constantinides, J., C. (1991).** Self-study and self regulation for ESL programs: Issues arising from the associational approach. In M. C. Pennington (Ed.), *Building better English language programs* (pp. 19–35). NAFSA.

*Keywords: self-study; participatory model; administrator; NAFSA; TESOL; professional organization; force field analysis*

Byrd and Constantinides describe the history of how NAFSA (for higher education) and TESOL (for all levels) developed similar approaches to influencing institutions internally through the use of self-study. Self-study is a study focusing on a single aspect of the program, continued cyclically but with evolving emphases through continuous data collection. It entails adjustment to changes in the outside environment, review and evaluation of current program practice as part of planning, and the tuning of external with internal evaluation. In order to choose the right design for self-study, ‘force field’ analysis can facilitate the process. This practice uncovers, both within and outside the program, all forces that are for versus against program self-study and change. Self-study helps the program to: (a) clarify goals; (b) identify existing problems; (c) enable learning about the program, procedures, and resource; and (d) identify and produce needed changes. It can also be used to familiarize new administrators with the program, build better understanding of the ESL program within a larger institution by participating in the institution’s accreditation reviews, provide information for external reviews, and provide ESL program staff a better understanding of the parent institution.

**Byrnes, H. (2008).** Owing up to ownership of foreign language program outcomes assessment. *ADFL Bulletin*, 39(2&3), 28–30.

*Keywords: university; US; foreign language; outcomes; assessment; ownership; transformative evaluation*

While outcomes assessment has traditionally been viewed negatively by many university foreign language departments, Byrnes argues that departments can take ownership of the process and use it to improve and strengthen their programs. Byrnes traces the professional discussion of this issue and recounts three important contributions toward the reconceptualization of outcomes assessment: John Norris’ 2006 ACTFL presentation on the transformative potential of outcomes assessment; the 2006 Modern Language Journal’s Perspectives column, which explores the why and how of outcomes assessment; and the University of Hawaii National Foreign Language Research Center’s 2007 summer institute on foreign language program evaluation for university faculty. Byrnes also introduces two case studies that contribute to the professional discussion. Both are examples of university foreign language programs taking ownership of the outcomes assessment process: Windham’s (2007) study of Elon University’s efforts to align their foreign language curriculum with the ACTFL guidelines; and Carstens-Wickam’s (2007) account of the role assessment played in departmental improvements at Southern Illinois University-Edwardsville.

**Cachey, T. (2014).** Reframing Assessment: Innovation and accountability between the global and the local. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 230–244). Cengage Learning.

*Keywords: outcomes-oriented assessment; useful program evaluation; chairperson perspective*

Theodore Cachey, the chairperson in the Department of Romance Languages and Literatures at the University of Notre Dame, describes the inception and sustainment of an outcomes-oriented assessment project. He discusses how the process of evaluation led to curricular revisions and innovations, enhanced collegiality and respect among faculty, and the reframing of goals and instructional practices within the department. Cachey argues that useful program evaluation may help language and literature departments achieve the reform outlined in the MLA Ad Hoc Committee on Foreign Languages (2007) report “Foreign languages and higher education: New structures for a changed world” as well as encourage faculty to become advocates for the value and importance of the study of foreign languages, cultures, and literatures within the humanities.

**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. <https://doi.org/10.1111/j.1944-9720.2012.01204.x>

*Keywords: CALL; motivation; quantitative; qualitative; college; Chinese; FL; questionnaires*

This report examines the impact that participating in an online learning community had on the language learning motivations of 44 university students of L2 Chinese. The researchers used Dörnyei’s framework of the L2 motivational self system to study how this online intervention affected the students’ motivation, as measured by responses on pre- and post-intervention Likert-item questionnaires. These quantitative findings were supplemented with students’ responses to open-ended questions about their perceptions of the online project itself, pointing specifically to motivating and demotivating features of their online participation. The article reports on the main activities of the students’ participation in the online learning community and the qualitative coding of students’ perceptions as themes emerged from the questionnaire findings. Findings point to the effectiveness of an online learning community for accessing resources, interacting with other Chinese language learners, and increasing the students’ levels of motivation related to L2 learning experience. Demotivating features of participation in the project were related to class organization, technology barriers, limited time, and the students’ workload in completing the project. Participation in the project was not found to affect the ideal L2 self or ought-to L2 self aspects of Dörnyei’s motivational self system.

**Cañado, M. L. P. (2017).** CLIL teacher education: Where do we stand and where do we need to go? In M. E. Gómez Parra, R. Johnstone (Eds.), *Bilingual education: Educational trends and key competences* (pp. 129–144). Ministerio de Educación, Cultura y Deporte. <https://www.researchgate.net/publication/320298606>

*Keywords: content and language integrated learning program, teacher education, Spain*

Cañado conducts a large-scale evaluation of the CLIL (Content and Language Integrated Learning) program, in a monolingual context in Spain. The major concern of the evaluation can be organized into three categories: (a) the training needs of teachers, students, and parents in three different monolingual communities in Spain (Andalusia, Extremadura, and the Canary Islands), (b) differences between the perceptions of the three cohorts (teachers, students, and parents), and (c) differences within the cohorts of teachers, students, and parents in terms of age, gender, type of school, and the level of studies.

The study employs triangulation designs that include multiple sources of information using multiple data collection methods (questionnaires and interviews) across multiple sites (primary and secondary schools in three different communities). The data obtained from the questionnaires were analyzed statistically while data coding, memoing, and conclusion drawing were employed to identify emerging patterns in the interview data.

Results showed that first, linguistic and methodological competence has been emphasized to a great degree, especially for non-linguistic area teachers and teachers with a B2+ level. Second, when it comes to scientific knowledge and collaborative competence, the results implied that teachers have made progress in familiarizing themselves with the knowledge of CLIL. However, reflective and developmental competence still needs to be addressed further as teachers were not participating in specific CLIL training. Based on these results, the authors suggest pedagogical decisions that are grounded in empirical evidence when designing and implementing courses.

**Carsten-Wickham, B (2008).** Assessment and foreign languages: A chair's perspective. *ADFL Bulletin*, 39(2&3), 36–43.

*Keywords: US; university; foreign language; outcomes; assessment; NCATE; teacher education; proficiency; standards; internal review; program improvement; case study*

Carsten-Wickham reports on the improvements made in the Department of Foreign Language and Literature (FLL) at Southern Illinois University-Edwardsville as a result of the department's participation in the NCATE process and an internal program review. Although focused on foreign language teacher education, the NCATE process revealed areas for improvement that benefited the entire foreign language program. During the process, the department reassessed its goals and objectives, brought its curriculum in line with national standards, and identified

necessary curricular changes. The process also helped the department identify study abroad as an area for further emphasis and development. Furthermore, because of NCATE and the internal review, the department was able to demonstrate the need for a state-of-the-art foreign language training center and obtain staffing support for the center from the university. Although originally wary of program assessments such as the NCATE, faculty in the FLL now see them as useful tools for program improvement.

**Celen, K. M., & Akcan, S. (2017).** Evaluation of an ELT practicum programme from the perspectives of supervisors, student teachers and graduates. *Journal of Teacher Education and Educators*, 6(3), 251–274. <https://eric.ed.gov/?id=EJ1207287>

*Keywords: English language teacher education, mixed method design, Turkey*

The study examines the English Language Teacher Education practicum in one state university in Turkey. The focus of the evaluation was on the strengths and weaknesses of the program, as well as the needs and problems related to learners' experiences. Based on the mixed method design, the authors collected data from senior students, university supervisors, and graduate students through Williams-Pettway's (2005) Survey of Teacher Education Programs (STEP) and focus group interviews.

The findings revealed that student teachers' awareness of cultural diversity and their technology skills were found to need more improvement, while the strengths of the program included the quality of instruction, student-teacher internship, and field experiences. In particular, there seemed to be a good relationship between cooperating teachers and student teachers. To further enhance the effectiveness of the teacher education program, the study suggested giving incentives to cooperating teachers so that they can focus on assisting the student teachers and providing students with more chances of field practice prior to student teaching.

**Cellante, D., & Donne, V. (2013).** A program evaluation process to meet the needs of English language learners. *Education*, 134(1), 1–8.

*Keywords: program evaluation, English language learners, teacher preparation*

Cellante and Donne used the process of program evaluation to evaluate the education program, provide information to make decisions on its ability to comply with mandates from the state education department, and develop or improve the program to meet the goals of the new initiative to meet the needs of English language learners. The program evaluation process began with a curricular review, included several evaluation measures, and was implemented across a four-year time span to gain state approval. The resulting evaluating process is conducted at a



small and private university and will aid other institutions to design a plan for implementing innovative programs.

**Chanda, C. (2008).** *Teaching and learning of English in secondary schools: A Zambian case study in improving quality.* Commonwealth Secretariat.

*Keywords: English; FL; quantitative; qualitative; questionnaires; focus group; interview; observation; instructional quality*

This book is based on Chanda's dissertation regarding the current nature of teaching and learning processes, constraints, and quality of educational outcomes in English language programs in Zambia. The author focuses on four primary areas: (1) commonly used instructional activities and strategies; (2) effectiveness of these activities and strategies; (3) limitations and constraints for using these strategies; and (4) suggestions for improving the teaching and learning of English in secondary programs. Data collected included teacher and student questionnaires, focus groups and interviews, and classroom observations. Of particular interest is chapter 4, "Ways of improving classroom teaching and learning," which is based on Chanda's field work and points to a variety of resource and logistical constraints affecting instructional quality, while also providing recommendations for program and curricular improvements.

**Chang, J.-Y., Kim, W., & Lee, H. (2017).** A language support program for English-medium instruction courses: Its development and evaluation in an EFL setting. *International Journal of Bilingual Education and Bilingualism*, 20(5), 510–528.  
<https://doi.org/10.1080/13670050.2015.1080658>

*Keywords: English-medium instruction courses, Korea, academic language skills*

Following the Korean government's language policy, an increasing number of universities have required undergraduate students to take English Medium Instruction (EMI) courses.

Chang et al. (2017) evaluates the program developed by the Center for Teaching and Learning (CTL) at Seoul National University for Korean students with no previous EMI experience, allowing them to learn the necessary knowledge and skills that are expected to enhance their academic language skills. The program consisted of workshops and tutoring attended by 330 students, and data were collected via questionnaires, pre- and post-tutoring surveys, tutors' weekly reports, and final evaluations. The findings showed that in general, the program seemed to be helpful for students' preparedness for taking EMI courses. Although the students' self-assessment did not improve significantly, the students still found the program helpful as it increased their self-confidence. To improve the program, the study suggests that the program be

tailored to individual majors and degrees. Most importantly, they suggest that a supportive environment is essential for students to succeed.

**Chase, G. (2006).** Focusing on learning: Reframing our roles. *Modern Language Journal*, 90(4), 583–588. [https://doi.org/10.1111/j.1540-4781.2006.00466\\_3.x](https://doi.org/10.1111/j.1540-4781.2006.00466_3.x)

*Keywords: Student learning outcomes assessment; cross-disciplinary; stakeholders; roles; responsibilities; university*

Chase agrees with and extends Norris's (2006) claim that a much needed conceptual shift in assessment practice (I.e., viewing assessment as an opportunity to support student learning rather than simply responding to accountability pressures) cuts across diverse disciplines in higher education. Chase advocates that different stakeholders of a department or program reconceptualize their roles and responsibilities to create a learning community of professionals and learners. Faculty members should actively engage in and commit to understanding and improving student learning to reflect each faculty member's unique strengths and views. Administrators also have an obligation to integrate and promote student learning as a core organizational goal, as well as to provide support structures for faculty development. And to create funding models that integrate student learning as a key component. Finally, Chase argues that students also need to take initiative in their own learning and provide feedback on their learning outcomes to the program.

**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 University Press.

*Keywords: US; Korean; foreign language; university; task-based language teaching; needs analysis; materials development; module development; utilization*

Chaudron et al describe the first stage (Needs Analysis) of a three-year, federally funded project on task-based teaching of Korean as a foreign language (KFL), conducted at the University of Hawaii at Manoa. The chapter is a detailed demonstration of how task-based needs analysis can be carried out and utilized for creating prototypical task-based instruction. First, unstructured interviews were conducted with instructors and a stratified random sample of students enrolled in KFL courses. These sought to obtain demographic information, reasons for studying Korean and for going to Korea, current and anticipated future Korean uses, language skills students expect to need, and necessary task performance abilities for language use in Korea and for future jobs. Based on the interviews, a questionnaire was formulated and administered to the entire KFL

student population at the university. Target tasks were identified from the survey results. For demonstration purposes, two target tasks were chosen (asking for directions and shopping for clothes); these were used to collect target discourse samples in the U.S. and in Korea. The process of identifying prototypical discourse structures within the samples is explained, and extensive discourse excerpts provided. The last section of the chapter demonstrates the application of findings for developing task-based language teaching modules, which consist of pedagogic tasks based on the two targets, and which increase in complexity. Included are a sample needs analysis questionnaire, consent form, and the model task-based language teaching module.

**Coleman, H. (1992).** Moving the goalposts: Project evaluation in practice. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 222–246). Cambridge University Press.

*Keywords: EFL; Indonesia; university; ODA; British Council; conflict resolution; evaluator's role; political*

Coleman provides an insider view on evaluation history and the related conflicts he experienced through developing the Key English Language Teaching (KELT) Project at Hasanuddin University, Indonesia. The project was funded by the Overseas Development Administration and overseen by the British Council. Coleman, as an incoming project officer not knowing beforehand the actual status of the project, encountered conflicting objectives and aims from different parties, each having a different expectation for evaluation. The original purpose of the project was to prepare materials and courses for pre-departure staff designated for overseas training. However, a needs analysis led to changes in emphasis towards (a) the attempt to modify undergraduates attitudes towards English, (b) the training of English teaching staff, and (c) the eventual creation of a course called Risking Fun. Coleman examined “the extent to which [the Project] achieved the objectives which it laid down for itself after it had begun to operate” (pp. 236-237). When it comes to evaluating a project where different objectives are expected and evolve over time, it is clear that interpretations about the project may differ dramatically. It falls to the evaluator to identify various objectives as they change over time, and to document how and why they changed and how they were achieved.

**Coombe, C., Al-Hamly, M., Davidson, P., & Troudi, S. (Eds.). (2007).** *Evaluating teacher effectiveness in ESL/EFL contexts*. The University of Michigan Press.

*Keywords: teacher evaluation; guidelines; standards; NCATE; performance indicators; self-evaluation; observation, portfolio; teacher training; ESL/EFL; K-12; university; adult education*

The book introduces various teacher evaluation practices in English as a foreign/second language programs around the globe. The 15 chapters are organized into four sections: teacher evaluation standards (Part One), case studies (Part Two), research in teacher evaluation (Part Three), and resources and tools (Part Four).

Part One covers two projects on teacher/professional standards in-depth: the development and implementation of professional standards for newly qualified teachers of English in Egypt, and the use of TESOL/NCATE standards as a resource for teachers to build professional independence in South Asia.

Part Two of the book compiles six case studies of teacher evaluation practices and guiding principles across diverse contexts, from a comprehensive NCATE accreditation review of an education department in the U.S. to teacher appraisal programs taking teacher-driven and collaborative approaches in Canada and the United Arab Emirates. Many chapters include example performance indicators and standards and descriptions of methodologies taken in each context.

Part Three of the book showcases four research projects that investigate various issues and practices of teacher evaluation, including (a) an action research study of adult ESL teachers' understanding and practices of multiliteracy; (b) a survey study of university students' perception of the usefulness, use, and focus of teacher evaluation; (c) a two-year study investigating development and training of teacher effectiveness from multiple perspectives (teachers, students, and trainers); (d) a survey study of teachers' attitudes towards three teacher evaluation methods (student evaluations of teaching, classroom observations by the administrators, and teacher portfolios).

The final part of the book examines tools for assessing teacher effectiveness, including a self-evaluation tool for pre-service teachers, a standards-based classroom observation tool, and a district-wide teaching portfolio assessment tool for formative purposes. Teacher trainers and administrators who conduct teacher evaluation will benefit from the book by reviewing the practical guidelines and ample examples of methods situated in various educational contexts.

**Dassier, J. P., & Powell, W. (2001).** *Formative foreign language program evaluation: Dare to find out how good you really are. Dimension 2001: The odyssey continues.* Selected proceedings of the 2001 Conference of the Southern Conference on Language Teaching, Birmingham, AL, 15-30.

*Keywords: US; university; foreign language; Spanish; French; language requirement; intrinsically motivated; questionnaire; focus group; proficiency; improvement; testing*

Formative evaluation is a constructive process to “form the foundation for decision-making on central issues of curriculum development and thus more effectively address the kinds of issues raised in the introductory scenarios” (p. 93). Dassier and Powell report on a two-year formative evaluation study of college-level French and Spanish language (four semesters) requirement courses at the University of Southern Mississippi (USM). The evaluation was initiated by faculty members (I.e., non-mandated) to “provide information, substantiate or reject intuitions, assess program impact, quality, and effectiveness,” (p.97) and to convince administrators (the chair of the department, the dean of their college, and others) of the value of such study. All students enrolled in second-semester and fourth-semester French and Spanish courses completed a questionnaire (demographic and attitudinal questions) and a College-Level Examination Program (CLEP) test. A spoken language proficiency test was also administered in the first year of the study, but not for the second year, due to cost. Students who took the CLEP test were randomly selected to volunteer in focus-group discussions to collect rich perspectives on issues that surround the required courses. Data suggested the need to: (a) create a system to identify true-beginners from false-beginners (require placement exams, review high school records, enhance advising, and/or create another 100-level class for true-beginners); (b) fill in the study gaps between high school, community college, and university language programs through cooperative articulation; (c) communicate curricular objectives and philosophies between USM and the schools students came from (external articulation); (d) clarify and articulate goals and curricular framework with greater coordination and internal consistency among instructors. The lack of validity of the CLEP-test, due to students’ low investment (“buy-in” effect), suggested the need to integrate an assessment mechanism with the existing curriculum for future on-going evaluation. This study demonstrates how intrinsically-motivated, on-going evaluation can lead to proposals for program improvements based on rich information; the article does not report on subsequent implementation of change.

**Davis, J. M., & McKay, T. H. (2018).** *A guide to useful evaluation of language programs.* Georgetown University Press. <https://doi.org/10.2307/j.ctvvngs5>

*Keywords: utilization-focused framework, practitioners, educators, administrators*

The book is a practical guide that helps practitioners, both educators and administrators, without little to no experience to conduct language program evaluation in a way that can lead them to a better evidence-based decision making. The book is organized around the steps that evaluators need to take when evaluating language programs. Before guiding the readers through the full range of stages of the evaluation, Chapter 2 introduces the major concepts and best practices of program evaluation. The key success factors of the evaluation include stakeholder involvement, effective communication, high-quality practices, and ample evaluation resources, among others.

Chapter 3 draws upon Patton's (2008) utilization-focused framework that consists of three major components of planning: identifying users of the evaluation, identifying evaluation uses of the information, and creating evaluation questions. Based on these components, the chapter introduces the practical details of utilization-focused evaluation. Chapter 4 concerns the identification of program indicators, which precedes the data collection procedure. Identifying indicators is especially important as they ensure evaluators collect the right types of information that answers the identified questions. Chapter 5 discusses the selection of data collection methods and suggests strategies that ensure the collection of useful information that suits stakeholders' needs.

The next three chapters (Chapter 6, 7, and 8) discuss three major data collection strategies: focus groups, interviews, and questionnaires, respectively. Each chapter consists of four parts: the purpose of the method, the strengths of the method, the data collection procedure, and best practices of the method.

Next, Chapter 9 concerns analyzing the data collected from questionnaires and focus groups. The authors provide guidance on how to analyze and interpret both numerical questionnaire data and textual information from focus groups. Additionally, the chapter covers threats to accuracy and trustworthiness of data.

Finally, Chapter 10 summarizes the key points from all the previous chapters, while Chapter 11 is an example of an evaluation plan that readers can consult. The example consists of the description of the plan, project users, indicators of the program, data collection methods, and the timeline of the project.

**Ebtesam, E., & Foster, S. (2019).** Implementation of CIPP model for quality evaluation at Zawia University. *International Journal of Applied Linguistics and English Literature*, 8(5), 106–115. <https://doi.org/10.7575/aiac.ijalel.v.8n.5p.106>

*Keywords: CIPP model, English language teaching program, sequential explanatory design, Libya*

In Libya, English education has been excluded from school curricula due to the intense political tensions between the Libyan government and Western countries, especially the United Kingdom. The study investigates the English language teaching (ELT) program at Zawia University in Libya, focusing on whether the program had ever been updated or validated. The study uses a sequential explanatory design with multiple data collection methods including questionnaires and semi-structured interviews. The quantitative data was analyzed using SPSS while the qualitative data was analyzed using content analysis.

The findings revealed that the English program was not validated and there seemed to be more weaknesses than strengths in terms of program delivery, teaching resources, language skills, and assessment of student work. To illustrate, the teaching was centered around the teachers, and English was taught mainly through the grammar-translation method. Furthermore, data from interviews revealed that many of the teaching materials were outdated and the classroom was lacking technological support. The assessment was also based on reading and writing skills, excluding the examination of students' speaking skills. Based on the findings, the study develops a framework that aims to improve the quality of the English language program.

**Ecke, P., & Ganz, A. (2014).** Student analytics and the longitudinal evaluation of language programs. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 62–82). Cengage Learning.

*Keywords: longitudinal evaluation; challenges of language program evaluation, German*

Peter Ecke and Alexander Ganz address the use of a variety of student analytics and longitudinal data to evaluate the German language program at the University of Arizona. Ecke, the director of the basic German language program, and Ganz, a doctoral student in Transcultural German studies, describe how the careful analysis of institutional and external data on student enrollments, language program offerings, and student profiles, accompanied by various internal student surveys, have helped the department respond to key challenges faced over the past decade.

**Efeoglu, G., Ilerten, F., & Basal, A. (2018).** A utilization focused evaluation of the preparatory school of an ELT program. *International Online Journal of Educational Sciences*, 10(4), 1309–2707. <https://doi.org/10.15345/iojes.2018.04.008>

*Keywords: utilization focused evaluation, English teacher preparatory program, English language teaching, Turkey*

The study explores an English teacher preparatory program, specifically the reading, grammar, and listening courses, provided by an English Language Teaching (ELT) Department at a state university in Turkey using Patton's (2008) utilization-focused evaluation (UFE). The study investigates participants' opinions of the courses offered through the ELT program, participants' suggestions for improving the courses, participants' perceptions of the effectiveness of the courses over two years, and whether the UFE framework applied to the first year's program had any effect on the following programs. A total of 38 students from a preparatory school of an English Language Teaching Department participated in the study. Based on the descriptive mixed

method design, the study collected data longitudinally over two subsequent years (2013-2014 academic years and 2014-2015 academic years) from multiple sources, including open-ended questionnaires, one-on-one interviews, and nominal group technique. The data from the three different sources were transcribed, categorized, and analyzed quantitatively.

The results of the first program evaluation were distributed to all stakeholders. Accordingly, some changes were made to make the program more tailored to students' needs. The second program evaluation revealed that some aspects of the program were improved (e.g., changing the course book) while there still remained some issues that needed to be addressed (e.g., technological support). The study concludes by highlighting the significance of follow-up evaluations that can compensate for the weaknesses of the otherwise effective UFE framework and ultimately lead the program to be more student-oriented.

**Ekkens, K., & Winke, P. (2009).** Evaluating workplace English language programs. *Language Assessment Quarterly*, 6, 265–287. <https://doi.org/10.1080/15434300903063038>

*Keywords: adult; ESL; workplace; standardized assessment; alternative assessment; learning journal; quantitative; qualitative*

This article begins with an overview of workplace ESL program models and standardized assessments commonly used to justify the funding of these types of programs. The purpose of the study was to evaluate the extent to which standardized assessments measure workplace ESL student learning, as compared to the learning measured by students' written journals after 30 hours of instruction. Twenty-one learners were recruited to participate in the study from three workplace sites offering a 10-week, workplace English course, focused primarily on improving the students' oral production and aural comprehension. Participants completed pre-test and post-test standardized workplace ESL listening and reading assessments before and after the 10-week course. Participants also completed a minimum of three learning journal entries per week to document what they found easy and difficult in English, as well as what they learned, how they used what they learned, and what they would like to learn. The learning journals were coded qualitatively for themes in participants' perceptions of their own learning. Gain scores for the standardized listening and reading assessments did not show significant improvement after 30 hours of instruction. However, the participants' perceptions of the impact of the workplace program, as reflected in their learning journals, indicate noticeable benefits for the development of their receptive and productive skills. The authors conclude that workplace ESL programs should collect qualitative self-report data from students, such as the learning journals, in addition to standardized assessment data to evaluate the impact of this type of program with these adult learners.



**Elder, C. (2005).** Evaluating the effectiveness of heritage language education: What role for testing? *The International Journal of Bilingual Education and Bilingualism*, 8(2&3), 196–212.

*Keywords:* Chinese; Arabic; Vietnamese; Australia; elementary and secondary school; heritage language; bilingual; testing; test use

The Australian state of Victoria provides funding to bilingual primary and secondary education programs with a stipulation to evaluate the effectiveness of the programs during the first three years of funding using pre-post testing and qualitative data. Based on Elder’s experience as an external evaluator of five bilingual programs (Chinese-English, Arabic-English, Vietnamese-English), the paper illuminates roles and issues of language testing in evaluating the effectiveness of heritage language (HL) bilingual programs. Some of the challenges Elder’s evaluation team faced were program stakeholders’ varying purposes and concerns, testing fairness, lack of available tests for HL learners, test quality checks, rater bias, timing of test administration, program-external factors, and a transient student population. These challenges affected test choice, test administration, scoring, and interpretation of test results in each evaluation. In order to effectively use tests as an evaluative tool for HL education, Elder provides six suggestions: (a) systematically document HL learners’ sociolinguistic background; (b) generate a test bank of “context-sensitive exemplar assessment tasks” (p. 208); (c) develop “validated standardized tests” (p. 209) for various HLs and for learners at differing proficiency levels; (d) gather HL learners’ longitudinal language development data to illuminate patterns and stages of language achievement; (e) set publicly-agreed test standards and principles to ensure that the test reflects educational context and sociolinguistic context surrounding HL programs; and (f) guide policy-driven evaluation of bilingual and HL programs by theory.

**Elder, C. (2009).** Reconciling accountability and development needs in heritage language education: A communication challenge for the evaluation consultant. *Language Teaching Research*, 13(1), 15–33. <https://doi.org/10.1177/1362168808095521>

*Keywords:* Australia; bilingual education; K-12; accountability; developmental evaluation; external; heritage language; Mandarin; Vietnamese; Arabic

Elder reflects on three evaluations of heritage language programs in government schools in Australia. She describes the circumstances of each evaluation, the challenges faced by the team members and their varying degrees of success. In lessons learned, she emphasizes the need to negotiate and clarify the following issues before the evaluation begins: resources and funds; the purpose, scope and audience of the evaluation; the roles of the evaluator and evaluands; and what will constitute evidence. She also stresses that evaluators need to be flexible and responsive to feedback from participants and stakeholders. Finally, Elder argues that while there is often tension between the accountability and developmental functions of evaluations, the two do not

have to be mutually exclusive. Thus, external evaluators who have been hired primarily for accountability purposes can conduct evaluations that also contribute effectively to internal program development. Building productive relationships and maintaining open and effective communication with participants and stakeholders are two key strategies for simultaneously addressing both external accountability and internal development needs.

**Elley, W. B. (1989).** Tailoring the evaluation to fit the context. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 270–285). Cambridge University Press.

*Keywords: pragmatic; method comparison; Comparative Analysis*

Elley offers pragmatic suggestions at various decision points in planning and implementing evaluation. The planning stage involves identifying evaluator, purpose, intended outcomes, design, sample size, sampling, and instrumentation. In order to tailor implementation, Elley recommends: (a) organize a committee to discuss the plans at each stage and ensure objectivity in data collection and analysis; (b) tailor time and effort according to the importance of information to be gained; (c) determine aims through instructional materials and lesson plans in lieu of clearly defined aims; (d) conduct pre- and post-test comparison or survey large representative samples as a baseline before the program is introduced, if there are no comparison groups; (e) consider the homogeneity of the populations and samples involved; (f) survey the assigned school to look for schools with similar student composition, or survey for potential counterparts at the next highest grade level for later comparison; (g) determine the weight of skills tested, source of test materials, question types, length of the test, and forms of the test; (f) pilot and improve the test items. The timing of administration and the clarity of test specifications (procedure) should be considered, as well as cautious marking of the test after administration. Monitoring of experimental and control groups is also necessary to assure implementation. When analyzing results, it is also important to consider loss of cases, equating of groups, ‘ceiling effects,’ differences between sub-groups or classes, and the behavior of extreme groups.

**Enkin, E., & Correa, M. (2018).** Evaluating learner and teacher perceptions of program outcomes in the foreign language major. *Electronic Journal of Foreign Language Teaching*, 15(1), 66–80. <https://e-flt.nus.edu.sg/wp-content/uploads/2020/09/enkin.pdf>

*Keywords: teacher perception, learner perception, foreign language outcomes, mixed methods, United States*

Based on a mixed-methods approach, the study investigates foreign language major seniors’ and instructors’ perceptions of teaching and learning foreign languages at a university in the United

States. The study is guided by three major topics: the alignment between learners' and instructors' perception of the program, the assessment of learners' needs, and teachers' understanding of those needs. 33 participants were sent an online survey of both quantitative and qualitative questions over four semesters. The participants included 24 students and nine instructors from three modern language departments: Spanish, French, and German.

Four conspicuous themes emerged from an analysis of the open-ended responses to the survey: (a) learners' ability to communicate, (b) learners' association with other cultures, (c) learners' expectations and needs of their programs, and (d) learners' preparedness for their futures. In general, the student's perceptions of the foreign language programs were in alignment with those of the teachers, and teachers were well aware of students' needs. Based on the results, the authors make several suggestions: more speaking opportunities for learners to improve their communication skills, more phonetics and translation courses, and more chances for learners to experience authentic tasks. The results would be informative for the language department's stakeholders who aim to improve their curricula in a way that can increase students' chances of future success.

**Ennemoser, M., Kuhl, J., & Pepouna, S. (2013).** Evaluation of a dialogic reading program to improve language proficiency in children with a migrant background. *Zeitschrift Für Pädagogische Psychologie*, 27(4), 229–239.

*Keywords: reading instruction; German as a second language instruction; migrants; preschool children; language proficiency*

In the present study researchers evaluated the validity of the findings that dialogic reading is an effective strategy to enhance young children's language proficiency (Mol, Bus, de Jong & Smeets, 2008) for preschool children with a migrant background when dialogic reading was conducted in a small group setting in German preparatory language courses. 45 preschool children with language deficits who had been assigned to a preparatory language course were included in the study. Based on a matching procedure, half of the children were assigned to the experimental group and subsequently received a dialogic reading intervention as a substitute for their conventional language lessons. The other half remained in their regular language course. Results suggest that dialogic reading, i.e. the consequent application of facilitative interaction techniques, is effective for second language learners in a German preschool setting. During the intervention, the experimental group displayed significantly larger increases in a standardized language test than the control group participating in a regular language course for children with a migrant background.

**Eskey, D. E., Lacy, R., & Kraft, C. A. (1991).** A novel approach to ESL program evaluation. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 36–53). NAFSA.

*Keywords: US; university; ESL; reliability; face validity; academic success*

The chapter reports on an evaluation at the American Language Institute (ALI), an ESL program at the University of Southern California (USC) that prepares students for academic English with a heavy emphasis on content. The students were mostly matriculated students who were placed into the ALI through a five-part program-specific examination. The program's advanced courses were tied to the needs of academic units (schools and departments at the institution). Therefore, both the ALI and related academic units were involved in the evaluation. The authors evaluated the effectiveness of the ALI by looking at the academic success of students who were released (exited) from the ALI, compared to other populations at the university. Further, the release criteria (based on writing skills) were validated by analyzing the relationship between ALI writing scores (on a nine-point scale) and GPA-based success rates. The results indicated that the ALI-released students were capable of academic success, while those who failed in the ALI but still enrolled in USC mostly dropped out (only 5 out of 55 successfully completed the program), supporting the validity of the ALI release criteria. This chapter demonstrates how the decision point of releasing students from an ESL program into mainstream university courses is an important criterion that calls for validation, and one that is apparently closely related to students' success in the university.

**Fall, T., Adair-Hauck, G., Glisan, E. (2007).** Assessing students' oral proficiency: A case for online testing. *Foreign Language Annals*, 40(3), 377–406.  
<https://doi.org/10.1111/j.1944-9720.2007.tb02865.x>

*Keywords: Oral proficiency assessment; online; ACTFL; French; German; Japanese; Spanish; large-scale; K-12*

Fall, Adair-Hauck, and Glisan report on their longitudinal project of developing, implementing, and validating a K-12 online district-wide oral student proficiency assessment called Pittsburgh Public Schools Oral Ratings Assessment for Language Students (PPS ORALS) (Pittsburgh, Pennsylvania). The project was funded by the U.S. Department of Education, and the goal was to create accessible, feasible, and easy online testing aligned with the ACTFL Oral Proficiency Guidelines. Teachers across the district and language consultants collaborated to create tasks and assessment rubrics for the PPS ORALS, a process which empowered and equipped teachers with a greater understanding of proficiency-based instruction. Teachers' involvement and rater training had a positive washback effect on the curriculum and classroom practices. In the Appendix, the authors include detailed examples of speaking tasks, a rubric, and a can-do check-list for different proficiency levels.

**Fenton-Smith, B., & Torpey, M. J. (2013).** Orienting EFL teachers: Principles arising from an evaluation of an induction program in a Japanese university. *Language Teaching Research*, 17(2), 228–250. <https://doi.org/10.1177/1362168813475946>

*Keywords: higher education; second language teachers; English as a second language instruction; Japan; teacher education*

Fenton-Smith and Torpey present the results of a program evaluation of a two-week induction for 22 new English teachers at a private foreign languages university in Japan. The views of a range of stakeholders were obtained (beginning teachers, experienced teachers and management), as were perspectives at different points in time (before and after induction, one semester later, one or more years later). The evaluation resulted in a clear picture of the strengths and weaknesses of the orientation program, which in turn led to the implementation of a range of measures to improve current practice. The findings also gave rise to the proposal of a framework outlining the major areas that all EFL orientations need to consider.

**Fox, R. P. (1991).** Evaluating the ESL program director. In M. C. Pennington (Ed.), *Building better English language programs* (pp. 228–240). NAFSA.

*Keywords: ESL; program administration; directors*

Fox describes previous literature on the evaluation of ESL program directors as “stress[ing] the concept of accountability through performance evaluation, professional development and personal growth, and reward for outstanding performance” (p. 235). He then questions how to conduct and who will conduct the evaluation of an ESL program director. Fox suggests forming a committee (e.g., the immediate dean, vice-provost, faculty, staff, sponsors, and students) and setting performance criteria. He cites 12 types of performance criteria (problem analysis, judgment, organizational ability, decisiveness, leadership, sensitivity, range of interests, personal motivation, educational values, stress, oral communication skills, and written communication skills) as a basis for developing the evaluation instrument. After responses are collected, analyzed, and reported, “one of the most important results of the evaluation is the development of an improvement plan by the ESL program director, which should form the basis of interim informal evaluations” (p. 238). Thus, evaluation of the director contributes additionally to further development of the program.

**Gattullo, F. (2000).** Formative assessment in primary (elementary) ELT classes: An Italian case study. *Language Testing*, 17(2), 278–288. <https://doi.org/10.1177/026553220001700210>

*Keywords: Italy; elementary; EFL; formative; implementation; teachers; interview; assessment*

This work-in-progress case study of an Italian elementary school (3rd and 4th grade) takes a discourse analytic perspective on formative classroom assessment practices of teachers. Within 150 hours of classroom interaction data, assessment events were identified, transcribed, and coded into nine teacher feedback categories: questioning/eliciting, correcting, judging, rewarding, observing process, examining product, clarification request, task criteria, and meta-cognitive questioning (in rank order). Gattullo emphasizes the value of the meta-cognitive questioning (lowest rank in data) which will make students better articulate their understanding and thus second, a context inventory should be conducted to make decisions on what issues to prioritize and how to carry out the program evaluation (covering factors such as availability of comparison groups, reliable/valid language measures, evaluation expertise, and instructional materials and resources; background of students and staff; student selection process, size, intensity, perspectives and purpose of the program; timing of evaluation; and the social and political climate of the program). Third, developing a preliminary thematic framework will also clarify the conceptual framework of the program, what the salient issues are, and what aspect is going to be evaluated. Fourth, a data collection design and system has to be selected based on the questions that need to/can be answered; feasibility information from the context inventory also may limit the methodology (a useful decision making chart for data collection design/system is provided as an example). Fifth, collect data based on the clear purpose of the design; data collection may be eclectic since new questions and issues can emerge and the purpose of the evaluation can evolve with the program. Sixth, ideally, analyze data with a “multiple analysis strategy [which] can strengthen the evaluation by avoiding the possibility of bias associated with any particular technique” (p. 36). Finally, formulate and tailor the evaluation report for a particular audience.

**Gieve, S., & Cunico, S. (2012).** Language and content in the modern foreign languages degree: A students’ perspective. *Language Learning Journal*, 40(3), 273–291.  
<https://doi.org/10.1080/09571736.2011.639459>

*Keywords: nUnited Kingdom; higher education; language culture relationship; curriculum planning; content area instruction; qualitative analysis*

Gieve and Cunico report on a small-scale qualitative study of students’ experience of their Modern Foreign Languages (MFL) degrees with particular regard to the relationship between language and content learning. It is framed by the identification in the recent Worton Report on MFL studies in UK higher education and elsewhere of a dualism between language and culture in MFL degrees, which is reflected in the structure of the curriculum, its delivery and staffing. While this study reports the views of only a small number of students in one university, it adds to our understanding of how students’ expectations on entering the degree programme (mainly being that they will improve their linguistic competence) are linked to their experience of the degree and their evaluation of what they have gained from it. Their weak appreciation of

connections between language form, language use and moments of culture (whether textual, cognitive or in the form of cultural practices) and intercultural communication, and of other beneficial outcomes of a modern languages degree, appears to be associated with a curriculum that does not promote an integration of language and content. To this extent the dualism identified in the literature appears to be associated with a negative aspect of these students' experience. Some suggestions are made for what could be done to manage the separation between language and cultural content.

**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(3), 505–540. <https://doi.org/10.1111/j.1944-9720.2009.01034.x>

*Keywords: university; accreditation; program development; self-efficacy; quantitative; questionnaire; program theory; foreign language; process; case study*

This article describes a university foreign language program evaluation focusing on student self-efficacy. The evaluation was part of a larger, summative evaluation being carried out for accreditation purposes. It was developed by a team of second language faculty to evaluate one of the department's core competency statements, "Students of Arabic, Chinese, French, German, Italian, Japanese, Portuguese, Russian, and Spanish will demonstrate confidence in using the second language in their language classrooms, and their future expectancies of their ability to use the second language in real-life contexts" (p. 506). The team developed a Likert-style questionnaire for students based on literature on self-efficacy and faculty input regarding outcomes and expectations. The questionnaire was administered to fourth semester students in all languages. Quantitative analysis was performed to determine the extent of students' perceived self-efficacy for various outcomes, and the team developed specific suggestions for improvement based on the results. Gorsuch notes that the process of developing the evaluation was as beneficial as the findings because it required faculty members from different language divisions to examine their assumptions about what students should be able to do and articulate a shared program theory. Furthermore, the evaluation successfully served two purposes: accountability and program development.

**Gottlieb, M., & Nguyen, D. (2007).** *Assessment and accountability in language education programs: A guide for administrators and teachers*. Caslon Publishing.

*Keywords: accountability; assessment; portfolio; English; bilingual education; K-12*

To begin, Gottlieb and Nguyen review national and local perspectives on accountability and assessment surrounding English language learner education (I.e., dual-language,

transitional-bilingual, and ESL programs). The authors propose an assessment framework called Balanced Assessment and Accountability System, Inclusive and Comprehensive (BASIC), a model implemented at the Schaumburg, Illinois, School District 54. Their model is an assessment and accountability framework that links state, district, program, and classroom-level assessments with curriculum and instruction. In planning for assessment, the model emphasizes the consideration of the effect of internal and external contextual factors on program design and assessment practices. Such factors include learning goals, benchmarks and standards, characteristics of the program constituents, and program mission and vision. Once contextual factors are identified and learning goals are listed, the next step is to match the purposes and types of assessment tools to each goal. Gottlieb and Nguyen suggest the use of student portfolios, called “pivotal portfolios,” which involve systematic collection of different types of student learning and achievement data based on agreed upon common assessment tools. The book showcases examples of the uses of the “pivotal portfolio” for classroom and program decision-making in dual language programs and transitional bilingual programs. The authors point out that “pivotal portfolio” data can be used not only for student assessment but also as a response to authentic (I.e., locally-relevant) accountability pressures and to improve instruction and student learning. The book includes worksheets and checklists to help educators design a contextualized assessment framework.

**Grim, F. (2010).** Giving authentic opportunities to second language learners: A look at a French service-learning project. *Foreign Language Annals*, 43, 605–623.  
<https://doi.org/10.1111/j.1944-9720.2010.01104.x>

*Keywords: community service learning; French; FL; college; standards; motivation; qualitative; journal; questionnaire*

The authors described a service-learning project for 25 L2 French university students who worked in local elementary schools to expose children to French as a foreign language. The project was based on the Alliance for Service-Learning in Education Reform’s standards for service learning, with the goals of exposing children to the French language and cultures, impacting the community through service, and building motivation to continue learning (and possibly teach) French among the university student participants, through applying their knowledge of French to the service learning experience. The project aimed to describe the impact of service-learning on the university students’ language learning motivation, their professional goals, and their roles in the community, through analysis of the students’ lesson plans and weekly journals. The evaluation itself was not linked to the Alliance for Service-Learning standards, but was based primarily on students’ comments in their journals and final questionnaires addressing the impacts that the project had on their levels of motivation and professional goals. Recommendations for conducting service-learning projects are included.



**Griva, E., & Sivropoulou, R. (2009).** Implementation and evaluation of an early foreign language learning project in kindergarten. *Early Childhood Education Journal*, 37(1), 79–87. <https://doi.org/10.1007/s10643-009-0314-3jj>

*Keywords: Greece; EFL; kindergarten; children; observation; tests; interview; field note*

Griva and Sivropoulou report on an evaluation study of a pilot, game-based, English as a foreign language intervention program implemented in two Greek state kindergarten classrooms. The evaluation focused on (a) the development of kindergarteners' English oral skills, (b) implementation of the intervention, and (c) teacher perceptions and experiences of the intervention. The authors administered pre- and post-tests (word production, event identification, word completion), took observation field notes, and conducted interviews with English language teachers and kindergarten teachers. Overall, the findings indicated the intervention's positive impact on student learning as well as teachers' and students' positive attitudes towards and experiences with the intervention.

**Grosse, C. U. (2004).** The competitive advantage of foreign languages and cultural knowledge. *Modern Language Journal*, 88(3), 351–373. <http://www.jstor.org/stable/3588783>

*Keywords: US; MBA; foreign language; culture; alumni survey*

Grosse describes how graduates from the Thunderbird business school, which requires a minimum of four semesters of foreign language study, view the advantage of their knowledge in foreign language and culture in the international business community. An online web survey was distributed through email to 2500 alumni who graduated between 1970 and 2002, and 581 responded. Over 80% of the respondents said that foreign language skills and cultural knowledge gave a competitive advantage, suggesting the value of foreign language and cultural knowledge in business, despite the mismatch between the languages studied at Thunderbird and the languages they needed in their workplaces. The article addresses the importance of foreign language in an MBA program in general. It also shows how graduates of a program can prove to be an important source of expertise for illuminating how learning in the program can be applied in the real world.

**Gruba, P., Cárdenas-Claros, M. S., Suvorov, R., & Rick, K. X. (2016d).** *Blended language program evaluation*. Springer. <https://doi.org/10.1057/9781137514370>

*Keywords: blended language program, program evaluation, argument-based framework*

The book walks readers, both practitioners and researchers, through theories and practices of evaluating blended language programs using the argument-based framework proposed by

Chappelle (2014). The book is organized around the four stages of the argument-based framework: argument planning, data collection, argument presentation, and argument appraisal.

Chapter 1 reviews three concepts, blended learning, evaluation of technology integrated learning, and language program evaluation as they are associated with blended language program evaluation. Based on this review, Chapter 2 introduces two approaches, developmental evaluations (Patton, 2011) and action research (Bax, 2011) that can be useful in evaluating blended language learning. Additionally, following Gruba and Hinkelman (2012), the authors discuss blended learning from the three levels: micro- (i.e., classroom), meso- (i.e., departmental), and macro- (i.e., organizational), with regards to four considerations: purpose, appropriateness, multimodality, and sustainability.

The first two phases of the framework, argument planning and gathering evidence, are discussed in chapters 3-5. In particular, the chapters discuss data collection methods including surveys, interviews, and observations in consideration of several ethical issues. In doing so, the authors emphasized that all the phases of data collection can be implemented at the micro- (Chapter 3), meso- (Chapter 4), and macro-levels (Chapter 5).

Chapters 6-9 present the third step of the framework (argument presentation) by demonstrating how a blended language program is evaluated in practice. Each chapter presents a case study conducted in Chile, Hawaii, Vietnam, and Australia, respectively, across three levels of structure (micro-, meso-, and macro-).

The last chapter asks what is learned from the evaluation and how to move forward. To answer these questions, the chapter addresses both the strengths and weaknesses of the argument-based approach when evaluating blended language programs.

**Gruba, P., & Chau Nguyen, N. B. (2019).** Evaluating technology integration in a Vietnamese university language program. *Computer Assisted Language Learning*, 32(5-6), 619–637. <https://doi.org/10.1080/09588221.2018.1527365>

*Keywords: Argument-based approach, technology, ethnography, document analysis, program evaluation, Vietnam*

The study adopts an argument-based approach and evaluates technology integration within an English language department in a Vietnamese university.

The authors collected data using multiple qualitative methods, including ethnography, document analysis, questionnaires, and semi-structured interviews. As an insider in the English department, Nguyen led the ethnographic research with the aim to provide more authentic insights of the department. For document analysis, 110 subject guides in the English department were analyzed

for references to technology use. 10 of the 49 lecturers in the department responded to the open-ended questionnaire, and four of the six academic administrators were interviewed.

The study presents the results of the data analysis to assess whether the evidence was sufficient to meet the assumptions of each inference. In terms of Evaluation and Explanation, the study found a lack of guidelines and policies in the macro and micro level through document analysis and insights from the questionnaire. The authors attributed such absence to inadequate professional development, low student digital literacy, and an inefficient responsibility distribution among administrators.

**Hajjaj, A., & Al-Najjar, B. (1989).** ESL program evaluation: Realities and perspectives. In J. E. Alatis (Ed.), *Georgetown University round table on languages and linguistics, 1989* (pp. 133–141). Georgetown University Press.

*Keywords: Kuwait; EFL; university; questionnaire; framework*

Hajjaj and Al-Najjar describe issues that arose in ESL program evaluation at Kuwait University (KU). They characterize ESL program evaluation in the 1980s by introducing the emerging notions of process and product evaluation, program-fair evaluation, and shared or negotiated evaluation. Further, they note that considerable attention was beginning to be paid to evaluating affective and cognitive aspects of learning, as well as inclusion of evaluation as an integral part of curriculum development. The authors present the results of survey research on the realities of ESL evaluation in Arabian universities, and they set out a framework for future ESL program evaluation there, including: (a) an overall comprehensive evaluation plan (not only testing) should be developed and communicated among stakeholders; (b) by monitoring progress throughout the program, evaluation will be an integral part of the learning process; (c) an evaluation should gather appropriate and relevant data; (d) evaluation should be action-oriented, feeding back into program development; and (e) the validity of evaluation study should be clarified.

**Hampel, R., & de los Arcos, I. A. (2013).** Interacting at a distance: A critical review of the role of ICT in developing the learner-context interface in a university language programme. *Innovation in Language Learning and Teaching*, 7(2), 158–178.  
<https://doi.org/10.1080/17501229.2013.776051>

*Keywords: language learning, language teaching, computer-assisted language learning, online learning, distance learning, learner–context interface*

Hampel and de los Arcos examine the introduction of new online technologies to support distance language learning in a higher education institution in the UK, charting the development

from using telephone conferencing in the 1990s to the implementation of Moodle and videoconferencing more recently. They use the sociocultural concept of the learner-context interface to emphasize the centrality of both learners and context in the design and delivery of technology-supported language courses rather than making the development of computer-mediated learning opportunities the main focus. Building on research and evaluation work carried out over more than 15 years, Hampel and de los Arcos analyze the issues that have arisen and that have affected change regarding technology and pedagogy. Central areas of investigation in terms of the learners were found to be interaction, learning communities, metacognition, literacy, affect and learner support; in terms of context they include task design, teacher roles and teacher skills. In the conclusion, limitations of the research and new developments are outlined.

**Hargreaves, P. (1989).** DES-IMPL-EVALU-IGN: An evaluator's checklist. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 35–47). Cambridge University Press.

*Keywords: checklist; method; model; purpose; agent; curriculum development; theoretical*

Hargreaves argues that approaches to curriculum planning have often treated “design, implementation, and evaluation” as a linear process, positing evaluation as a post-hoc matter. He proposes a cyclical integrated view and illustrates a checklist of twelve mutually dependent factors that can be utilized when planning an evaluation: target audience (non-specialist / specialist), purpose (formative / summative), focus (direct / indirect), criteria (global / relative), method (a priori / empirical), means and instrument (a priori / empirical), agents (internal / external), resources (staffing / funding), time factors (timing and time scales), findings (nature and status of findings), presentation of results (formats), and follow-up (action). Integration of evaluation to curriculum is essential for all stages of curriculum development.

**Harklau, L., Norwood, R. (2005).** Negotiating researcher roles in ethnographic program evaluation: A postmodern lens. *Anthropology and Education Quarterly*, 36(3), 278–288. <https://www.jstor.org/stable/3651396>

*Keywords: ethnography; evaluator role; postmodernism; stakeholder; secondary education; minority learners*

Harklau and Norwood discuss how evaluators' roles and reflexivity are shaped by and relative to institutional and societal discourses. In the official role of external evaluators, Harklau and Norwood conducted an ethnographic evaluation of a month-long summer college readiness program (including ESL and math among other subjects) for underrepresented middle school students. The researchers reflect on the fluid multiple roles they were positioned in by the

stakeholders of the program: as insiders, as outsiders, as colleagues, as teaching staff, as program benefactors, and even as ornamental researchers. Throughout the study, Harklau and Norwood negotiated their evaluator roles and power with the stakeholders of the program and occasionally resisted their positionality. Taking a postmodern perspective, they suggest that: (a) evaluation is a performative act, in particular recognizing that “science is a representation, not a transparent reality” (p. 285); and (b) policy makers should acknowledge multiple ways of knowing, including the value of ethnography as an evaluation method.

**Harris, J. (1990).** The second language programme-evaluation literature: Accommodating experimental and multifaceted approaches. *Language, Culture, and Curriculum*, 3(1), 83–92.

*Keywords: Ireland; Irish; bilingual education; elementary school; experimental; multi-faceted; observation; questionnaire*

Harris discusses how the purpose of evaluation differs between two distinct evaluative approaches, experimental and multi-faceted. The experimental approach is often used for theory-oriented evaluation, which generates theory, tests hypotheses, and seeks causal relationships, rather than making practical decisions about an individual program. Many bilingual education studies carried out in Canada are theoretically-oriented evaluations that respond to policy shaping purposes. The multi-faceted approach focuses on short-term, applied, practical, decision making, though it may eventually shape policy as well. Harris presents an example of a nation-wide evaluation of Irish-language programs in Irish primary schools to illustrate how a long-term study can take both approaches, depending on the focus. The evaluation began as a test-development project, with an “output-oriented quasi-experimental approach” (p.87) for decision-making and policy orientation. However, a thorough investigation of expectations of schools and a greater focus on process issues (classroom observation and student, teacher, and parent questionnaire) emerged, reflecting the pedagogic concerns of participants. In a follow-up evaluation, a hypothesis testing study was carried out to verify “the relationship between general ability, amount of naturalistic use of Irish in school and achievement in different aspects of spoken Irish” (pp. 90-91). The accommodation of both approaches can bring greater depth and generalizability to program evaluation.

**Harris, J. (2009).** Late-stage refocusing of Irish-language programme evaluation: Maximizing the potential for productive debate and remediation. *Language Teaching Research*, 13(1), 55–76. <https://doi.org/10.1177/1362168808095523>

*Keywords: Ireland; Irish language; national language; language policy; context; political; data analysis; case study; criterion-referenced; assessment; K-12*

Harris discusses a series of Irish language program evaluations in Ireland and their role in the public debate on language policy and program development. Focusing on two studies in particular he describes how adjustments were made during the evaluation process to achieve a more comprehensive understanding of the findings and prevent their misuse or misinterpretation. For example, when faced with potentially problematic findings, the evaluation team performed a more thorough analysis of the data and looked at contextual factors that helped explain the findings. By clarifying and contextualizing the data, the evaluators were able to produce reports that contributed constructively to the public debate on Irish language education policy. Harris emphasizes the political nature of evaluation and calls on evaluators to consider the political implications of their work and take responsibility for clarifying findings so that they are not misinterpreted.

**Hedge, T. (1998).** Managing developmental evaluation activities in teacher education: Empowering teachers in a new mode of learning. In P. Rea-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 132–158). Longman.

*Keywords: university; teacher training; self-evaluation; developmental evaluation*

Hedge highlights the outcomes of developmental evaluation activities (also provided as Appendices) in a 50-hour core teacher education course in a postgraduate programme. Two case studies show how the use of self-evaluation activities assists teachers in developing awareness about managing group work through collaborative tasks. “The reflective investigation and developmental evaluation activity [through experiential learning] can build strong motivation among participants on teacher education courses” (p. 150).

**Heining-Boynton, A. L. (1990).** The development and testing of the FLES program evaluation inventory. *Modern Language Journal*, 74(4), 432–439. <https://doi.org/10.2307/328517>

*Keywords: US; elementary school; FLES; rating scales; testing*

Heining-Boynton reports on the development and testing of a multifaceted FLES program evaluation instrument, a survey distributed to FLES teachers, principals, administrators, students, elementary classroom teachers, and parents. The instrument covered issues that the FLES program faced historically, as well as concerns elementary foreign language programs had at the time: teacher qualifications, goals and objectives, pedagogy, articulation, homework and grades, parent support, FLES teacher acceptance by colleagues, workload, at-risk students, and student satisfaction.

**Herzog, M. (2003).** Impact of proficiency scale and the oral proficiency interview on the foreign language program at the Defense Language Institute Foreign Language Center. *Foreign Language Annals*, 36(4), 566–571. <https://doi.org/10.1111/j.1944-9720.2003.tb02146.x>

*Keywords: Interagency Language Roundtable scale; history; Defense Language Institute Foreign Language Center*

In this article, Herzog describes (a) the impetus behind and history of the Interagency Language Roundtable (ILR) scale; (b) the way the ILR scale is used as a guiding framework for curriculum, instruction, and assessment at the Defense Language Institute Foreign Language Center; and (c) the tasks, content domains, and linguistic accuracy illustrated in the ILR scale. The analysis of the proficiency level descriptors is particularly useful for those who are interested in proficiency scale development or understanding the ACTFL Proficiency Guidelines and the ILR scale.

**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(5), 37–54.

*Keywords: US; university; Mandarin; utilization focused; web-based; hybrid; improvement; formative; questionnaire; interview; participatory; case study*

annotation: This study reports on the evaluation of a web-based, beginning-level conversational Mandarin course at the University of Hawaii at Manoa in Fall 2005. The evaluation use of this study was to improve the courses and generate a deeper understanding of how the program worked. Users of the evaluation included one tutor and one instructor in the program, as well as several university language departments and a language resource center (the NFLRC). The evaluation questions were, “How satisfied are students with each part of the course,” “What are classroom- and individual-related motivation factors,” and “What are the students’ needs?” The evaluation consisted of several overall steps, including identifying program constituents and analyzing their relationship with the program, examining evaluation documents, implementing improvements, asking students to re-evaluate the course, and planning future evaluations and implementation stages. The authors report on two evaluation stages. In the first, evaluation documents are examined and actions to improve the course are taken. In the second, students re-evaluate the course and provide input on the improvement actions from the first stage. Data-collection methods in the first stage included questionnaires and structured interview data. Second-stage data-collection methods included survey questionnaires, instructor observations, and communication with students. From the first stage, the authors found an overall positive attitude towards instructors, the tutor, and the course; that students were happy with the textbook and CD as well as the design and content of the online course components; negative responses to the difficulty of the listening workbook, lack of oral interaction opportunities, and technical

difficulties; and need for longer speaking tasks and opportunities to review and repeat material. Second-stage findings were that instructors and tutors were highly rated; there was a desire for more practice with natural conversation; there were issues with the difficulty level and topics of material; there was poor adaptation to online learning; there was a positive attitude towards improvement actions; some textbook topics were seen as useful but not others; and students had trouble with some online tasks. The authors conclude by providing lessons learned and attribute the possibility of the evaluation to the local context focus and use of existing sources, a strong evaluator, and the use-focused nature of the U-FE approach adopted.

**Horwitz, E. K. (1985).** Formative evaluation of an experimental foreign-language class.  
*Canadian Modern Language Review*, 42(1), 83–90.

*Keywords: US; university; French; classroom observation; interview; curriculum; formative; innovation*

Horwitz illustrates the formative evaluation of an experimental French as a foreign language class with a communicative focus at the University of Illinois Urbana-Champaign. The evaluation was conducted by an outside evaluator using structured interviews and systematic classroom observation. An observational coding system captured how many turns each student took (response, question, or comment), whether the utterance was spontaneous or not, whether the utterance was in English or the target language, and whether the utterance was long or short. By comparing the characteristics of students' communicative behavior in one activity with another, the teacher can modify the new activity to effectively engage students. Besides quantitative aspect of utterances, the coding system can be modified to examine the quality of the utterance, such as to include types of feedback and error modification. The results of Horwitz's use of this observational coding system were triangulated with student and teacher interviews. As a result of the formative evaluation, the instructor incorporated student initiated topics, adapted grammar lessons (due to students' concerns about the disadvantage of a communicative approach for the second semester), and confronted student's habitual off-topic comments. Horwitz argues that, when a teacher is trying to implement a new approach, formative evaluation can provide useful and timely feedback (if the observation system is not too time consuming or labor intensive). In particular, it can help to monitor both teacher and students' classroom behavior, inform instructional modifications, adjust pedagogical tasks, and last but not least, raise unanticipated issues in time to try to solve them.



**Houston, T. (2005).** Outcomes assessment for beginning and intermediate Spanish: One program's process and results. *Foreign Language Annals*, 38(3), 366–376.  
<https://doi.org/10.1111/j.1944-9720.2005.tb02223.x>

*Keywords: US; university; Spanish; foreign language; outcomes assessment; ACTFL; standards; oral proficiency; survey; placement exam; portfolio; case study*

Houston describes a university Spanish language program's process for articulating and assessing learning outcomes. Using the ACTFL Guidelines and the Standards for Foreign Language Learning in the 21st Century as rough guides, the program developed both proficiency goals and general program goals. To assess outcomes, the program looked at student gains on the placement exam, student satisfaction surveys, and oral proficiency interviews and tasks. The proficiency assessments indicated that the program was generally successful, although there was some conflicting data on students' grammar skills. However, the student satisfaction surveys showed that students felt some general program goals were not met. The program used this feedback to make improvements.

**Hudson, T. D. (1989).** Mastery decisions in program evaluation. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 259–269). Cambridge University Press.

*Keywords: criterion-referenced; testing; mastery decision*

Hudson addresses the use of criterion-referenced measurement (CRM) to assess student performance (mastery or non-mastery) in relation to program goals for program evaluation purposes. Mastery testing establishes the absolute standing of students' performance against the objectives of instruction. Hudson discusses issues of reliability/dependability of CRM, in which consistency of decisions can be resolved through various statistical approaches. He also suggests that content-based or Contrasting Groups methods may inform the validity of decision standards in CRM. The process of developing CRM through input from instructors, material writers, and administrators can strengthen the curriculum by calling for a rationalization of instructional goals and methods. Further, by analyzing the results of CRM, "the evaluator can determine the extent to which the program is i) producing the desired results and ii) realistic in its goals" (p. 265). A major difference between Bachman's (1989) and Hudson's approach to criterion-reference testing is whether testing involves reference to other programs or to the program itself. The choice of approach may depend on whether there is a strong need for generalization in evaluation outcomes (e.g., for accountability purposes).

**Imani, S. I. (2013).** Evaluation of modular EFL educational program (Audio-visual materials translation & translation of deeds & documents). *English Language Teaching*, 6(4), 8–17. <http://dx.doi.org/10.5539/elt.v6n4p8>

*Keywords: EFL educational program, program evaluation, translation courses*

Imani evaluated the Modular EFL Educational Program from five fundamental criteria including: Admission Requirements, Program Content, Program Resources, Program Instruction/Evaluation Methods, and Graduation/Employment Requirements. Methodologically, the study is based on the requirements of both qualitative and quantitative research paradigms. A sample of teachers enjoying at least five years of offering both courses took a 22-item Likert-scaled questionnaire accommodating subcategories of the five macro criteria followed by open-ended written protocol commenting spaces for qualitative data. The findings revealed controversies over the all the macro-criteria and compatibility of the program with these well-established standards; suggesting exercise of comprehensive revisits and modifications in all aspects of the program as a whole.

**Issa, B. I., Koronkiewicz, B., & Faretta-Stutenberg, M. (2022).** Second-language writing in university-level basic language programs: A survey of student and instructor beliefs. *Foreign Language Annals*, 55(2), 383–407. <https://doi.org/https://doi.org/10.1111/flan.12609>

*Keywords: second language writing, L2 writing assessment, students' perceptions, university-level, United States*

The study investigates the (mis)alignment between instructors' and students' perceptions of L2 writing assessments. Data were collected from 183 students and 28 instructors in the Spanish language programs across three different public universities: The University of Tennessee, Knoxville, The University of Alabama, and Northern Illinois University. The participants were emailed an online survey with questions that were categorized into four parts: who decides the format of assignments, why the assignments should be part of the curriculum, what assignments should be included, and how the assignments should be distributed.

The results showed that there was a consensus between the two groups regarding all four components of the survey. In particular, both believed that focusing on grammar and vocabulary knowledge was essential in writing assignments. However, there seemed to be a misalignment between instructors and students in the pedagogical purpose of writing assignments. Based on these findings, the study suggests that the instructors focus on grammar and vocabulary knowledge when assessing writing assignments for greater uniformity in the process of grading. Additionally, reflecting students' high needs for creative writing, the study suggests developing

assignments that are different in format from traditional essay writing, such as brochures, posters, or social media posts.

**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)* [PDF document]. University of Hawai‘i, Second Language Teaching & Curriculum Center.

*Keywords: USA; Japanese; heritage language; university; needs analysis; survey*

The authors report on the findings of a learner needs analysis conducted for the lower-level two year Japanese as a foreign language (JFL) program at the University of Hawaii at Manoa. A Performance-Based Testing committee, which consisted of nine full-time lower-division Japanese instructors, led the needs analysis project. The committee was tasked to (a) identify language learning domains for the lower-level JFL courses, (b) create a test item bank aligned with the language learning domains, and (c) establish grading criteria and scoring procedures for the performance-based test. The needs analysis was conducted to identify the language learning domains (i.e., area, theme, and task), and to investigate if any differential views existed between teachers and students as well as within sub-groups of teachers (i.e., experienced versus inexperienced, native versus non-native). In order to capture opinions from a large sample, a survey was conducted. A student survey was administered to students enrolled in 100- and 200-level courses, and all lower-level course instructors were invited to respond to the teacher survey. The article ends with a list of content and task domains, derived from the needs analysis, that require attention and discussion for the department to proceed with curricular reform. The student and teacher questionnaires are appended.

**Jacobson, P. L. H. (1982).** Using evaluation to improve foreign language education. *Modern Language Journal*, 66, 284–291.

*Keywords: improvement; accountability; triangulation; personal factor; political; utilization*

Jacobson describes constraints on the availability of valid evaluative information, shows how to aid foreign language education through program evaluation, and provides suggestions for improving the utilization of evaluation information. Jacobson refers to summative evaluation as “the most authoritative and defensible information,” while ongoing formative evaluation is “an integral part of a foreign language program is a sine qua non for providing valid data to decision makers” (p. 288-289). She advocates the use of evaluability/evaluative assessment (to determine the likelihood of an evaluation success), needs assessment (to determine the gap between the desired status and the current status), and implemental evaluation (to determine the reality of the implementation process). When it comes to utilization of evaluation, personal factors

(responsibility, leadership, enthusiasm, determination, etc.), political climate, and format of the report all interact to determine whether the evaluative outcomes will be utilized by the stakeholders of the program.

**Janssen, G., & Restrepo, S. (2019).** A utilization-focused program evaluation of an ERPP tutoring service at one Colombian university. In J. N. Corcoran, K. Englander, & L. Muresan (Eds.), *Pedagogies and policies for publishing research in English* (pp. 19–35). Routledge. <https://doi.org/10.4324/9781315151229>

*Keywords: utilization-focused program evaluation, English for research publication purposes, tutoring program, Colombia*

The study evaluates an English for research publication purposes (ERPP) tutoring service at a private university in Colombia. The university is devoted to improving its ranking, resulting in high investments to increase academic publications in top-tier journals. With this goal in mind, the university developed the ERPP tutoring service, a one-on-one tutoring program between an expert and a novice writer.

The chapter addresses the findings from the program evaluation of the tutoring service using the utilization-focused evaluation framework across three iterations (2015-2, 2016-2, and 2017-1 semesters). The specific foci of the evaluations were (a) the degree to which the service is conducted as planned and (b) issues that arose during the implementation of the service that needed to be addressed.

The data was collected from nine PhD tutees using a research archive for each writer and their information. The archive was supplemented with video-recorded, semi-structured interviews at the end of the program. The researchers then compared this data with the writing service's stipulated preconditions to determine the appropriateness of the preconditions and the usefulness of the service, as well as any structural issues of the program. The findings revealed that the initial structure of the program was not flexible enough. The second iteration of the study supported the finding that more flexibility was needed to satisfy the users of the tutoring service. During this iteration, not just Ph.D. students but M.A. students as well as professors were also supported by the service in the process of developing academic papers. Based on these findings, the chapter suggests developing a formalized structure that can support tutors' and program coordinators' positions and reduce their workload.

**Jenks, F. L. (1991).** Designing and assessing the efficacy of ESL promotional materials. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 172–188). NAFSA.

*Keywords: ESL; promotional materials; administration*

Jenks discusses the purposes and the effectiveness of ESL program promotional materials, such as videos, brochures, program advertisements, posters, newspapers and newsletters. English language programs that target international clients need to seek ways to improve their promotional strategy/material through constant formative assessment that seeks to balance cost and effectiveness. The effectiveness of each type of promotional material can be evaluated by utilizing the following strategy: (a) place a code in the application page of the brochure to track distribution; (b) chart the number and country source of the returned preprinted forms; (c) attach a tear-off pad or postal card to the poster to later tally the response; (d) tally the number of inquiries through newspaper and newsletters (author notes that the effectiveness of video is difficult to assess). Not covered here are online advertisements and web pages, which were not prominent at the time of the publication.

**Johnson, R. K. (Ed.) (1989).** *The second language curriculum*. Cambridge University Press.

*Keywords: curriculum planning; ends/means specification; program implementation; classroom implementation; faculty development*

This book offers a collection of papers arguing for a cohesive curriculum and emphasizing the interdependence of various stages (curriculum planning, specification of ends and means, program implementation, and classroom implementation) throughout the development and evaluation process. Evaluation is understood as “necessary and integral part of each and all of the stages” (p. xiii). The chapters in the book cover all aspects of curriculum, and some focus on evaluation. Chapters include: “A decision-making framework for the coherent language curriculum” (Chapter 1: Johnson); “Syllabus design, curriculum development and polity determination” (Chapter 2: Rodgers), “DES-IMPL-EVALU-IGN: an evaluator’s checklist” (Chapter 3: Hargreaves); “Needs Assessment in Language Programming: From Theory to Practice” (Chapter 4: Berwick); “The Role of Needs Analysis in Adult ESL Programme Design” (Chapter 5: Brindley); “Service English Programme Design and Opportunity Cost” (Chapter 6: Swales); “Faculty Development for Language Programs” (Chapter 7: Pennington); “The Evolution of a Teacher Training Programme” (Chapter 8: Breen & Candlin, Dam, & Gabrielsen); “Appropriate Design: The Internal Organisation of Course Units” (Chapter 9: Low); “Beyond Language Learning: Perspectives on Materials Design” (Chapter 10: Littlejohn & Windeatt); “Hidden Agendas: The Role of the Learner in Programme Implementation” (Chapter 11: Nunan); “The Evaluation Cycle for Language Learning Tasks” (Chapter 12: Breen); “Seeing the Wood AND the Trees: Some Thoughts on Language Teaching Analysis” (Chapter 13: Stern);

“Language Program Evaluation: A Synthesis of Existing Possibilities” (Chapter 14: Brown); “The Development and Use of Criterion-Referenced Tests of Language Ability in Language Program Evaluation” (Chapter 15: Bachman); “Mastery Decisions in Program Evaluation” (Chapter 16: Hudson); and “Tailoring the Evaluation to Fit the Context” (Chapter 17: Elley). This collection, and its many examples, will be of key interest to anyone who is concerned with developing the various components of language programs. Only chapters that cover the issues of language program evaluation are annotated here.

**Karava-Doukas, K. (1998).** Evaluating the implementation of educational innovations: Lessons from the past. In P. Rea-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 25–50). Longman.

*Keywords: Greece; secondary; innovation; overview; trends; methods; participatory evaluation; communicative language teaching; EFL*

Karava-Doukas examines the implementation issues associated with language program innovation, and she provides an example implementation study of an English teaching curriculum innovation in secondary schools in Greece. Some of the factors that influence successful educational innovations include: (a) teacher’s attitudes and beliefs towards education (attitude clarification and refinement); (b) the clear articulation of an innovation proposal (specified goals and means in non-technical terms); (c) teacher training (systematic, ongoing, and long-term training which clarifies teacher beliefs, makes teachers innovators, and accommodates teachers’ existing knowledge); (d) communications and support (administration and peer support); and (e) the compatibility of the innovation with contingencies and constraints of the classroom and wider educational contexts (time, resources, organizational constraints, teachers’ perception of needs, and teaching style). The Greek EFL case study used classroom observation, questionnaires, interviews, and reports of classroom practice to reveal the implementation of innovation. Key findings included the gap between an intended innovative curriculum and actual classroom practice, as well as the disjuncture between a communicative approach and teachers’ beliefs about and understanding of the approach.

**Kennedy, C. (1988).** Evaluation of the management of change in ELT projects. *Applied Linguistics*, 9, 329–342.

*Keywords: innovation; management; stakeholders; theoretical; process*

Kennedy addresses the notion of innovation theory in program evaluation, considering a program as a systematic organization where various factors interact. He suggests “[In] evaluating any project we should be concerned not only to evaluate the outcome of the project but the process of

innovation itself, the stages it passes through, from the identification of a problem to the selection of the innovation and its final incorporation, acceptance, and diffusion” (p. 329). To create an innovative change in the curriculum, the program manager/developer has to: (a) understand the underlying attitudes and beliefs of the program stakeholders; (b) monitor and adjust the process of change; (c) investigate the relationship of the process to the outcomes; (d) incorporate any information found that can make the change; and (e) “return to projects some time & to see whether the change has been incorporated to the system” (p. 330). Kennedy also emphasizes that the innovations need to be contextually adaptable to local conditions, and that all participants be involved/consulted so that they see what benefits would be gained by the innovation. In classroom innovation, cultural, political, administrative, educational, and institutional factors interrelate and cannot be ignored. A good match in terms of feasibility, acceptability, and relevance between innovation and the existing program can generate acceptance from the stakeholders. Kennedy concludes with questions that can be asked throughout the innovation process for evaluating change management.

**Kennedy, C., & Miceli, T. (2013).** In Piazza online: Exploring the use of wikis with beginner foreign language learners. *Computer Assisted Language Learning*, 26(5), 389–411.  
<https://doi.org/10.1080/09588221.2013.770035>

*Keywords: computer assisted language learning; Italian as a second language, student attitudes; cooperative learning, wikis*

In this paper, Kennedy and Miceli report on the evaluation of the integration of wikis into their first-year Italian course with the aim of encouraging out-of-class practice and fostering students’ sense of class community, right from the start of their learning. The evaluation showed that, although the students created several attractive and interesting pages, they did not appreciate the wikis as much as we had hoped: there were technical hitches, many found collaboration dynamics challenging, and most developed little interest in participating in a cross-campus online group. The evaluation found no relationship between the students’ perceptions of the wiki work and their gender, initial confidence or frequency of use of computer-mediated communication (CMC) tools. However, those who, on entering the course, placed greater importance on interaction with other students, and a sense of community in class, showed greater appreciation of the wiki experience. From these findings, the researchers draw implications for improving their approach to integrating wiki work into their program.

**Kiely, R. (1998).** Programme evaluation by teachers: Issues of policy and practice. In P. Rea-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 78–104). Longman.

*Keywords: UK; EAP; Europe; university; overview; methods; participatory evaluation; ethnography*

Kieley bridges ethnography with program evaluation. He presents a case study evaluation of a 12-week EAP program in a British university, for both improvement and accountability purposes. The evaluation is based on ethnographic methodologies (interviews, classroom observation, field notes, questionnaires, structured discussions, and program documents). Kieley also seeks to inform the relationship between evaluation and pedagogy through the use of evaluation dialogues between the teacher and the students in a classroom. This approach transforms both students and teachers into active participants in a democratic classroom evaluation process.

**Kiely, R. (2006).** Evaluation, innovation, and ownership in language programs. *Modern Language Journal*, 90(4), 597–601. [https://doi.org/10.1111/j.1540-4781.2006.00466\\_8.x](https://doi.org/10.1111/j.1540-4781.2006.00466_8.x)

*Keywords: assessment; ownership; marketization; EFL; university; UK*

Kiely stresses that engagement in meaningful in-depth evaluation depends on educators' perceptions of their departmental culture, disciplinary orientations, and their own professional roles and expertise as teachers and academics. As an illustrative example, he describes an internally driven program evaluation and innovation studies of English as a Foreign Language and English for Academic Purposes programs at Thames Valley University in the UK. Three major motivations for conducting the evaluation studies were faculty members': (a) awareness of the need to respond to the changing market situations in education (e.g., enrollment, needs); (b) understanding of assessment as a central activity to inform curriculum and internal and external stakeholders about the program; (c) buy-in for conducting research on student learning. Kiely concludes that an educator's sense of ownership of the program is the key factor in generating cycles of program evaluation, development, and innovation.

**Kiely, R. (2009).** Small answers to the big question: Learning from language programme evaluation. *Language Teaching Research*, 13(1), 99–116. <https://doi.org/10.1177/1362168808095525>

*Keywords: UK; university; EAP; ESL; learning; program development; context; innovation; ethnography; case study*



Kiely's article explores the concept of learning from program evaluation. He discusses various issues that can hinder learning, including multiple and sometimes conflicting purposes for evaluation, lack of interaction and collaboration between stakeholders, and lack of clarity about learning from versus learning through evaluation. He also traces the historical trends of program evaluation and their limitations with regard to their contributions to learning. Kiely advocates for evaluations that focus more on understanding the contextual features of programs. Three contextual features that he views as particularly important are innovation, teachers at work, and the quality of the student learning experience. To demonstrate the impacts of the various issues and contextual features on evaluation, he analyzes an evaluation of learning materials being used in an EAP program at a British university. His analysis includes an ethnographic study of the evaluation which reveals not just what was learned from the evaluation, but also learning opportunities that were missed. Finally, Kiely concludes that to maximize learning, program evaluation needs to become "a socially-situated cycle of enquiry, dialogue, and action" (p. 99).

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

*Keywords: history; case study; method; framework; teacher-led; management-led; impact analysis; standard; development; management; SLA; research; ESL; EAP; immersion; Africa; Asia; US; Europe; Canada; Australia*

This book introduces principles and procedures that may be adapted to a variety of language program evaluation contexts. It consists of four sections: background (Part 1), case studies (Part 2), framework (Part 3), and resources (Part 4). In Part 1, the nature of program evaluation (chapter 1), its history and developments in design, methodology (chapter 2), context, and use (chapter 3), and theory development in language learning (chapter 4) are discussed. Kieley and Rea-Dickins argue that Evaluation is about the relationships between different program components, the procedures and epistemologies developed by the people involved in programs, and the processes and outcomes which are used to show the value of a program – accountability – and enhance this value – development (p. 5). Program evaluation is also complex, and the authors address five challenges: (a) the clarification of the purpose of evaluation, (b) the articulation and understanding of stakeholders' values and factors that mediate them, (c) the identification of evaluation criteria, (d) collection of valid data, and (e) assurance of evaluation use for program development, management, and research advancement. In Part 2, seven case studies illustrate program/project context, aim, and scope, as well as evaluation design, procedures, sample instruments, and implementation. These cases enable readers to understand the relationship between specific programs and evaluation practices, and they provide useful bases for relating evaluations to other language program contexts. The case studies vary widely, including: (a) a nation-wide study of teachers' English language skills in an EAL/ESL context (chapter 5); (b) a multinational evaluation of the language component of the Science Across

Europe project in secondary schools (chapter 6); (c) a large-scale evaluation of the contribution of native speaker teachers in secondary schools in Hong Kong (chapter 7); (d) a multi-site evaluation of foreign language teaching pilot programs in primary education in Ireland (chapter 8); (e) a quality management evaluation (chapter 9) and an evaluation of students' experiences (chapter 10) in an EAP program at a British university; (f) a document evaluation of national and state assessment standards for EAP learners in different contexts (Canada, USA, and Australia) (chapter 11); and (g) an impact evaluation by external evaluators of the Centre for Canadian Language Benchmarks (chapter 11). These case descriptions enable understanding of why each evaluation study was conducted (evaluation for accountability, learning, sense-making, curriculum development) and how analytic frameworks were adopted and applied (ethnographic, survey, classroom observation, document evaluation). In addition to the case studies, chapter 12 provides a comprehensive discussion of how stakeholder participation can be sought in evaluation. Part 3 examines three different types of impetuses for evaluation: large-scale evaluation (chapter 13), teacher-led evaluation (chapter 14), and management-led evaluation (chapter 15). When conducting a large-scale evaluation, understanding the construct, design (validity, questions, procedures, analysis), and implementation (capacity, constraints, and ethical issues) are especially important. Teacher-led evaluation is effective when: (a) it is linked with pedagogic concerns; (b) teachers perceive a need for change and/or perceive evaluation as opportunity for improvement; and (c) there is sufficient time and teachers are involved in quality management. Sample teacher-led evaluation projects are described, focusing on a course textbook, a curriculum innovation, and a teacher education research methods course. Particularly relevant for language programs in the US, who face an evaluation mandate from external accreditation bodies, is the discussion of management-led evaluation. The involvement of management may develop and facilitate the use of links between program evaluation and management processes of performance assessment and professional development (p. 255). Sample evaluations of an EAP program in South Africa (a ten-stage procedure), and of management and use of resource centers in Eastern and Central Europe by the British Council, are used to illustrate example frameworks here. The last section (Part 4) provides a list of resources that may be useful for extending knowledge about evaluation in general, including books, journals, professional associations, research ethics, electronic mailing lists, and internet resources.

**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). Mouton de Gruyter.

*Keywords: applied linguistics; accountability; materials evaluation; methods evaluation; teacher evaluation; stakeholder participation; contextualization; evaluation practice applied linguistics; accountability; materials evaluation; methods evaluation; teacher evaluation; stakeholder participation; contextualization; evaluation practice applied linguistics;*

*accountability; materials evaluation; methods evaluation; teacher evaluation; stakeholder participation; contextualization; evaluation practice*

The authors describe trends of language program evaluation practices, discuss issues related to program evaluation in language education and research, and outline ways research and theory developments in applied linguistics can contribute to the praxis of program evaluation.

Over the past years, purposes of evaluation in language education expanded from theory-building, monitoring, and compliance to program development and improvement. Programs are now seen as a dynamic interplay among various stakeholders (e.g., learners, teachers, institution) and program elements. As the dynamism of programs is unpacked, the nature of the evaluation work should be multifaceted. To meet various evaluation demands in a complex program context, Kiely and Rea-Dickins suggest that evaluators should (a) involve diverse stakeholders to ensure ownership of evaluation; (b) “view effectiveness in diverse ways” (p. 679); and (c) utilize diverse and efficient ways to provide a comprehensive account of the program under study.

The authors draw attention to two issues in language program evaluation. One is the preponderance of evaluation studies that are embedded in decontextualized, short-term second language acquisition research. Another concern is the nature of evaluation practices within the scope of accreditation and accountability mandates, which often narrowly prescribe evaluation design, do not allow creativity and innovation, and overemphasize learner satisfaction.

In the final section, Kiely and Rea-Dickins discuss how development of language learning theories, conceptual analysis, research methods, and tools in applied linguistics can be instrumental for program evaluation practices.

**Kim, Y. (2012).** Implementing ability grouping in EFL contexts: Perceptions of teachers and students. *Language Teaching Research*, 16, 289–315.  
<https://doi.org/10.1177/1362168812436894>

*Keywords: Korea; FL; English; middle school; ability grouping; quantitative; qualitative; questionnaire; teacher; student; perceptions*

This study looks at the implementation of ability grouping in English as a foreign language classes in Korean middle schools, as well as teacher and student perceptions of the benefits and problems associated with it. Separate questionnaires were distributed to teachers (n=55) and students (n=754) in 19 schools, and suggestions for improving ability groupings were elicited from both participant groups. Participants’ descriptions of implementation, perceptions, and needed improvements were analyzed qualitatively for common themes that emerged from the data. Student responses to Likert-scale questions about perceptions were analyzed quantitatively

to identify patterns in perceptions among students placed in different ability groups. Results showed that students who were placed in higher ability groups tended to have slightly more positive perceptions of ability grouping and that teachers were more motivated to work with higher ability than lower ability groups. There was wide variability in how ability grouping was implemented and perceived among the different schools. The author concludes that ability grouping seems to be associated with inequitable educational practices in Korea, and recommends that future studies collect data from other stakeholders and incorporate outcome measurements to provide evidence of ability grouping effectiveness. Instruments are included in the appendices.

**Klee, C., Melin, C., Soneson, D., & Tarone, E. (2014).** From frameworks to oversight: Components to improving foreign language program efficacy. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 131–153). Cengage Learning.

*Keywords: longitudinal program evaluation; student-learning outcomes; outcomes benchmarks; interpretive competence; interpersonal competence; presentational competence*

Carol Klee, Charlotte Melin, and Dan Soneson describe the development of proposed educational outcomes and recommendations from the Second Language Acquisition Working Group at the University of Minnesota. This committee of faculty experts developed institutional benchmarks in foreign languages that aimed to align intellectual content in general education with language development goals in FL study. The authors, including the director of the language center and directors of the Spanish and German studies programs, suggest that longitudinal program evaluation and oversight with input from various stakeholders and experts in the field of second language acquisition is essential for meaningful change to occur. They conclude that reorienting FL programs to encompass larger intellectual goals may reemphasize the relevance of FL learning in 21st-century higher education.

**Kondo-Brown, K., Davis, J. McE., & Watanabe, W. (2014).** Evaluation capacity building in college language programs: Developing and sustaining a student exit survey project. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 15–40). Cengage Learning.

*Keywords: postsecondary program evaluation, student's exit survey initiative, assessment capacity, CLLL departments*

Kimi Kondo-Brown, John McE. Davis, and Yukiko Watanabe describe a graduating student's exit survey initiative in the College of Language, Linguistics, and Literature at the University of Hawai'i at Manoa. This survey, designed to elicit feedback on educational experiences and outcomes from all graduating students across some 30 programs, provided a means for jump-starting evaluation practices in response to accreditation pressures, but with a primary goal of providing useful feedback to faculty and programs. The authors (an associate dean and, at that time, two PhD students specializing in program evaluation) highlight key stages in the survey development and implementation-critically including extensive consultation with program directors, chairs, and other insiders-along with issues, challenges, and lessons learned as well as meaningful and productive outcomes for faculty and staff.

**Koyalán, A. (2009).** The evaluation of a self-access centre: A useful addition to class-based teaching? *System*, 37, 731–740. <https://doi.org/10.1016/j.system.2009.09.016>

*Keywords: Turkey; EFL; university; self-access center; autonomy; questionnaire; observation*

The self-access center (SAC) at the Izmir University of Economics in Turkey provides students with text-based and audio-visual English resources (e.g., books, magazines, newspapers, DVDs), self-learning English activities (e.g., exam papers, practice sheets, exercises), and space for self-study (audio-video lab and study space). An evaluation was conducted to showcase the degree of success and effectiveness of the facility. More specifically, the evaluation focused on (a) learning practices at SAC, (b) perceived value of SAC, (c) SAC's facilitative role in classroom learning, (d) SAC's resource and activity needs, and (e) SAC's impact on student learning, students' approach to learning, and student autonomy. With these evaluation foci in mind, data collection involved student and staff questionnaires and observations of the center. Based on the findings, the article provides some suggestions for SAC's improvement.

**Lai, C., Zhao, Y., & Wong, J. (2011).** Task-based language teaching in online ab initio foreign language classrooms. *Modern Language Journal*, 95(s), 81–103. <https://doi.org/10.2307/41413421>

*Keywords: TBLT, task-based language teaching, Chinese, online, CFL, implementation, innovation, synchronous, asynchronous, background survey questionnaires, reflection blogs, classroom observations, recordings, evaluations, oral performance, task cycle*

This study investigates task-based language teaching (TBLT) in a beginning-level, online Chinese foreign language course for high-school students. Lai, Zhao, and Wang note that evaluations of how TBLT is implemented in online classrooms are relatively scarce, particularly online classrooms with beginning-level language learners. The purpose of this study is twofold:

to examine the reactions to TBLT of online ab initio (complete beginner) teachers and students, and to study the issues that emerge from task-based implementation in this beginning-level, online context. Participants consisted of four teachers, student enrollees in an online TBLT course, and student enrollees in a non-TBLT course. The TBLT online course consisted of synchronous and asynchronous components, the former consisting of 12, one-hour, task-based sessions with a teacher, and the latter relying heavily on a non-task-based e-textbook. Background surveys, weekly reflection blogs, classroom observations and recordings of synchronous sessions, course-final evaluations, recordings of production on an oral performance task, and weekly debriefings and an end-of-program interview with teachers were the data-collection tools. Results for the first research question showed an overall satisfaction with the TBLT online course, that some students in the TBLT course became less reliant on their teacher and more autonomous in their own learning, that the TBLT course had differential effects on students' oral production, that students lacked the skills and strategies for dealing with TBLT, and that the features of the online platform of an online course were important. For the second research question, the authors found designing a TBLT syllabus and implementing a task cycle, managing collaborative tasks, the delay in sound transmission of the online platform, and exclusive target-language use to be challenges to implementing TBLT in an online setting. Towards the end of the study, the authors outline some of the advantages to working with TBLT in an online context, noting similar and different issues between TBLT online and face-to-face in the classroom, and making a number of suggestions potentially applicable to both contexts. They note that the online learning environment, however, has great potential for task-based implementation by enabling individualized instruction, reducing cognitive load for from-scratch beginners, and promoting student engagement.

**Lee, J. (2009).** ESL student teachers' perceptions of a short-term overseas immersion programme. *Teaching and Teacher Education*, 25, 1095–1104.  
<https://doi.org/10.1016/j.tate.2009.03.004>

*Keywords: English; FL; professional development; student teachers; study abroad; quantitative; qualitative; questionnaire; observation; journal; essay*

This article describes the program design, implementation, and evaluation of a six-week study abroad immersion program in New Zealand for 15 pre-service English FL student teachers from Hong Kong. The study examined participants' perceptions of the impact the program had on their English proficiency, cultural awareness, teaching skills, and personal development, by collecting data regarding the program's academic studies, field experiences, homestay experience, and community/cultural activities. Data collection included personal background questionnaire and essays, mid-program evaluation, field observations, participant journals, final reflection papers, and two post-program questionnaires (Likert-scale and open-ended). Findings

present benefits, challenges, and recommendations for implementing an immersion program for student teachers.

**Lett, J. A. (2005).** Foreign language needs assessment in the US military. In M. H. Long (Ed.), *Second language needs analysis* (pp. 105–124). Cambridge University Press.

*Keywords: US; Defense Language Institute; foreign language; needs analysis; proficiency; subject matter experts; reliability; validity*

This chapter reports on how the Defense Language Institute Foreign Language Center (DLIFLC), in the US, has been conducting a systematic foreign language needs assessment for setting foreign language proficiency requirements of different career fields. The purpose is to: (a) assure that funds are well spent for educating students to an appropriate level; (b) manage military linguists (identify how they are deployed, assign appropriate tasks, set criteria for keeping the job, etc.); and (c) identify what proficiency level is required to accomplish distinct military jobs. Lett describes analysis procedures in detail and discusses issues of reliability and validity. In particular, he reports on how career group subject matter experts (SMEs) discussed job tasks, conditions, and standards. He also proposes two different methods for understanding reliability of judgments about proficiency and task requirements (see below). In addressing validity, he suggests (a) assuring that the task list is reflecting the SMEs' understanding of their career fields, and (b) comparing the task requirements (high and low proficiency tasks) with supervisors' perceptions of task performance by high and low proficiency individuals. The reliability and validity resolutions for task and language specification can be utilized for any needs analysis study, in addition to verifying interpretations through multiple sources and methods. Modified split-half procedure: Two groups of SMEs each discuss the tasks and tasks' language requirements. Later they compare the results and reach consensus. Surrogate or partial test-retest design: Show a video-taped discussion to another group of participants and compare the proficiency judgment between groups.

**Li, B., & Tin, T. B. (2013).** Exploring the expectations and perceptions of non-native English speaking students in masters level TESOL programs. *New Zealand Studies in Applied Linguistics*, 19(2), 21–35.

*Keywords: second language teachers; student attitudes; nonnative speakers; TESOL; teacher education*

Second language teacher education has been regarded as central to ensuring the quality of the English learning experience of many students around the world. In recent years, an increasing number of non-native English speaking teacher trainees have gone to English speaking countries

to attend post-graduate level teacher education programs. One central consideration is to what extent these non-native English speaking teacher trainees receive adequate preparation from these programs. Li and Tin investigate this pivotal concern by evaluating one Masters TESOL program in an English speaking country in the Pacific Region. They employed qualitative evaluation, using in-depth interviews with a group of non-native English speaking participating students from Asian backgrounds, focusing on their expectations and perceptions of the program. Findings demonstrate that the program has many strengths, such as improving English reading and writing proficiency, cultivating subject knowledge related to applied linguistics and promoting research engagement. However, there are also reported weaknesses with the most recurrent one reported being lack of practice teaching. The findings indicate a need for language teacher education programs which both incorporate exploration of students' expectations and establish built-in procedures for student evaluation of the entire program rather than just of separate courses.

**Li, L., & Walsh, S. (2011).** Technology uptake in Chinese EFL classes. *Language Teaching Research*, 15, 99–125. <https://doi.org/10.1177/1362168810383347>

*Keywords: China; English; FL; CALL; teachers; quantitative; qualitative; survey questionnaire*

This large-scale study focused on the existing use and impact of information and communications technology (ICT) in Chinese secondary schools, particularly related to how Chinese EFL teachers are trained to implement ICT in EFL instruction. Data collection involved a survey questionnaire that was administered using a probability multi-staged sampling strategy of 450 teachers (89% response rate) and follow-up focus group interviews of 33 teachers. The teacher questionnaire focused on four main areas: Personal background and teaching experience, ICT environment in the instructional setting, ICT skills and training, and current ICT use (opinions, attitudes, expectations). Based on their responses to the questionnaire, teachers were classified as either ICT or non-ICT. Findings showed that although most ICT and non-ICT teachers regularly accessed computers at home and in the workplace with a high degree of competence, the implementation of ICT in the EFL setting is relatively low for both. Results point to both positive and negative factors impacting ICT integration, with the implication that more training be provided to EFL teachers to integrate CALL in EFL instruction. The survey questionnaire and the focus group questions are included in the article appendices.

**Liskin-Gasparro, J. E. (1995).** Practical approaches to outcomes assessment: The undergraduate major in foreign languages and literatures. *ADFL Bulletin*, 26(2), 21–27.

*Keywords: US; foreign language; outcomes assessment; portfolio; oral proficiency test; interview; university*



Liskin-Gasparro emphasizes using outcomes assessment to incorporate a reflective component into the program and to give directions for change and improvement. She describes the development and use of instruments for assessing students' content knowledge, attitudes about the program, postgraduate activities, and linguistic knowledge, skills, and performance. She also notes how the intention to measure both the growth/process of student learning and specific skills has introduced the use of portfolios, which can be aligned with departmental objectives. Two case studies of the development of foreign language outcomes assessment are presented. The University of Iowa (department of Spanish and Portuguese), utilized a variation of the simulated oral performance assessment (later replaced by a Spanish speaking test), a writing assessment (later replaced by a portfolio), an exit interview, and a questionnaire to enrolled students and alumni to reveal learners' needs. The second case study at Bates College (Spanish and French sections of the department of classical and Romance languages and literatures) was an internally-motivated assessment plan using portfolios. The appendices include an exit interview protocol from the University of Iowa and an example portfolio program from Bates College.

**Liskin-Gasparro, J., & Vasseur, R. (2014).** Designing an embedded outcomes assessment for Spanish majors: Literary interpretation and analysis. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 83–110). Cengage Learning.

*Keywords: utilization-focused assessment; Spanish; student learning outcomes; undergraduate majors, rubric development*

Judith Liskin-Gasparro and Raychel Vasseur present a case study that describes the department's design and implementation of a rubric to assess the content knowledge and literary interpretation skills in writing of graduating Spanish majors at the University of Iowa, following a university mandate that all academic departments develop student learning outcomes for their undergraduate majors. Liskin-Casparro, a faculty member and program director, and Vasseur, a PhD student, describe the benefits and challenges of utilization-focused assessment as well as the process of engaging departmental faculty in its development and implementation.

**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(1), 35–54. <https://doi.org/10.1177/1362168808095522>

*Keywords: US; K-12; reading; ESL; outcomes; quasi-experimental; multiple methods; context; data collection; data analysis; NCLB; political*

Llosa and Slayton describe their evaluation of the Waterford Early Reading Program, a reading intervention program for at-risk kindergarten and first grade students in an urban school district in California. The evaluation sought to (a) determine the effectiveness of the program on reading ability, (b) understand how the program was being implemented, and (c) look specifically at its effectiveness and implementation for English language learners. Llosa and Slayton provide a detailed description of the context, the quasi-experimental design, and the findings. They then discuss conditions and strategies that allowed them to successfully carry out the evaluation and provide results that were useful to various stakeholders and decision makers. They stress that in a difficult political environment, evaluators can make their studies more useful by expanding the focus beyond outcomes measurement. Instead, by utilizing multiple methods of data collection, incorporating qualitative data, and thoroughly investigating the context as well as the outcomes, evaluators can better understand the reasons for certain outcomes and frame findings and recommendations in a way that they can be appropriately acted upon.

**Lo, Y. (2010).** Implementing reflective portfolios for promoting autonomous learning among EFL college students in Taiwan. *Language Teaching Research*, 14, 77–95.  
<https://doi.org/10.1177/1362168809346509>

*Keywords: Taiwan; English; FL; college; portfolio; quantitative; qualitative; autonomous learning; questionnaire; project evaluation*

This study examines the use of learner-centered portfolios in an EFL university course titled “Journalistic English” in Taiwan, in order to identify the challenges that students and teachers face in implementing portfolios for language learning purposes. Data collection consisted of pre-course and post-course self-evaluation questionnaires (Likert-scale and open-ended questions) regarding their perceptions and experiences with the portfolio project, and the researcher’s weekly field notes regarding students’ feedback, challenges, and personal reflections. Prior to the course, none of the 101 participating students had completed a portfolio or learner-centered project before. The article explains the process of designing and implementing the portfolio project and related assessment criteria, and then analyzes the students’ reactions to their experience, pointing to the challenge of implementing critical thinking skills and student-centered learning in EFL portfolio projects. The pre- and post-course self-evaluation questionnaires are included in the article appendices.

**Long, M. H. (1984).** Process and product in ESL program evaluation. *TESOL Quarterly*, 18, 409–425. <https://doi.org/10.2307/3586712>

*Keywords: product; process; summative; formative; classroom observation; second language acquisition*

Long, a second language acquisition researcher, draws upon a classroom research perspective to inform language program evaluation. He describes the necessity of looking at “what is actually going on in classrooms as opposed to what is thought to be going on” (p. 422) to supplement the evaluation of program/classroom products. Many of the comparative studies on teaching methodology in the 1970s and 1980s tended to focus on student outcomes alone, and to not distinguish between evaluation and assessment. However, “the product evaluations cannot distinguish among the many possible explanations for the results they obtain because they focus on the product of a program while ignoring the process by which that product came about” (p. 413). Long defines process evaluation as “the systematic observation of classroom behavior with reference to the theory of (second) language development which underlies the program being evaluated” (p. 415), and he distinguishes it from formative evaluation. Long’s process evaluation focuses the evaluation on illuminating real classroom behaviors in order to understand the program more comprehensively.

**Long, M. H. (2005a).** Methodological issues in learner needs analysis. In M. H. Long (Ed.), *Second language needs analysis* (pp. 19–76). Cambridge University Press.

*Keywords: needs analysis; sources; methods; triangulation; outsider; insider; expert; non-expert; validity; reliability; questionnaire; interview; journal log; language audit; ethnographic; observation; tests*

Long provides a comprehensive overview of methodological issues in needs analysis (NA), focusing on the sources of information, methodologies for eliciting information, and source × method combinations for improved interpretations. Some of the sources of information used in previous NA studies include learners, teachers, applied linguists, domain experts, (un)published literature, and other sources. Long also discusses advantages and disadvantages of various methodologies including expert and non-expert intuitions, interviews, questionnaire surveys, language audits, ethnographic methods, observations, journal logs, and tests. He then presents an example NA study on identifying tasks and language demands of airline flight attendants, revealing the importance of collecting data from various sources (e.g., flight attendants in the study), and not relying solely on experts’ (applied linguists in the study) intuitions. Each data collection method (written introspections, unstructured interviews, and surreptitious recordings of target discourse) showed different advantages (time, labor, specificity, and conciseness) in relation to different types of data (task types, lexis, and language use). Interactions of source and method were also found: written information best described the tasks and language; insiders compared with outsiders provided richer information on tasks and technical terms; written introspection was a more efficient way to collect information on tasks and language use than were unstructured interviews, but this applied for outsiders and not for insiders; and surreptitious recordings were better for baseline data on language use. More study on NA research itself in

different contexts is called for to understand the effective use of sources and methodologies, and to attend to reliability and validity of information/interpretations.

**Long, M. H. (Ed.) (2005b).** *Second language needs analysis*. Cambridge University Press.

*Keywords: needs analysis; validity; reliability; SLA; rationale; methodology; case studies*

This book provides useful insights and examples for those who intend to use needs analysis for research purposes and/or for curriculum development and evaluation. In his introduction, Long provides a rationale for conducting needs analysis to inform effective course design and to hold programs accountable. While needs analysis may vary by context, the book outlines a comprehensive research approach, principally through discussion of methodological considerations (especially chapter 1) that can be generalized to most contexts. Collected case studies (chapter 3-11) reflect needs analyses applied in a variety of societal, occupational, vocational, and academic domains, and they emphasize the link between setting, purpose, and methodology. For foreign language educators, chapter 3 (“Foreign language needs assessment in the US military” by Lett) and chapter 7 (“A task-based needs analysis of a tertiary Korean as a foreign language program” by Chaudron et al.) may be of key interest. Other cases cover societal-level language needs identification for policy shaping (chapter 2), English-language needs of hotel maids in Hawaii (chapter 4) and journalists in Spain (chapter 6), foreign-language (German) needs in business firms (chapter 5), Dutch-language needs of foreign professional footballers in the Netherlands (chapter 7), target task identification of naturalization interviews in the U.S. (chapter 9), dialogue analysis of coffee service encounters in Hawaii (chapter 10), and analysis of small-talk features in the workplace in New Zealand (chapter 11).

**Loughrin-Sacco, S. J., Matthews, S. A., Sweet, W. M., & Miner, J. A. (1990).** Reviving language skills: A description and evaluation of Michigan Tech’s summer intensive French course. *ADFL Bulletin*, 21(2), 34–40.

*Keywords: US; French; immersion; intensive course; false beginners; curriculum; questionnaire; placement; university*

The authors report on a curriculum and its evaluation, in the context of a summer intensive French program at Michigan Technological University. A previous ethnographic study had revealed that 56% of the students who had experience studying French for at least a year (false beginners) enrolled in elementary French courses to pull up their grade point average. In order to revive false beginners’ language skills and entice them to enroll in higher level classes, Michigan Technological University set up two-week intensive immersion courses in French, German, and Spanish. The authors first describe the French program in detail (schedule, student-to-teacher

ratio, content, course materials), report on the results of the students' course evaluations, and give suggestions on how the program can be applied in other contexts. For the evaluation, an ETS advanced-placement test was administered pre- and post-immersion to view the success of the course; an end-of-the-course questionnaire was also administered to the students to collect their impressions of the course and their motivation towards foreign language study. Based on the survey, the program changed the schedule for some of the activities that received low ratings. Most of the students from the French and Spanish intensive course continued their foreign language study at the intermediate or advanced level, indicating partial fulfillment of the initial purpose of the program.

**Lynch, B. K. (1990).** A context-adaptive model for program evaluation. *TESOL Quarterly*, 24(1), 23–42. <https://doi.org/10.2307/3586850>

*Keywords: context-adaptive model; framework; method; reporting; audience; purpose; university*

Lynch proposes a context-adaptive model for program evaluation following seven steps. First, the reasons for program evaluation will be different for each audience; thus, to ensure that the audience gets the most out of the evaluative outcomes and can make use of information later, the audience and their goals have to be identified. Second, a context inventory should be conducted to make decisions on what issues to prioritize and how to carry out the program evaluation (covering factors such as availability of comparison groups, reliable/valid language measures, evaluation expertise, and instructional materials and resources; background of students and staff; student selection process, size, intensity, perspectives and purpose of the program; timing of evaluation; and the social and political climate of the program). Third, developing a preliminary thematic framework will also clarify the conceptual framework of the program, what the salient issues are, and what aspect is going to be evaluated. Fourth, a data collection design and system has to be selected based on the questions that need to/can be answered; feasibility information from the context inventory also may limit the methodology (a useful decision making chart for data collection design/system is provided as an example). Fifth, collect data based on the clear purpose of the design; data collection may be eclectic since new questions and issues can emerge and the purpose of the evaluation can evolve with the program. Sixth, ideally, analyze data with a “multiple analysis strategy [which] can strengthen the evaluation by avoiding the possibility of bias associated with any particular technique” (p. 36). Finally, formulate and tailor the evaluation report for a particular audience.

**Lynch, B. K. (1992).** Evaluating a program inside and out. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 61–99). Cambridge University Press.

*Keywords: ESP; Mexico; university; quantitative; qualitative*

Lynch describes both quantitative and qualitative approaches that were utilized for formative and summative evaluation of the University of Guadalajara (UdeG)/University of California, Los Angeles (UCLA) Reading English for Science and Technology (REST) Project, which sought to improve reading skills for chemical engineering students.

Quantitative data consisted of a norm-referenced test (The English as a Second Language Placement Exam (ESLPE), fill-in-the-blank cloze) and criterion-referenced test (multiple-choice cloze). However, the ESLPE was the only pre- and post-test completed by both the treatment and control groups. As for the qualitative data, teacher/researcher journals, administrative logs, observation and program documents, interviews, questionnaires, and meeting notes were utilized. The data were coded and reduced into an Effects Matrix and Site Dynamics Matrix to “characterize the various outcomes and changes associated with the program” (p. 78) and to solicit and discuss the dilemmas and problems perceived. One problem the researcher faced was that the qualitative data collection was intensively done for the REST (experimental) program, but not for the control group, who did not receive any English instruction.

The researcher launched the project with evaluation in mind, which is not always the case in language curriculum development projects; his approach thus demonstrates how to incorporate evaluation from the outset. The REST project also makes obvious that collecting both quantitative and qualitative data enables a richer interpretation of any program. Lynch also calls attention to the need for linking intended analyses with the kinds of data collected, though the additional link between data analyses and their use for evaluative purposes is less clearly articulated.

**Lynch, B. K. (1996).** *Language program evaluation: Theory and practice*. Cambridge University Press.

*Keywords: EFL; university; method; measures; research design; positivistic; naturalistic; quantitative; qualitative; context adaptive model; REST project; Mexico*

Lynch provides an overview of the conflict between quantitative (positivistic) and qualitative (naturalistic) research paradigms in program evaluation, and he finds a middle ground by introducing the context-adaptive model (CAM) for evaluation, “a flexible, adaptable heuristic – a starting point for inquiry into language education programs that will constantly reshape and redefine itself, depending on the context of the program and the evaluation” (p. 3). Lynch begins by highlighting the potential contribution of program evaluation research to the field of second

language acquisition (SLA). Language program evaluation is “the systematic attempt to gather information in order to make judgments or decisions” (p. 2), for justifying and/or improving a program that targets the development and use of learners’ language abilities (chapter 1). Chapter 2 discusses the debate between positivistic and naturalistic paradigms, and describes the history of how the two resulting methodologies have been utilized in program evaluation. Lynch encourages careful articulation of what counts as evidence in the two different paradigms, and presents the associated issues of internal and external validity in chapter 3. In the following chapters, quantitative and qualitative research designs (chapter 4 and 5), data gathering, and analysis procedures (chapter 6 and 7) are presented through examples of how past program evaluations utilized those methods. Lynch argues that combining the two approaches to program evaluation research will allow an evaluator to acquire rich and thorough information about the program. In the case of Reading English for Science and Technology (REST) Project in Mexico, where a mixed-methods approach was used for evaluating an English for specific/academic purposes program, qualitative data helped to explain apparent contradictions in quantitative data (chapter 8). In chapter 9, Lynch provides a useful checklist of the seven steps that CAM follows: (a) determine the audience, goals, and purpose of evaluation; (b) gather information for a context inventory; (c) identify preliminary thematic framework (emergent themes and issues); (d) select appropriate data collection designs/methodology to answer the questions; (e) collect data following the design chosen; (f) analyze and interpret data (conduct multiple perspectives negotiation for increasing relevance and understanding); and (g) tailor the reports for intended audiences.

**Lynch, B. K. (2000).** Evaluating a project-oriented CALL innovation. *Computer Assisted Language Learning*, 13(4-5), 417–440.  
[https://doi.org/10.1076/0958-8221\(200012\)13:4-5;1-E;FT417](https://doi.org/10.1076/0958-8221(200012)13:4-5;1-E;FT417)

*Keywords: Australia; CALL; university; foreign language; observation; questionnaire; documents; logs; qualitative; validity; context-adaptive model*

Lynch describes an evaluation of the Project-Oriented Computer Assisted Language Learning (PrOCALL) innovation in foreign language courses (Chinese, French, German, Indonesian, Japanese) at the University of Melbourne, where he acted both as an evaluator and as an administrator. Following the context adaptive model, the goals, audiences, preliminary thematic framework, and data collection design/system were determined. However, the evaluation evolved with the project implementation. The thematic framework and research questions were revised based on changing perspectives of the participating teachers and the project director. The data were collected from class documents (emails, web pages, other texts), documentation of project (brochures, ethics and grant proposals), teacher and director’ s logs, teacher interviews, student focus group interviews, semi-structured classroom observations, quality of teaching surveys, and open-ended student feedback questionnaires. In addition to data analysis done by the researcher,

Lynch enhanced the validity of interpretations not only by triangulating data from multiple resources, but by consulting the project participants for interpretation and going back to the data for “counter examples and rival explanations” (p. 437). Some recommendations for teaching contexts similar to the ProCALL approach are offered.

**Lynch, B. K. (2003).** *Language assessment and programme evaluation*. Edinburgh University Press.

*Keywords: assessment; evaluation; paradigm; purposes; design; measures; analysis; interpretivist; qualitative; positivist; quantitative; validity; ethics*

Lynch highlights the relationships and differences between assessment and evaluation, and he presents “the range of paradigms (positivist vs. interpretivist), perspectives, designs, purposes, methods, analyses, and approaches to validity and ethics that currently define language assessment and programme evaluation” (p. vii). He defines evaluation as “the systematic inquiry into instructional sequences for the purpose of making decisions or providing opportunity for reflection and action” and assessment as “the range of procedures used to investigate aspects of individual language learning and ability, including the measurement of proficiency, diagnosis of needs, determination of achievement in relation to syllabus objectives and analysis of ability to perform specific tasks” (p. 1). He also posits evaluation as a super-ordinate category. Lynch claims that the purposes of assessment (assessment for decision making and/or learning) and evaluation (summative and formative purposes) interact with different methodological approaches. These he simplifies into two paradigm clusters, positivist (seeks to determine objective causal relationships) and interpretivist (seeks to understand socially constructed complex and fluid relationships). The chapters in the book flesh out how these paradigms influence assessment and evaluation designs (chapter 2), instruments (chapters 3 and 5), analyses (chapters 4 and 6), and validity and ethics (chapter 7).

**Ma, Q. (2008).** Empirical CALL evaluation: The relationship between learning process and learning outcome. *CALICO Journal*, 26(1), 108–122.  
<https://doi.org/10.1558/cj.v26i1.108-122>

*Keywords: EFL; China; university; CALL; vocabulary; learning process; learning outcomes*

Ma argues that an evaluation of a computer-assisted language learning program should examine both learning processes and learning outcomes, as well as the relationships between the two. Ma conducted an evaluation study that explored the degree of relationship between various user actions (learning process) and receptive and productive vocabulary retention (learning outcome) in a computer-assisted English vocabulary learning program called WUFUN. Participants were



50 first-year Chinese university students and were non-language majors. The study employed a pre-, immediate post-, delayed post-test design. While participants used WUFUN for a total of one class period (100min), the computer recorded user actions (e.g., amount of time spent on the exercises, initial and final scores on the exercises). Pearson correlation coefficients indicated that learning outcomes were related to the following user actions: the final score on the exercises, the number of target words in the answers to the exercises, and the number of words looked up. Step-wise regression analyses further showed that the three user actions indicated above predicted either a receptive or a productive vocabulary retention. Receptive learning was better predicted by the user action variables compared to productive learning.

**Mackay, R. (1988).** Position paper: Program evaluation and quality control. *TESL Canada Journal*, 5(2), 33–42.

*Keywords: definition; principles; stakeholders; evaluator role; utilization; pragmatic*

The article offers useful and practical guiding principles for deciding the who, the what, and the why of program evaluation. Mackay defines program evaluation as the “purposeful and systematic collection, analysis, and interpretation of information about one or more components of a particular program” (p. 34) to resolve practical programmatic concerns. Throughout, Mackay takes a utilitarian and pragmatic approach to program evaluation centered around what he refers to as “principal stakeholders” (p. 35), or those responsible for the program. Following Mackay’s approach, the evaluator’s role is to serve the needs of the principal stakeholders or other stakeholders who are affected by evaluation when no principle stakeholders are clear. Serving the principle stakeholders includes facilitating discussion of the program components to be evaluated, the evaluation purposes and uses, and the appropriate timing of evaluation, among other important decisions necessary for program evaluation. To support principle stakeholders, Mackay advises the evaluator to provide information (evidence) that is responsive, timely, relevant, credible, and comprehensible. This article is a good starting point for beginning evaluators, particularly when considering the roles and responsibilities entailed in conducting useful and effective program evaluations.

**Mackay, R. (1994).** Undertaking ESL/EFL programme review for accountability and improvement. *ELT Journal*, 48(2), 142–149.

*Keywords: Indonesia; intrinsically motivated; extrinsically motivated; improvement; accountability; ODA; program/project based review model; framework*

Mackay distinguishes extrinsically-motivated evaluation, which is a top-down bureaucratic approach to evaluation, mainly for accountability purposes, from intrinsically-motivated

evaluation, which is undertaken by program personnel with a focus on the improvement of the program. Based on his experience on the evaluation of projects in fourteen language centers in Indonesia, Mackay proposes a program/project-based review model. This model is based on an intrinsically-motivated evaluation system but seeks to satisfy both (internal) improvement and (external) accountability demands. The model involves the following steps: (a) conceptualization of the program/project as a whole; (b) review of the program/project components and possible focus (e.g., staff, resources, curriculum etc.) within each component; (c) identification of key areas of each focus (e.g., quality of teaching, each course, etc.); (d) setting performance indicators of the key areas to estimate effectiveness; (e) appropriate and credible data collection on each performance indicator; and (f) examination and interpretation of data among program/project personnel to arrive at judgments on the strengths and weaknesses for each key area. The gathered information can serve interests and concerns of both the program/project staff and the bureaucracy.

**Mackay, R., & Bosquet, M. (1981).** LSP curriculum development: From policy to practice. In R. Mackay & J. D. Palmer (Eds.), *Language for specific purposes: Program design and evaluation* (pp. 1–28). Newbury House.

*Keywords: LSP; curriculum development; program maintenance; needs analysis; questionnaire*

Mackay and Bosquet offer suggestions on how the process of educational decision making and curriculum development can be achieved, presenting examples of constraints from language for specific purpose programs. There are three stages in operating curriculum development. (1) At the pre-program development stage, the educational goal and its rationale will be decided by the administrative body in authority, and the intention to develop the program will be diffused to all stakeholders. The main purpose of this dissemination is to ensure and maximize the chances of obtaining valuable information. (2) At the program development stage, various factors that affect the program are identified, weighed, and contemplated. The processes of the developmental stage are similar to that of the task-based language teaching framework. The five phases that are involved in this stage are: (a) Basic information gathering phase, where learners' needs and target situations are identified; (b) Goal-specification phase, where gathered information is transformed to specify objectives; (c) Production phase, where materials and tests will be created based on appropriate target language samples identified from the needs analysis, syllabus is specified, and appropriate methodological procedures are devised and implemented; (d) Teacher-training phase, where teachers are trained in the new innovations, and where students' and teachers' perceptions on the effectiveness of the program are utilized to adjust pedagogical instruments; and (e) Trial phase, where formative and summative evaluation take place. (3) At the program maintenance and quality control stage, the quality of instruction, appropriateness of the goals, teacher training, and testing procedures are monitored. The three stages are stated in a hierarchical fashion, although they may intertwine with each other throughout the development

process. Appended are useful examples of a student needs survey to elicit problems in listening skills, and a teacher feedback questionnaire for lesson evaluation.

**Mackay, R., Wellesley, S., & Bazergan, E. (1995).** Participatory evaluation. *ELT Journal*, 49(4), 308–317.

*Keywords: Indonesia; ODA; participatory; performance indicator*

Mackay, promoting a participatory approach to project evaluation, acted as an external consultant for the English Language Teaching Projects Unit (ELTPU) in twenty-five language centers in Indonesia. A performance indicator (PI) framework was introduced at the ELTPU workshop as a diagnostic approach to program evaluation. The PI framework starts with diagnosing the focus of the project, by dividing it into key areas and defining performance indicators for each. The indicators are then broken down into two to three critical ‘themes’ to indicate what information is needed from evaluation. The relative strength and weaknesses of each PI found from the analysis will provide future direction for the language center, directed at long-term sustainability. A key factor in the PI framework is the participation of stakeholders. The stakeholders are involved in all decision-making processes, making evaluative outcomes contextually appropriate. The advantage of the PI framework is that it provides opportunities for the stakeholders to collaboratively clarify and reflect on the issues, which can lead to self-awareness-raising and empowerment for the staff.

**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-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 111–131). Longman.

*Keywords: EFL; participatory model; self-evaluation; Programme Based Review; performance indicator*

Mackay, Wellesley, Tamasan, and Bazergan illustrate how a language center in Indonesia benefited by undergoing the process of participatory self-evaluation called Programme-Based Review (PBR). This approach emphasizes “participative monitoring and evaluation activities initiated within the unit to facilitate periodic or continuous improvement by programme staff themselves” (pp. 111-112). Based on its intrinsic motivation, PBR can generate direct and relevant information for the continuous improvement of program management and teaching, as well as information for satisfying accountability demands of supervisory bodies. The contextually-bound performance indicators (Pis) clarify how the program is expected to work based on the language center goals and policies, the Pis are used as a framework for making

evaluative judgments. PBR also provides opportunities for stakeholders to take an informed role in planning and implementing evaluation, making suggestions based on evaluative outcomes, and acting for improvement.

**Markee, N. (1996).** *Managing curricular innovation*. Cambridge University Press.

*Keywords: innovation, implementation, change management, evaluation, change model, organizational development, change agent, ESL, task-based language teaching, TBLT, CATI project, diffusion-of-innovations, Project Framework matrix, linkage model*

In chapter 1, Markee notes that the gradual accumulation of both practical and theoretical knowledge in past decades has led to numerous innovations in language education. This book is concerned with educational change—what makes one pedagogy, methodology, textbook, or teaching exercise successful where others have failed—aims to equip teachers with the practical and theoretical knowledge to help understand these change phenomena. Markee introduces the CATI project, which he uses as a case study to illustrate common problems that arise in different sociocultural and educational contexts.

In chapter 2, Markee discusses six innovations in language teaching that highlight obstacles accompanying attempts at educational change, including the British Council’s international development work, the notional-functional syllabus, the process syllabus, the natural approach, the procedural syllabus, and task-based language teaching (TBLT). Markee describes each innovation, discusses factors and attributes that promoted or inhibited innovation, and lays out implications of each for educational change. Markee concludes by synthesizing all six innovations, noting that factors affecting curricular innovation include sociocultural context, comprising systemic and ethical constraints and personal characteristics of potential adopters; models and strategies of change used to manage the innovation process; and the perceived newness of an idea or practice. He lists attributes that can be used to analyze the promotion or inhibition of innovations, including relative advantages, compatibility, complexity, observability, trialability, form, explicitness, originality, adaptability, and feasibility.

In chapter 3, Markee introduces a theoretical framework for understanding innovation and the uptake of educational change—“Who adopts what, where, when, why, and how” (Cooper, 1982; 1989). Markee then breaks down each component of the theoretical framework. In defining cultural innovation, he highlights differences between innovation and change, provides insights from management theory, draws attention to links between development and innovation, distinguishes between primary and secondary innovations, and discusses the subjective nature of the ‘newness’ of innovations. For the “how” component, Markee outlines five different change models: the social interaction model; the center-periphery model; the research, development, and diffusion model; the problem-solving model; and the linkage model. He describes each and discusses their relative strengths and weaknesses.

In chapter 4, Markee re-introduces the CATI project. This chapter is divided into five sections. In the first, Markee provides the rationale behind the CATI project, its history and sociocultural context, and issues of generalizability versus understanding. In the second, Markee discusses the model of curricular innovation that underpins the CATI project; it draws on critical pedagogy and makes use of strategic, tactical, and operational planning levels to guide the innovation process. In the third, Markee discusses the strategic planning level in more detail and notes how a Project Framework matrix (Alderson, 1992) can be used to incorporate aims, goals, and evaluation criteria. He discusses the evaluation criteria of the CATI project, weakness with historically expert-driven approaches to evaluation, the CATI project's change model, and the ten attributes of innovation adoption (from chapter 3) in relation to TBLT. In the fourth and final sections, Markee describes the tactical and operational planning stages, both of which offer loci for innovation.

In chapter 5, Markee makes the case that any curricular innovation must also be managed using a model of organizational development. In addition to developing the professionalization of teachers or teaching assistants, for instance, we must also be mindful of educational capacity. Markee adapts a version of Everard and Morris' (1990) organizational development model to manage the CATI project, which has four components: communicating, knowing, monitoring, and evaluating. Communicating allows project participants to communicate the aims and goals of a project, as well as their development and research activities, with one another. Knowing consists of means for providing participants with knowledge about innovations being implemented via teacher development and research activities. Monitoring has to do with the formative evaluation of a project while evaluating with (summative) project success or failure. Markee discusses the CATI project in relation to the organizational development model. He notes that communicating is perhaps the single most deciding factor in innovation success or failure and proposes this four-part model to promote primary innovations.

In chapter 6, Markee opens by framing the CATI project within three issues and problems that affect the evaluation of any educational change: whether an evaluation should be done by insiders or outsiders, what kinds of data to gather to evaluate the project, and ways to overcome practical problems that crop up during evaluations. Markee notes that teachers are key figures in the innovation and implementation process, and so their participation is crucial. Action research, teachers' journals, and survey questionnaires are used to understand to what extent the CATI project met its aims and objectives. Markee then draws our attention to two issues that arose during an evaluation by Hutchin (1992) of the CATI project that affected its implementation; teaching assistants (TAs) did not understand that the purpose of materials development was to help professionalize them as teachers, and there was a lack of communication between the project director and TAs regarding fundamental project goals. Later, Markee discusses the evaluation of primary and secondary innovations of the CATI project. Markee concludes this chapter by noting that the success of the project came about through analysis of its failures at the

start, the innovation within the CATI project became sustainable despite regular TA turnover, action research is ongoing, and materials are continually revised and used by participants.

Markee concludes by distilling nine principles about curricular innovation and management from the CATI project. These principles are as follows: curricular innovation is complex, the principal job of a change agent is to effect desired change, good communication among project participants is key, a successful implementation of educational innovations is based on a strategic approach to managing change, innovation is a messy and unpredictable process, effecting change takes longer than originally anticipated, change agents' proposals are likely to be misunderstood (at least at first), it is important for implementers to have a stake in the implementation, and it is important for change agents to work through opinion leaders to help in influencing peers. Markee concludes the chapter by calling for more research into how innovations are implemented and sustained in different contexts and experimental research to complement the perspective on change management presented in this book.

**Martinez Agudo, J. d. D. (2017).** What EFL student teachers think about their professional preparation: Evaluation of an English language teacher education programme in Spain. *Australian Journal of Teacher Education*, 42(8), 62–76.  
<https://doi.org/10.14221/ajte.2017v42n8.5>

*Keywords: pre-service English teacher, teacher education, English as a foreign language context, teacher education program, Spain*

Cognizant of the dearth of evaluation of pre-service English teacher education programs, this study investigates the strengths and weaknesses of the teacher education program from the perspective of student teachers in the EFL context. The study specifically asks whether the current program satisfies the needs of the student teachers, and how it helps the teachers to be well-prepared for teaching English.

The study, based on both quantitative and qualitative methods, collected data through a questionnaire adapted from Peacock (2009) as well as semi-structured interviews. The 58 student teachers were enrolled in the four-year EFL teacher education program at the University of Extremadura located in Spain. They were in their final year of the program, taking the second practicum course in which the student teachers are required to teach in real classrooms.

Ten interviews were conducted to triangulate the quantitative data from the questionnaire.

The results showed that the students were generally satisfied with the program in terms of its flexibility in utilizing different teaching methods, the students' ability to use diverse teaching materials, its promotion of teacher reflection, and the coherence of the program elements. In particular, student teachers found it helpful to reflect on their past teaching experiences. At the

same time, the survey results also revealed that students felt unsure about whether the program satisfied their needs. This, according to the author, may be because they had little understanding of their needs.

In terms of the weakness of the program, around 55% of the students thought that the program was lacking adequate training in developing their English skills. Interview data corroborated these survey results. Student teachers generally thought that the program had its strong points in developing students' pedagogical knowledge. The knowledge gained from the courses helped them enhance their teaching skills and confidence in real classrooms. In contrast, many students pointed out that their English skills should be improved to a great degree. Part of the reason, according to many students, was that (a) students take the practicum course four years into the program and (b) there are not enough courses that enhance their understanding of the English language. Other weaknesses of the program included the lack of training in classroom management skills and the high emphasis on theories rather than practices. Based on these results, the author suggests giving students more opportunities to practice teaching skills in addition to emphasizing linguistic and management skills.

**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). Heinle Cengage.

*Keywords: Spanish; FL; college; quantitative; qualitative; questionnaire; interview; observation; document analysis; beliefs; specific purposes*

This book chapter examines the perceptions of university instructors, administrators, and students regarding the existence of separate tracks for a Spanish language for specific purposes (LSP) program. Using questionnaires, interviews, class observations, and document analysis, the study aimed to identify student needs, teacher and administrator beliefs, and the relationship between students' beliefs and the teacher/administrator beliefs regarding Spanish LSP. Four program administrators completed interviews, 15 teachers completed a pedagogical views questionnaire, and 129 students in three different tracks (general, specific, and advanced) completed a beliefs about language learning (BALLI) questionnaire. Findings about beliefs from the questionnaires differed within the groups and between the teacher and student groups related to motivation for taking courses for specific purposes, the existence of separate tracks, and differences between the goals of the program and the methods used to achieve those goals. The study provides implications for course design that is based on student needs analysis and discourse/genre analysis, and allows for flexibility given certain contextual factors that may otherwise conflict with teachers' beliefs. The questionnaires are included in the appendix.

**Martinez-Lirola, M., & Rubio, F. (2009).** Students' beliefs about portfolio evaluation and its influence on their learning outcomes to develop EFL in a Spanish context. *International Journal of English Studies*, 9(1), 91–111. <https://doi.org/10.6018/ijes.9.1.90661>

*Keywords:* Spain; university; EFL; portfolio; student evaluation; learning outcomes; impact on learning; survey

The authors examined whether undergraduate students' attitudes and motivations toward portfolio assessment affect learning outcomes. Two survey studies took place at two universities in Spain with students who were mostly new to portfolio assessment. The study examined (a) students' motivation and attitude towards portfolio use, (b) perceived efficacy of learning using portfolio assessment, and (c) advantages and disadvantages of using portfolio assessment. Part of the study involved portfolio (experimental) and non-portfolio (control) groups, but it remains unclear what the perceptual differences were between the two groups, since the comparative results were not reported in this study. The study findings should be interpreted with caution, as the study design and the constructs of the questionnaire lack clarity.

**Mathews, T., J., & Hansen, C., M. (2004).** Ongoing assessment of a university foreign language program. *Foreign Language Annals*, 37(4), 630–640. <https://doi.org/10.1111/j.1944-9720.2004.tb02429.x>

*Keywords:* US; French; German; Spanish; university; portfolio; NCATE; ACTFL; assessment; top-down; accreditation

Mathews and Hansen report on the process and results from the first-year of assessment of foreign language programs (French, German, and Spanish) at Weber State University. All students in lower level courses (each semester) and all potential teaching majors and minors take modified or unmodified ACTFL Oral Proficiency Interviews (OPIs). The ACTFL OPI was selected as the primary assessment tool due to the implementation of assessment standards in 2004, developed by the National Council for Accreditation of Teacher Education (NCATE) and ACTFL. Beyond this immediate demand, the foreign language department in this study took initiative in 1998 to develop a departmental mission statement, and they set student learning outcomes in reference to National Standards. In 2000, they proposed a senior assessment portfolio (a computerized oral proficiency test and writing samples) and created rating rubrics. For this study, all graduating seniors' portfolios were assessed for their achievement of stated outcomes. The purposes of the evaluation study were: (a) to examine to what extent the department's curriculum and requirements are helping the students to maximize proficiency; (b) "to check department's progress in incorporating the National Standards into the curriculum" (p. 630); and (c) to assess students' achievement of the departments' outcomes and judge whether the goals are reasonable.



**Matthies, B., F. (1991).** Administrative evaluation in ESL programs: “How’m I doin’?” In M. C. Pennington (Ed.), *Building better English language programs* (pp. 241–256). NAFSA.

*Keywords: ESL; program administration; directors; professional development*

Similar to Fox (see chapter 11), Matthies illustrates the importance of directors’ on-going improvement throughout their career by building on areas highlighted via evaluation. Program administration can be understood in comparison with three distinct types of criteria: (a) professional guidelines and other ESL institutions (national standards); (b) parent institution’s mission (the institution); and (c) the ESL program itself. Based on survey research, the author identifies the following six key job skills for administrators: communicating, planning, educating, organizing, evaluating, and negotiating. Support staff, instructional staff, students, and parent institutions that are directly related to the ESL director, should be able to evaluate administrative effectiveness. Matthies provides several evaluation measures as examples: self-evaluation of current job status, formal/informal feedback, written feedback, checklist response (Appendix A), and student survey (Appendix B).

**McAlpine, D., & Dhonau, S. (2007).** Creating a culture for the preparation of an ACTFL/NCATE program review. *Foreign Language Annals*, 40(2), 247–259.  
<https://doi.org/10.1111/j.1944-9720.2007.tb03200.x>

*Keywords: NCATE; accreditation; assessment; OPI; electronic portfolio; university*

McAlpine and Dhonau reflect on and illustrate the acculturation process for NCATE/ACTFL review at the Department of International and Second Language Studies at the University of Arkansas at Little Rock. Since its inception in 2002, the ACTFL/NCATE Program Standards for the Preparation of Foreign Language Teachers has heightened teacher education programs’ attention to quality assurance of teacher candidates’ knowledge, skills, and dispositions of language, culture, and literature, knowledge of assessment, and professionalism. McAlpine and Dhonau provide six suggestions from their experience undertaking NCATE/ACTFL accreditation: (a) foster collaboration and engagement from both the foreign language department(s) and the College of Education; (b) establish capacity, infrastructure, and culture of oral proficiency testing (ACTFL oral proficiency test); (c) familiarize faculty with the Standards for Foreign Language Learning; (d) revise and align curriculum around the ACTFL/NCATE Content Standards; (e) implement an assessment system to gather evidence on the content standards; and (f) create data/artifacts management system utilizing technology, such as electronic portfolios. The article also provides a useful timeline and activity list for preparing for NCATE/ACTFL program review as an Appendix.

**McDonough, K., & Chaikitmongkol, W. (2007).** Teachers' and learners' reactions to a task-based EFL course in Thailand. *TESOL Quarterly*, 41(1), 107–132.  
<https://doi.org/10.1002/j.1545-7249.2007.tb00042.x>

*Keywords: Thailand, EFL, TBLT, task-based language teaching, reactions, evaluation, learning notebooks, observations, course evaluations, interviews, field notes, learner independence, grammar, academic needs, evaluation cycles, implementation, checklist*

This study examines teacher and learner reactions to a task-based EFL course at Chiang Mai University (CMU) in Thailand. In response to policy changes by the Thai government, the English department of CMU revamped the integrated-skills courses to foster cross-cultural communication, promote life-long learning, and achieve personal and academic goals. To do this, a team of Thai researchers created a task-based course and then subsequently piloted and revised it over a 12-month period. Course content and tasks were developed to reflect learners' real-world interests and to promote cognitive and metacognitive learning strategies. Each task consisted of nine, five-minute classes. Learning in the course was assessed via three oral and written task performances, in-class quizzes, and a final exam. Participants in this study were 35 L1 Thai EFL learners in the English department between 17 and 19 years-old. Thirteen teachers participated in the course. The researchers wanted to know about Thai teacher and learner reactions to the task-based course, and, if they had concerns, how those were addressed. Data-collection instruments were post-task evaluations, learning notebooks, observations, course evaluations, interviews, and field notes. Data was analyzed qualitatively by examining the entire corpus and pulling out themes that reflected teachers' and learners' reactions. The researchers found that the course increased learner independence; was lacking in grammar teaching and learning, but this perception decreased over time; and that the course was relevant more for academic than non-academic purposes. Concerns were addressed in future iterations of the task-based course (e.g., through workshops). Implications of the study are that students' real-world needs can be academic ones, that teachers and learners require support when transitioning away from traditional L2 teaching methods, and task-based courses should provide opportunities for teachers to respond to learners' needs as they arise. The authors remind us that cycles of systematic evaluation are needed in task-based courses, and they provide a checklist for administrators and instructors implementing and/or evaluating task-based courses.

**Middlebrook, G. C. (1991).** Evaluation of student services in ESL programs. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 135–154). NAFSA.

*Keywords: student service; administration; ESL*

Middlebrook addresses an administrative aspect in evaluating language programs, namely, the need to evaluation student services components. He provides lists of useful evaluative questions

to be asked for each component (recruitment, admissions, orientation, employment, advising, financial aid, and housing) of student services in an ESL program. The guidelines he proposes feature the need for: (a) “thoughtful, pragmatic and well-articulated institutional policy” (p. 148) that will allow the program to specify the goals and objectives; (b) staff to possess “requisite skills and knowledge” (p. 148); and (c) adequate funding for student services components to function as stated in program policy.

**Milleret, M. (2008).** The trials and tribulations of comprehensive program evaluation. *ADFL Bulletin*, 39(2&3), 44–48.

*Keywords: US; university; Portuguese; needs analysis; program improvement; program development; qualitative; quantitative; survey; focus group; IRB; case study*

Milleret recounts a program evaluation which was part of a program development project in a Portuguese program at a US university. The program development project was initiated by members of the department in an attempt to address three issues: poor course articulation, the need for program growth, and concern about the needs of Spanish-speaking students in the program. The needs analysis evaluation used both qualitative and quantitative data and the main data collection instruments were surveys, focus groups, and program documents. The evaluation team encountered numerous challenges during the evaluation including very limited funds, poor responses to focus group requests, and delays due to the IRB process. Despite the challenges, the evaluation produced data to guide program development including the design of new courses to meet student interests and the development of a special course sequence for Spanish-speaking students. The Portuguese program continues to utilize data from the evaluation as it aims for continued growth and improvement.

**Miranda, J. L. W., Wells, J. C., & Jenkins, A. (2019).** Preparing special education teacher candidates to teach English language learners with disabilities: How well are we doing? *Language Teaching Research*, 23(3), 330–351.  
<https://doi.org/10.1177/1362168817730665>

*Keywords: special education (SPED), teacher education program, learners with disabilities, utilized-focused framework, United States*

The study investigates special education teacher preparation programs in the U.S. using an utilized-focused framework (Patton, 2008). The purpose of the evaluation was to investigate the degree to which special education (SPED) teacher preparation programs prepared teacher candidates from the special education department in a public university to teach English language learners (ELLs) with disabilities. In particular, the evaluation asked (a) whether

teachers gained the necessary knowledge and skills for teaching ELLs with disabilities, (b) whether the programs consisted of the necessary components, and (c) the extent to which the programs improved the teacher's perceived competence to teach ELLs with disabilities.

Based on the literature review, the authors developed a framework that organizes and analyzes the data comprising four content elements: foundations of second language acquisition, sociocultural and political aspects of ELLs, practices and strategies of ELLS, and assessment practices of ELLS.

Data collection was conducted using a document review, faculty questionnaire, a faculty focus group, and a teacher candidate survey. The findings showed that there were inconsistencies in the reported ELL content and the actual content of the course materials. The curriculum maps also revealed that components of teaching ELL were not persistently integrated in the programs. In particular, among the four content elements, the foundations of second language acquisition and assessment practices were incorporated into the program the least. The lack of a cohesive and organized approach to SPED led to teacher candidates' low confidence levels when teaching ELLs with disabilities as well as their lack of necessary knowledge.

**Mitchell, R. (1989).** Second language learning: Investigating the classroom context. *System*, 17(2), 195–210. [https://doi.org/10.1016/0346-251X\(89\)90033-X](https://doi.org/10.1016/0346-251X(89)90033-X)

*Keywords: Scotland; French; Gaelic; bilingual education; secondary; elementary; foreign language; communicative approach; classroom observation; interview; assessment; action research; retrospective; political*

Mitchell first reviews foreign language classroom research (1976-1986) and then introduces an evaluation study of Gaelic-English bilingual programs (1984-1986), undertaken at the University of Stirling, Scotland. The purpose of each project varied. First, the classroom research studies were mainly concerned with the actual instructional practices (teachers' use of the target language, skills used in instruction that influence learners' FL experience, the use of message-oriented activity types, routinization of communicative methodology) and teachers' views on methodology, during the shift towards communicative language teaching in Scotland. Some of the instruments created for the classroom research included systematic observation instruments, teacher questionnaires and interviews, assessments of students' achievement and attitudes. Second, suffering from political tensions between the Local Education Authority and the Scottish Education Department over continuation of the Bilingual Education Project (BEP), an evaluation study of the Gaelic-English bilingual program was undertaken. This independent retrospective evaluation of bilingual programs utilized teacher and parent interviews, qualitative classroom observation, and assessment of students' Gaelic and English proficiency to document the implementation of the BEP and to examine learners' achievements and factors that influence their achievements. This language program evaluation study is one of the early examples of a

non-experimental approach to program evaluation. Mitchell notes that “Evaluators have a duty to address all process and product variables which are important for participants in the programme, and not only those few which can be experimentally linked; the “clear picture” may involve ethnographic portrayals as well as quantitative accounts: and evaluation reports must be informative not only for decision-makers in the specific context of the programme under study, but also for others considering the future development of similar programmes” (p. 207).

**Mitchell, R. (1990).** Evaluation of second language teaching projects and programmes. *Language, Culture, and Curriculum*, 3(1), 3–15.

*Keywords: Scotland; bilingual education; Gaelic; English; qualitative*

Mitchell argues for language program evaluation studies to move away from experimental and quasi-experimental research, even if internal validity is supported by the systematic gathering of information about the process of language acquisition (as suggested by Long, 1984). She suggests this shift due to the complex nature of what is actually involved in a program. Thus, studies that are not bound by experimental expectations can address a wider range of questions. As an example of a non-experimental study, she illustrates an evaluation with a many-faceted design which took place at the Gaelic-English bilingual primary school program in Scotland. Evaluative techniques included unstructured classroom observations, teacher and parent interviews, and writing and speaking tests. Mitchell emphasizes that evaluation is not about revealing the causal relationships between a small number of variables, but rather inferring the most likely relationships of complex events that interact by monitoring intended outcomes, identifying the unexpected, and proposing untried solutions (p. 11)

**Mitchell, R. (1992).** The “independent” evaluation of bilingual primary education: A narrative account. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 100-140). Cambridge University Press.

*Keywords: Scotland; Gaelic; English; elementary; bilingual education; retrospective; political; outsider*

Mitchell presents an independent, retrospective evaluation of a Gaelic-English bilingual primary school program in Scotland, where Mitchell and her colleague, being ‘outsiders’, had to negotiate a rough political climate during the planning and implementation of an evaluation project. The instruments were restricted by local administrative bodies. Evaluators were not to question students’ attitudes towards culture. Further, the Scotland Education Department required use of a quantitative-experimental methodology. As a result of these restrictions, the evaluators’ array of options was curtailed. In order to select ten site schools for study, a

preliminary interview survey with the head teachers of all schools was conducted, querying the history of their school involvement with the bilingual education project (BEP), attitude towards BEP policies, and the present status of program implementation. Two classrooms within each school were selected, observed, and participants interviewed. Students' Gaelic and English proficiency, perceptions, and attitudes towards the BEP (but not culture) were measured. In addition, a limited number of parents were interviewed for their knowledge, involvement, and attitudes towards BEP. This case provides a good example of the potential confrontation between evaluators and educational/political authorities, especially when the evaluation proceeds with high stakes on the line.

**Moreno-Lopez, I., Saenz-de-Tejada, C., & Tami, K. S. (2008).** Language and study abroad across the curriculum: An analysis of course development. *Foreign Language Annals*, 41, 674–686.

*Keywords: FL; content and language integration; college; study abroad; federal funding; survey; teachers; perceptions; quantitative*

This article describes the design and pilot implementation of one college's Title VI-funded project, Integrated Intensive Courses Abroad (IICA), which includes content courses, taught in the target language by faculty representing different academic disciplines, and a 3-week study abroad experience in the middle of the semester-long course. Considerations for the program implementation include the identification of eligible students, the cross-disciplinary training of content faculty to offer the course in the target language, the identification of the study abroad site and service learning component, and the seven expected outcomes as articulated by the Title VI grant. In addition to a discussion of the college's ability to meet the expected programmatic outcomes, the article reports the results of a post-course Likert-scale survey completed by 8 of the IICA professors regarding their perceptions of the effectiveness of the program for improving students' L2 proficiency and content knowledge. No student results or perceptions are included in this report. Recommendations for future improvements are made, with a particular emphasis on the needed language and content cross-training of the faculty teaching these integrated courses.

**Morris, M. (2006).** Addressing the challenges of program evaluation: One department's experience after two years. *Modern Language Journal*, 90(4), 585–588.

*Keywords: assessment; multiple languages; electronic portfolio; modified SOPI; interview; university*

Morris reflects on an evaluation of how well the department of Foreign Languages and Literatures department at Northern Illinois University prepares its majors. The main data collection tools were: (a) a best-works electronic portfolio that consisted of a number of different components including artifacts that demonstrated learners' ability in cultural understanding, reading, speaking, and writing; (b) a written-exit questionnaire; (c) a follow-up alumni questionnaire; and (d) a modified SOPI test. In choosing appropriate instruments, the faculty at Northern Illinois University carefully modified existing instruments to reflect the departmental student learning outcomes. Evaluation results informed curriculum improvement, including the creation of a new course and changes in class scheduling, and also raised questions about the efficacy of the evaluation methods for cultural learning and oral proficiency. Morris's narrative illustrates the organic, responsive, and reactive nature of program evaluation (or "messiness" in Morris's words). Engaging in program evaluation also led to increased faculty awareness and communication about the program.

**Murdoch, G. (2000).** Introducing a teacher-supportive evaluation system. *ELT Journal*, 54(1), 54–64. <https://doi.org/10.1093/elt/54.1.54>

*Keywords: EFL; United Arab Emirates; university; personnel evaluation; teacher performance review; teacher appraisal; system review; survey*

Murdoch advocates the necessity of a supportive and improvement-oriented performance review, and outlines five features of what he calls a progressive teacher-performance review. He argues for a teacher-performance review that (a) promotes reflective practice; (b) empowers and motivates teachers by actively engaging teachers in the evaluation process; (c) examines various elements of professional activity via multiple sources and perspectives; (d) incorporates students' views on teacher performance; (e) fosters collaborative dialogue between teachers and their supervisor/director. To illustrate how such principles can be operationalized, the author describes a teacher personnel-evaluation system at the United Arab Emirates (UAE) University English program. The article also summarizes survey findings of teachers' perceptions on the performance evaluation system and procedures at the UAE University.

**Myers, L., & Lindsay, N. (2014).** Do we speak the same language?: The iterative development of an institutionally mandated foreign language assessment program. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 208–229). Cengage Learning.

*Keywords: assessment program; university-wide mandate;*

Lindsay Myers and Nathan Lindsay's, the assistant vice provost of assessment and the coordinator of the first-year French program at the University of Missouri, St. Louis, provide a case study describing the implementation of an assessment program within the Foreign Language and Literatures department in response to a university-wide mandate. Temporal, financial, and organizational challenges are presented as well as positive outcomes within the department, such as collaborative reflection among faculty and staff and learning to "speak the same language" as administrators and external stakeholders.

**National Research Council of the National Academies. (2007).** *International education and foreign languages: Keys to securing America's future*. The National Academies Press.

*Keywords: Title VI; Fulbright-Hays; national security; global economy; large-scale evaluation; language centers; federal programs*

This is a national-level evaluation report of the Title VI and Fulbright-Hays (Title VI/FH) programs, federal programs that support higher education with the aim of building foreign "language abilities or knowledge of world regions and international issues" (p. 60). The National Academies formed the Committee to Review the Title VI/FH International Education Programs.

Part One of the book reviews the need for foreign language, area, and international expertise in the U.S., the historical context and implementation of Title VI/FH programs, and the unique roles Title VI/FH has in relation to other federal programs, such as Title VI/FH's broad role compared to other federal programs aiming at capacity building for national security.

Part Two addresses the effectiveness and adequacy of eight key performance areas, referred to as program missions, defined by the U.S. Congress. Examples of the eight areas include "reducing shortages of foreign language and area experts" (p. 2) and "supporting research, education, and training in foreign languages and international studies" (p. 2). The Committee first created a logic model specifying "resources/inputs, activities, outputs, [short-term and long-term] outcomes, and impact expected of the program" (p. 313-314). The assessment of the expected outcomes was restricted due to insufficient data and few systematic program evaluation studies. For this reason, the Committee extracted data from only a few program evaluation studies, program monitoring data, grant applications, funding data, commissioned papers and targeted analyses, written commentaries from experts, public testimonials, and site visits. Although the evidence was too limited to draw conclusions and recommendations, the Committee indicated that Title VI/FH programs have played an important role as foundation-builder for "internationalizing higher education" (p. 242). Furthermore, Title VI/FH programs were shown to have focused universities' attention on building pedagogical resources and teaching foreign languages and area studies with particular focus on the less commonly taught languages.



The final section, Part Three, offers the Committee's recommendations for future Title VI/FH programming strategies. The appendices provide detailed information on legislative history of Title VI/FH, logic models, summaries of Title VI/FH evaluation studies, and site visit interview questions.

**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.  
<https://doi.org/10.1016/j.tate.2011.07.001>

*Keywords: French; Tahitian; heritage language; primary; proficiency test; quantitative*

This study examined the impact that a Tahitian heritage language program had on early elementary school students' proficiency in Tahitian and French oral and reading skills from kindergarten to first grade. An experimental group of 81 children received one hour of Tahitian instruction per day, while a control group of 44 children received only French instruction. Data collection included a parents' demographic and language background questionnaire and a battery of vocabulary, oral communication, and reading tests in French and Tahitian. Results from the language tests indicated that the experimental group's proficiency in Tahitian improved without negatively impacting their proficiency in French, leading the authors to conclude that the heritage language program does not have subtractive effects on the students' bilingual abilities. Recommendations are made for Polynesian language and cultures programs that offer a range of curriculum and instructional opportunities for the different types of students in attendance.

**Norris, J. M. (2006).** The why (and how) of assessing student learning outcomes in college foreign language programs. *Modern Language Journal*, 90(4), 576–583.

[https://doi.org/10.1111/j.1540-4781.2006.00466\\_2.x](https://doi.org/10.1111/j.1540-4781.2006.00466_2.x)

*Keywords: student learning outcomes; assessment; evaluation; measurement; definition; university*

Norris calls for college foreign language (FL) educators to reconceptualize student learning outcomes (SLO) as a means of assuring not only educational quality and effectiveness but also improvement, development, and even defense/survival of existing programs. Norris' reconceptualization also stresses evaluation as an opportunity for programs to define and articulate their values. In order to conduct SLO assessment that is informative and action-oriented, Norris argues (based on learning from evaluation work to date) that it must be participatory, feasible, useful, credible, relevant, timely, understandable, and clear to the intended users of assessment. To achieve such useful SLO assessment practices, Norris suggests three key

steps. First, FL departments must resolve terminological confusion by distinguishing among measurement, assessment, and evaluation. Norris emphasizes that SLO assessment should be free from technocracy of measurement issues and should have a system that bridges learner information and program use, a use that “[helps] educators deliver better programs and [helps] students achieve valued learning outcomes” (p. 582). Second, FL professionals should respond to the accountability movement as an opportunity to build assessment capacity and rethink and redirect programs. Thirdly, educators need to foster programmatic and evaluative thinking for program development as well as understanding the role of assessment in their programs. By way of example, Norris introduces three examples of FL program evaluation at the university level that focus on SLOs: Byrnes (2002), Dassier & Powell (2001), and Liskin-Gasparro (1995).

**Norris, J. M. (2008).** *Validity evaluation in language assessment*. Peter Lang.

*Keywords: placement test; program evaluation; validity; improvement; curriculum; C-test; German; university*

Norris illustrates the challenges that face college foreign educators in evaluating and ensuring the quality of educational assessments. Instead of applying conventional validity criteria from psychometric traditions to evaluate educational assessments, Norris advocates the reconceptualization of assessment validation as “validity evaluation.” Validity evaluation requires: (a) the treatment of an educational assessment as a coherent program; (b) the acceptance of a variety of purposes for assessment validity evaluation; (c) the prioritization and contextualization of evaluation purposes; and (d) the selection and articulation of evaluation methods to meet prioritized purposes. To exemplify validity evaluation, Norris applied the framework to the development and implementation of a placement assessment program in the Georgetown University German Department. Facilitated by Norris, the department specified the intended uses of assessments in their program, developed an assessment program aligned with the curriculum, implemented and revised the placement assessment program, and sustained the assessment program by taking action based on evaluation findings. The comprehensive validity evaluation of the assessment program not only had positive consequences for the innovative curriculum but also helped the faculty members to reconceptualize and transform assessment and educational practices in their program.

**Norris, J. M. (2009).** Understanding and improving language education through program evaluation: Introduction to the special issue. *Language Teaching Research*, 13(1), 7–13. <https://doi.org/10.1177/1362168808095520>

*Keywords: improvement; accountability; context; participation; multiple methods*

In his introduction to a special issue on evaluation, Norris emphasizes the need to focus on the use of evaluation for improving language programs and teaching practices. He argues that the increased demand for evaluation resulting from greater emphasis on accountability provides an opportunity for evaluators to increase awareness of evaluation for developmental purposes. Norris provides an overview of the five articles in the journal; each one an example of how evaluators had to respond to various constraints and pressures, yet managed to conduct evaluations that were useful to stakeholders, sensitive to political contexts, and provided valuable information for program improvement purposes. Despite their diverse contexts, the examples had the following characteristics in common: (a) the participation of language teachers in the evaluation; (b) the use of multiple methods of data collection, which provided for a better understanding of the various factors and perspectives involved; (c) the contextualization of findings to prevent their misuse or misinterpretation; and (d) communication with stakeholders to increase the likelihood that findings are understood and used. In proposing next steps for the field, Norris advocates for evaluators to take a more proactive, instructional role, helping stakeholders, participants, and audiences see evaluation as a means of improving language programs and teaching practices, rather than only as an accountability tool.

**Norris, J. M. (2016).** Language program evaluation. *The Modern Language Journal*, 100, 169–189. <http://www.jstor.org/stable/44135002>

*Keywords: language program evaluation, applied linguistics, accountability evaluation, student learning outcomes assessment, internal program evaluation.*

This review traces the development of language program evaluation within the field of applied linguistics, which is interdisciplinary in nature. A few waves have provided the momentum for the development of program evaluation: accountability evaluation, student learning outcomes assessment, and internal program evaluation efforts. Based on the nature and the origin of program evaluation, the author then prospected the future of program evaluation in light of innovations and challenges. On the one hand, innovations will be driven by adapting to the heuristics of other domains and responding to the dynamic nature of language programs. On the other hand, a dearth of training in program evaluation, and tensions related to the subjectivity of the evaluation and its repercussions will pose challenges to the potential of language program evaluation. The review concludes by highlighting that program evaluation contributes to our understanding of theories and facilitates meaningful application.

**Norris, J. M. (2015).** Thinking and acting programmatically in task-based language teaching: Essential roles for program evaluation. In M. Bygate (Ed.), *Domains and directions in the development of TBLT: A decade of plenaries from the international conference*. John Benjamins.

*Keywords: TBLT, task-based language teaching, design, implementation, development, evaluation, program, pragmatism, experiential learning, U-FE, utilization-focused evaluation, holistic*

In this article, Norris recalls that task-based language teaching (TBLT), in its most complete form, is not simply an additional technique for the language educator's toolbox but is rather an architecture for the whole-scale design, implementation, development, and evaluation of educational programs. Norris reminds us that TBLT's philosophical and educational origins in pragmatism, process philosophy, and experiential learning advocated for an educational enterprise that emphasized the indivisibility of knowledge and its application, as well as considered the many—often messy—factors at play in the context of holistic, real-world educational settings. Norris discusses several research studies of both proponents and critics of TBLT alike, remarking that, while desiring to address program-level components (e.g., needs, materials, assessment, etc.), neither side has looked at TBLT holistically. Such a tradition then begs the question of how to treat the 'big picture' of TBLT and where to start. Norris proposes program evaluation as a framework for guiding task-based inquiry and innovation into program-scale phenomena. Program evaluation, with its emphasis on use and users, is decision-oriented; offers ways to identify, prioritize, and resolve issues for program stakeholders; is multi-methodological; and acts as a frame of reference for coping with the messy and situated realities of programs on real time-scales. Norris then turns to five published examples of task-based evaluation, contextualized in actual programs, to illustrate the diverse uses, methods, and contributions of each to an understanding of TBLT as a holistic venture. From his discussion of these studies, Norris concludes by calling for a formative approach to TBLT evaluation inquiry focusing on the design, implementation, and understanding of language programs as well as systematic narratives that document TBLT in practice in order to inform our understanding of those task-based ideas being used; how design and implementation decisions are ultimately made; and the individuals and circumstances that determine actual task-based classrooms.

**Norris, J. M., Davis, McE. J., Sinicrope, C., & Watanabe, Y. (Eds.). (2009).** *Toward useful program evaluation in college foreign language education*. National Foreign Language Resource Center.

*Keywords: university; Chinese; French; German; Italian; Portuguese; Spanish; teacher training; major; study abroad; student learning outcomes; program change; program value; needs; intercultural competence; portfolio; focus group; alumni survey; interview*

This edited book is one product of a federally funded project entitled “the Foreign Language Program Evaluation Project” (principal investigator: John Norris). The book is comprised of nine chapters. The first chapter by Watanabe, Norris, and Gonzalez-Lloret reports on a nation-wide survey study that investigated current program evaluation practices, needs, issues, and capacities in U.S. college foreign language (FL) programs.

Chapters two through eight are example evaluation projects in various language programs with unique evaluation purposes, intended uses, focus, and data collection methodologies. The authors of the case studies participated in a Summer Institute on program evaluation hosted by the National Foreign Language Resource Center and implemented a utilization-focused approach to program evaluation in their local FL programs. Their evaluations focused on various elements of a program, for example, curriculum design and effectiveness (chapter 2: Milleret & Silveira), student learning outcomes (chapter 5: Walther; chapter 6: Grau Sempere, Mohn, & Pieroni), teacher training needs (chapter 4: Zannirato & Sánchez-Serrano), value of a FL major (chapter 8: Pfeiffer & Byrnes), study abroad experiences (chapter 7: Ramsay), and program visibility and associated administrative changes (chapter 3: Loewensen & Gómez). In addition to different designs and uses of evaluation, interesting here are the transformational effects of evaluation, which authors articulate in the lessons learned section of each chapter.

In the final chapter, Davis, Sinicrope, and Watanabe summarize positive program evaluation practices that were observed across the seven case studies and suggest strategies to respond to and act on internal and external evaluation needs, ways to approach use-driven outcomes assessment, and some considerations in professionalizing evaluation in the field of foreign language education.

**Norris, J. M., & Mills, N. (2014).** Innovation and accountability in foreign language program evaluation. *AAUSC series issues in language program direction*. Heinle.

*Keywords: integration of professional standards; university benchmarks; departmental goals; outcomes assessment in language program evaluation*

The volume provides language program directors with an overview of innovative methodologies, guidelines, and frameworks in language program evaluation and includes topics such as the integration of professional standards, university benchmarks, departmental goals, and outcomes assessment in language program evaluation. Ideas and examples emanating from the volume should equip language program directors with tools and knowledge to help innovate and otherwise transform their language programs in response to the pressing needs to do so. The volume is organized in two overarching sections. In the first section (chapters 1 – 5), readers are presented with various methodologies, guidelines, and frameworks for language program evaluation. In the second section of the volume (chapters 6 – 10), readers are presented with

various language program evaluation initiatives and their relation to professional, institutional, and departmental goals and identities.

**Obando Guerrero, G. V., & Sánchez Solarte, A. C. (2018).** Learners' satisfaction in two foreign language teacher education programs: Are we doing our homework? *HOW Journal*, 25(1), 135–155. <https://doi.org/10.19183/how.25.1.436>

*Keywords: foreign language teacher education programs, English as a foreign language, French as a foreign language, learners' satisfaction, Colombia*

In a quest to investigate how participants evaluate a teacher education program, the study explores learners' satisfaction in two government-accredited foreign language teacher education programs at a Colombian public university. One program had both English and French as foreign languages while the other program only had English as a foreign language with Spanish as a native language. 70 students in total were randomly chosen from the two programs. The chosen students included learners from all different stages in the program. The study conducted a survey that included both closed and open-ended questions related to learners' satisfaction with the program.

The results revealed that students were generally satisfied with both of the programs. Advanced learners (i.e. students in their last year), seemed to be more satisfied with the program compared to the learners in their first or second year. This, according to the authors, was because the advanced learners have already taken courses across the fields (from methodology to assessment) and understood their role as teachers.

Students also made some suggestions for improving the efficacy of the program. First, they mentioned the importance of observing the diverse roles of teachers inside the classroom. Second, they commented on the inconvenience of a large class as the large class size prevented them from being engaged in classroom activities as well as interacting with other students. Third, students thought that the class should be conducted in the target language to improve their target language skills. Some students who planned to pursue graduate studies expressed that they need to learn test-taking skills to gain high scores in standardized tests such as TOEFL. Fourth, students also mentioned that the classroom size affects the quality of feedback, underscoring the significance of assessments that are diverse, authentic and challenging. Last but not least, the students also suggested that their voices be reflected in actual class contents through teacher evaluation. Based on these results, the authors conclude the paper by emphasizing the importance of giving attention to learners' satisfaction when evaluating teacher education programs, so that positive changes may be introduced.

**Oliver, K., Kellogg, S., & Patel, R. (2012).** An investigation into reported differences between online foreign language instruction and other subject areas in a virtual school. *CALICO Journal*, 29, 269–296.

*Keywords: secondary; FL; CALL; quantitative; qualitative; survey; student perceptions; teacher perceptions*

This large-scale study used quantitative and qualitative survey methods to capture high school foreign language learners' perceptions of L2 courses offered online. End-of-course evaluation surveys (Likert-scale) were completed by 559 L2 students to measure their perceptions of learning, teacher preparation, course content and teaching, and critical non-teaching factors. Results of this survey indicated that the foreign language online courses were rated significantly lower than online courses in other content areas. A follow-up survey (open-ended) was completed by 119 students and 19 teachers, and provided more in-depth explanations for the lower ratings of L2 courses related to limited interaction with teachers and classmates, difficulty of foreign language content, and student characteristics. Participants' open-ended responses from the follow-up survey were coded and analyzed to formulate recommendations for future design and implementation of online L2 course instruction and for training L2 teachers delivering online L2 courses. The original and follow-up surveys are included in the article appendices.

**Palmer, A. (1992).** Issues in evaluating input-based language teaching programs. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 144–166). Cambridge University Press.

*Keywords: German; university; experimental curriculum; Krashen; comparative method; testing; attitude; journal; questionnaire*

Palmer describes various decision-making issues he and his colleagues faced during the evaluation of an eight-month experimental first-year German course at the University of Utah. The study examined whether applying Krashen's input and affective filter hypotheses to language teaching was feasible, productive, and appealing. Attitudinal information about the program was obtained through journals and activity ratings by the students, on-going conversations with teachers, papers written by the teachers, and a questionnaire to the students, teachers, and administrators. Traditional language tests, and students' self-ratings of their performance, measured learners' listening, speaking, reading, writing, grammar, and vocabulary. Teachers perceived the input-driven approach feasible, while students, over time, felt the need for output practice. The test results were analyzed statistically, revealing that the control group performed better than the experimental group. The evaluators faced many dilemmas between interpretability and practicality, and they had to make compromises in deciding how to test; thus, "due to constraints on time, money and personnel, tests had to be easy to develop, administer, and score" (p. 152). Test design issues potentially led to apparent discrepancies between the

instructed learners. Such methodological issues led the author to embark on a series of further evaluation studies.

**Pardo-Ballester, C. (2012).** CALL evaluation: Students' perception and use of LoMasTv. *CALICO Journal*, 29, 532–547. <https://doi.org/10.11139/cj.29.3.532-547>

*Keywords: Spanish; FL; college; quantitative; qualitative; CALL; questionnaire; proficiency tests; student perceptions*

This large-scale study examined 539 Spanish L2 university students' perceptions of the effectiveness of the web-based multimedia program LoMásTV for L2 learning. These intermediate-level Spanish L2 learners completed a pre-course background questionnaire and a post-course perception questionnaire (open-ended and multiple choice questions) at the end of the semester to measure their level of enjoyment in using the program for L2 learning and the perceived usefulness of the program. Participants' perceptions were measured based on the criteria from Chapelle (2001): language learning potential, learner fit, meaning focus, authenticity, positive impact, and practicality. Participants also completed pre- and post-tests on listening and speaking. The participant self-reported perceptions of improvement and the post-test results both indicated gains in listening and speaking. Recommendations for using web-based video materials in L2 instruction are included in the article.

**Pawan, F., & Thomalla, T. G. (2006).** Making the invisible visible: A responsive evaluation study of ESL and Spanish language services for immigrants in a small rural county in Indiana. *TESOL Quarterly*, 39(4), 683–705. <https://doi.org/10.2307/3588527>

*Keywords: ESL; U.S.; Spanish; community service; immigrant; responsive evaluation; participatory; SWOT analysis; purposive sampling; stakeholder*

Pawan and Thomalla report the implementation and results of a responsive evaluation of ESL and Spanish language services, initiated by the County Alliance for Community Education (CACE) in a rural county in Indiana. The impetus for the evaluation study was a prediction that the number of immigrants and immigrant workers is likely to grow, due to the influx of immigrants in neighboring counties and a dairy company's plan to locate its facility in the county. The purposive sample of participants included sponsors (the board members and staffs of CACE), community leaders, service providers, and language service clients. Initial interviews were conducted with 12 stakeholders to elicit concerns and issues, and to set up standards for evaluation. The standards, then, served as a guide to elicit information from a larger population (63 individuals). Multiple sources of information about the language service providers, such as interview notes (notes were checked by the interviewee for accuracy), activity observations, and



documents and multimedia analysis, were utilized for triangulation. Two meetings with ten representative stakeholders were held to jointly discuss the report in terms of strengths, weaknesses, opportunities, and threats (SWOT). The SWOT analysis enabled stakeholders to engage in the interpretation of findings. Lastly, the authors laid out short-term and long-term recommendations to stakeholders. The responsive evaluation approach involved collaborative decision making between the evaluation specialist and the stakeholders, and it offered insights from multiple perspectives into the existing situation and the complexities of providing language services.

**Patton, M. Q. (2022).** *Utilization-focused evaluation* (5th ed.). Sage publications.

*Keywords: utilization-focused evaluation (UFE), Guiding, Useful, Inspirational, Development, and Evaluate (GUIDE) framework, sustainability, equity*

The book comprises three parts: part 1 (chapter 1 to chapter 4) explains what Utilization-focused evaluation (U-FE) is and introduces essential concepts and processes of U-FE. Part 2 (chapter 5 to chapter 16) details the GUIDE framework and 10 operating principles. Part 3 (chapter 17 to chapter 20) brings in the modern context in which U-FE can be used and the goal of this book. The book claims that evaluations should strive for sustainability, equity, and transformation not only globally, but also locally. The book aims to help students who are interested in evaluation by introducing major concepts and to help professional evaluators by refreshing their memories and showing a new approach to evaluation.

Chapter 1 explains what U-FE is. Chapter 2 shows the types of minimum specifications (henceforth MIN SPECS) essential for an evaluation to be considered utilization-focused. Chapter 3 illustrates the evaluative thought process for U-FE. Chapter 4 explains the importance of accountability in evaluation. Chapter 5 provides the Guiding, Useful, Inspirational, Development, and Evaluate (GUIDE) framework. Chapter 6 to Chapter 16 introduce 10 U-FE operating principles: Contextualize, personalize, customize, align, prioritize, engage, strategize, adapt, mobilize, and reflect and learn. Chapter 17 gives some examples of what hinders the evaluation process: *Posttruth* (the state in which objective facts are less effective than appeals to personal beliefs and emotions in shaping public opinion), and *infodemic* (information + epidemic; the tendency for high-spreading information to combine accurate information with misleading interpretations and allegations). Chapter 18 summarizes the book by making a case that evaluation can contribute to a more equitable and sustainable world.

**Peacock, M. (2009).** The evaluation of foreign-language-teacher education programmes. *Language Teaching Research*, 13(3), 259–278.  
<https://doi.org/10.1177/1362168809104698>

*Keywords: Hong Kong; teacher education; program improvement; university; EFL; qualitative; quantitative; questionnaire; interview; case study; evaluation procedure*

This article describes an evaluation of a foreign-language teacher education program at the City University of Hong Kong. The evaluation aimed to determine the program's strengths and weaknesses, and how well it met the needs of the students. Based on review of the evaluation and foreign language teaching literatures, the evaluator developed his own evaluation procedure. The steps for the procedure were: (a) review the literature and produce a set of questions; (b) establish appropriate sources of data for the setting; (c) choose and design data collection methods and instruments; (d) collect and analyze each set of data against the questions; (e) construct an account by relating each interpretation to the others (p. 262). The evaluation used both qualitative and quantitative data collection methods, including questionnaires, interviews, and document analysis. After discussing the results of the evaluation and recommendations for program improvements, the author reflects on the strengths and weaknesses of the procedure, and makes suggestions for improving the evaluation in the future.

**Pellettieri, J. (2011).** Measuring language-related outcomes of community-based learning in intermediate Spanish courses. *Hispania*, 94, 285–302.  
<https://doi.org/10.1353/hpn.2011.0041>

*Keywords: Spanish; college; FL; questionnaire; quantitative; qualitative; community service learning; willingness to communicate; motivation*

This article reports on an evaluation of the impact that participation in a community-based Spanish language learning program had on university learners' willingness to communicate (WTC) in Spanish, motivation, and attitudes toward communication in Spanish both inside and outside the classroom. Most of the 18 participants served as English tutors for Spanish-speaking students throughout an academic year and completed mini-research projects as well. Pre- and post-program data were collected using a Likert-scale questionnaire that measured Spanish WTC, self-ratings of Spanish oral communication ability, L2 communication anxiety, motivation, attitudes of integrativeness, and frequency of Spanish use. The post-program questionnaire also included open-ended questions gauging students' perceptions and opinions of the experience. Results show significant gains in WTC in Spanish, increases in motivation and integrativeness, and stronger associations between Spanish learning goals and WTC. The questionnaire is included in the article appendix.

**Pennington, M. C. (1991).** *Building better English language programs: Perspectives on evaluation in ESL.* NAFSA.

*Keywords: administrator; faculty evaluation; class observation; student service; self-study*

The book is a collection of articles that discuss approaches to English as a second language (ESL) program evaluation (system construction, self-study), evaluation of curriculum and content (participatory placement and cultural aspects of ESL programs), non-instructional aspects (student services, database construction), and administrative aspects (administrators and teachers). The book is particularly broad in its coverage of diverse elements that constitute language programs and enable them to function. The chapters include: “Developing effective evaluation systems for language programs” by Brown and Pennington (chapter 1); “Self-study and self-regulation for ESL programs: Issues arising from the associational approach” by Byrd and Constantinides (chapter 2); “A novel approach to ESL program evaluation” by Eskey, Lacy, and Kraft (chapter 3); “Unifying curriculum process and curriculum outcomes: The key to excellence in language education” by Pennington and Brown (chapter 4); “Participatory placement: A case study” by Spaventa and Williamson (chapter 5); “Evaluation of culture components in ESL programs” by Winskowski-Jackson (chapter 6); “Evaluation of student services in ESL programs” by Middlebrook (chapter 7); “Creating and operating a statistical database for evaluation in an English language program” by Ponder and Powell (chapter 8); “Designing and assessing the efficacy of ESL promotional materials” by Jenks (chapter 9); “Procedures and instruments for faculty evaluation ESL” by Pennington and Young (chapter 10); “Evaluating the ESL program director” by Fox (chapter 11); “Administrative evaluation in ESL programs: How’m I doin’?” by Matthies (chapter 12). Chapters include useful appendices, such as checklists for evaluating cultural components of a program, classroom observation sheets, sample C-tests for placement purpose, and faculty evaluation instruments.

**Pennington, M. C., & Brown, J. D. (1991).** Unifying curriculum process and curriculum outcomes: The key to excellence in language education. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 57–74). NAFSA.

*Keywords: model; Curriculum Process Model; Curriculum Outcomes Model; quality control; needs analysis; objectives; testing; materials; teaching; consistency; efficiency; effectiveness*

Curriculum development is a cyclical process of interrelated activities, including needs analysis, objectives setting, testing, materials, teaching, and evaluation. In the Curriculum Process Model (Brown, 1989), evaluation is a “process devoted to continually improving each component of a program on the basis of what is known about all other components separately as well as collectively” (p. 65). In addition to the Curriculum Process Model, Pennington and Brown add another dimension, the Curriculum Outcomes Model, which takes a quality control approach to

ensuring excellence in language programs. A language program will achieve targeted outcomes if consistency, efficiency, and effectiveness are unified at all programmatic stages (as covered in the Curriculum Process Model). By “developing a more unified vision of the curriculum and greater cooperation among members of a language program” (p. 71), the purpose of evaluation will also be clarified. Evaluative outcomes, in turn, contribute to a more unified understanding of the program.

**Pennington, M. C., & Young, A. L. (1991).** Procedures and instruments for faculty evaluation in ESL. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 191–227). NAFSA.

*Keywords: ESL; faculty development; faculty evaluation*

Pennington and Young address the evaluation of faculty as a way to reinforce program quality. They reflect on the use of different kinds of assessment at different stages of the teachers’ career. Pennington and Young claim there are two kinds of instruments that can be utilized, depending on the purpose: fluid instruments (conversations, letters, essay questionnaire) and fixed instruments (fixed response questionnaire, rating scales, tests, and various descriptive data). Both types have advantages and disadvantages to fully respond to the purpose of evaluation, multiple instruments from multiple resources should be used. They suggest four steps for a performance evaluation interview: (a) substantiate performance; (b) reach understanding of (teaching) job requirements and responsibilities; (c) gain acknowledgement of the issues discussed in the interview; and (d) set goals, action steps, and a time-table for professional development towards the goals. Teachers are an essential part of any educational program, and faculty evaluation can offer substantial information for further development. Appendices include samples of an essay questionnaire, rating scales, a teacher observation form, a form for self-evaluation of a lesson, a student evaluation form, faculty standards of performance, categories for evaluation of research/teaching/service, and a format for annual teacher performance review.

**Plakans, L., & Burke, M. (2013).** The decision-making process in language program placement: Test and nontest factors interacting in context. *Language Assessment Quarterly*, 10(2), 115–134. <https://doi.org/10.1080/15434303.2011.627598>

*Keywords: placement tests; English as a second language tests; higher education; English as a second language instruction*

Plakans and Burke describe a study of decision making during test use in the context of a university intensive English language program. Over a period of 21/2 years, data were collected by audio-recording placement sessions in which the program director and an instructor made

decisions based on a three-part placement exam, TOEFL scores, course grades, instructor evaluations, and other information. Analysis of the sessions revealed four major areas impacting test use and the decision-making process: (a) test performance and score factors, (b) student factors, (c) test user factors, and (d) program factors (i.e., number and size of levels, curriculum, textbooks). These four areas represent factors in the placement process that are clearly about tests (test performance and score) and factors that have less relation to tests or scores such as test users and programmatic factors. The test and nontest factors interacted in the placement process to navigate borderlines between levels, confirm placements, and adhere to programmatic constraints.

**Plass J. L. (1998).** Design and evaluation of the user interface of foreign language multimedia software: A cognitive approach. *Language Learning & Technology*, 2(1), 35–45.  
<http://dx.doi.org/10125/25031>

*Keywords: CALL; model; evaluation criteria; user interface design*

Plass reviews various models and approaches for evaluating user interface design of foreign language multimedia software. He defines interface design as “the process of selecting interface elements and features based on their ability to deliver support for the cognitive processes involved in the instructional activities facilitated by the application” (p. 45). Among many approaches, he proposes an integrated cognitive and pragmatic approach to user interface design and evaluation that is theory-driven, pragmatic, and domain-specific (i.e., contextualized). According to Plass, setting domain-specific evaluation criteria involves the following four steps: “(1) Identify relevant skills, competencies, and domain knowledge. (2) Identify activities that cultivate and develop these skills, competencies, and knowledge. (3) Identify the cognitive processes involved in these activities. (4) Assess the level of support for these cognitive processes provided by the application and its user interface” (p. 48).

**Ponder, R., & Powell, B. (1991).** Creating and operating a statistical database for evaluation in an English language program. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 155–171). NAFSA.

*Keywords: ESL; data collection; database*

Evaluation involves systematic data collection over extended periods of time. The purpose and utility of the database motivate the types of variables included: to track individuals, to inform language learning theory and pedagogical decisions, and to make administrative and business decisions. The authors list four typical narrative cases of problems that frequently arise in language program evaluation (e.g., placement) and provide solutions by using an example

statistical database. Ponder and Powell suggest that it will be wise to conceptualize what variables one will be collecting and decide what kind of record keeping (management) system, database format, and analysis will be used, before the implementation of evaluation.

**Prabhu, N. S. (1987).** *Second language pedagogy*. Oxford University Press.

*Keywords: Communicational Teaching Project, CTP, Bangalore, India, implementation, innovation, sense of plausibility, meaning-focused activity, EFL, task, pre-task, marking, task cycle, TBLT, Structural-Oral-Situational, S-O-S, syllabus, procedural, reasoning-gap, information-gap, opinion-gap, eclecticism*

In chapter 1, Prabhu introduces the Communicational Teaching Project, or CTP—an exploratory teaching project that took place in primary and secondary schools in southern India over periods of time ranging from one to three years. The CTP was exploratory in that it did not seek to ‘test’ a particular approach or methodology of English language teaching (ELT); its purpose was to develop teaching procedures through both the teaching process and in ongoing, regular professional debate about ELT in India. He then talks about how the impetus behind the CTP was an intuition that grammatical competence was developed automatically through the creation of conditions for meaning-focused learning. Next, Prabhu talks about generalizability in relation to the CTP, stating that its insights may not be relevant for the implementation of procedures all over India nor irrelevant for other contexts. Lastly, English speakers, the ‘status’ and perceptions of English, the role of English in certain spheres, and resources for teaching in India are discussed.

Chapter 2 is broken up into three sections. The background section discusses early trends and perceptions of ELT in India, generally, and in the early phase of the CTP, specifically. Prabhu discusses the transition from earlier grammar-based approaches to Structural-Oral-Situational (S-O-S) teaching procedures, perceived shortcomings with S-O-S, mixed reactions to the later introduction of notional-functional syllabuses (Wilkins, 1972) and discourse views on language (Widdowson, 1978), and teachers’ exposure to other approaches to ELT being implemented overseas. All of this led to a re-examination of the S-O-S pedagogy and the desire for new teaching procedures that focused on meaning and avoided sequencing. In the principles and procedures section, Prabhu discusses the CTP in its early and later years. Early years were marked by uncertainty with new teaching procedures, negative responses from learners, and conflicting perceptions. Over time, the CTP team devised a set of tasks; a three-part task cycle, including pre-task, task, and marking phases; an understanding of how language was used throughout lessons and phases; and a conceptualization and clarification of the way students dealt with language and meaning, leading to four different language categories and the understanding of task as a meaning-focused activity. Later years saw the application of tasks to subsequent years, the use of tasks with beginning learners, the creation of a bank of tasks,

participation in review seminars to re-examine pedagogic assumptions and perceptions, and an evaluation by Beretta and Davies (1984). In the last section, Prabhu illustrates a few tasks, including railway timetables, instructions to draw, interpreting rules, and beginners' tasks.

In chapter 3, Prabhu discusses how teaching evolved throughout the course of the CTP project. In the reason-gap section, Prabhu lays out three activity types that featured in project teaching: reasoning-gap, information-gap, and opinion-gap activities. In the pre-task and task section, Prabhu discusses how the pre-task and task phases benefited the project and project participants. In the teacher's language section, Prabhu discusses how teachers tailored their vocabulary decisions and comprehension questions to help students in task completion. In the learners' language section, Prabhu notes how learners used various resources and strategies for conveying meaning; he also distinguishes between students' productive, borrowing, and reproductive English uses. In the final section, incident correction, Prabhu highlights the way teachers corrected students' mistakes or errors during the project. He discusses error correction in terms of systematic (public, explanatory correction activities) and incidental (token-based, responsive, facilitative, and transitory) correction.

In chapter 4, Prabhu discusses the concepts of language learning behind the CTP's teaching procedures. Learners' preoccupation with tasks was seen as a way to develop linguistic competence through sustained engagement with meaning-focused activity. In doing meaning-focused activity, learners handled meaning-content, which involved both conscious attention to meaning and unconscious attention to linguistic structures. Prabhu refers to system-development as a process of the interplay between deployment and acquisition, wherein attempts to comprehend result in the deployment of abstract structures, and the act of deployment further develops those structures. As for rule-focused activity, Prabhu reminds us that the structure or state of an internal system is beyond description in a language grammar, and the order in which structures are presented in a grammar do not mimic developmental trajectories. Planned progression refers to the assumption in using grammars, either for implicit or explicit purposes, that language is learned discretely and in a particular order. Pre-selection is a result of planned progression, which results in form-focused activity and the provision of a teacher with both language samples and opportunities to use those samples. Prabhu contrasts meaning-focused activity with meaningful practice; the former deals with the treatment of meaning only and language use as the need arises, whereas meaningful practice refers to a combination of meaning and form. Language awareness refers to instances in which learners pay attention to form but do so incidentally in dealing with meaning. In the comprehension and production section, Prabhu compares deployment in comprehension on four factors: private versus public displays, partial versus incomplete states, explicitness and commitment, and the use of learner versus interlocutor use or reliance on extra-linguistic resources. In the groupwork section, Prabhu notes that the CTP project did not use groupwork. This decision stemmed from a desire to adhere to the original principle of sustained work with meaning-focused activity but

also came to reflect the idea that learners need to encounter ‘superior data’ (e.g., from the teacher) and may be comfortable losing face with the teacher but not peers.

In chapter 5, Prabhu discusses implications for task-based syllabuses and pedagogical materials. In the first section, Prabhu notes how the CTP’s syllabus was seen as a way to share teaching experience. Prabhu then discusses the generality of the CTP syllabus and how it could be modified along the following parameters to vary task complexity: information provided, reasoning needed, precision needed, familiarity with constraints, degree of abstractness, moving from information- to reasoning-gap tasks, moving from oral to written tasks, and varying the amount and complexity of language used in presenting tasks. Next, Prabhu refers to an illuminative construct as one that sheds light on what is to be learned. The syllabus can be used as a tool of organizational control; it provides supervisory control in institutions and can be a foundation for common examinations. The syllabus as a document of public consent section has to do with the function of a syllabus for making explicit teaching intentions to the public. In the simple and sophisticated syllabuses section, Prabhu states that syllabuses should be as general or specific in so far as they support classroom practice. In the materials section, Prabhu notes that materials in task-based teaching are resources for teachers and not just textbooks. In the coverage section, Prabhu discusses how task-based teaching as a means for promoting long-term learner growth is not useful for gauging how much language has been covered. In the teaching aids section, Prabhu notes the limited resources available for the CTP project, but cautions against correlating teaching resources with educational quality. In the final section, Prabhu discusses students’ learning from non-native English speaking (NNES) teachers and the status of English as an international language.

In chapter 6, Prabhu discusses his own view of the CTP project and pedagogic innovation. He questions the statutory implementation of teaching methods on the grounds that teaching methods crucially hinge on teachers’ pedagogic perceptions. In the sense of plausibility section, Prabhu states that there is a web of factors influencing what teachers do in the classroom or the method to which they ascribe. This web is formed through the routines in which teachers and learners interact with one another and perform their various roles, as well as teachers’ perceptions of how teaching practices result in learning outcomes. Sense of plausibility can be un-engaged, resulting in the routinization of classroom practices, or engaged, in which a teacher’s sense of plausibility (the perceived plausibility of methods or practices) is open to change and may result in professional growth. In the impact of innovations section, Prabhu talks about new perceptions of pedagogy in relation to statutory implementations. Prabhu then discusses a host of factors that may lead to the adoption or rejection of new perceptions or innovations. In the language teaching specialism section, Prabhu discusses how specialism (e.g., applied linguistics) is a combination of identifying, developing, and articulating perceptions as well as uncovering ways to engage these perceptions in the teaching community, which may in turn affect perceptions and teaching procedures. In the eclecticism section, Prabhu discusses how eclecticism is needed and outlines four concepts associated with eclecticism. The chapter



concludes with a few comments on how innovation may refresh teaching procedures that have become overly routinized and may renew teachers' sense of plausibility.

**Rarick, D. (2009).** Resuscitating university language programs in the global age: The international engineering program at the university of Rhode Island. *ADFL Bulletin*, 41, 113–124. <https://doi.org/10.1632/adfl.41.1.113>

*Keywords: college; French; German; Spanish; FL; professional degree; cross-disciplinary; qualitative; case study; survey*

Many undergraduate professional degrees have no foreign language course or credit requirements. This case study describes the University of Rhode Island's collaborative degree established between an international engineering program (IEP) and a university foreign language and literature department. This five-year IEP gives students a BA in French, German, or Spanish and a BS in an engineering discipline, thus bolstering enrollments in foreign language courses. The article describes the development, funding, program design, recruitment, post-program job placement, and other related outcomes of the program. Results from a post-program survey of 221 graduates are briefly reported.

**Rea-Dickens, P., & Germaine, K. P. (1998).** The price of everything and value of nothing: Trends in language program evaluation. In P. Rea-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 3–19). Longman.

*Keywords: UK; Europe; overview; trends; methods; participatory evaluation*

Rea-Dickens and Germaine illustrate the growth of interest in program evaluation in the 1990s, evident from the increase of publications on evaluation (including macro- and micro- evaluation studies), the emergence of an active professional evaluation community, and the establishment of various external accreditation organizations (especially in the UK). With the expansion of evaluation functions (accountability, developmental, awareness-raising, and management), encouragement to use a variety of triangulated data elicitation methods, and engagement of various stakeholders in the process of evaluation, program evaluation has become much more dynamic than early traditions that focused on predetermined measurable outcomes from an empiricist paradigm. Evaluation is argued here to be information/knowledge generation for short-term immediate use and for policy shaping, thereby building bridges across domains and stakeholders to “promote professional development and validity” (p. 16).

**Rea-Dickins, P. (2001).** Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing*, 18(4), 429–462.  
<https://doi.org/10.1177/026553220101800407>

*Keywords: EAL; English; assessment; elementary; process; classroom observation; feedback; teachers*

Rea-Dickins presents a working framework of processes (planning, implementation, monitoring, and recording/dissemination stages) and strategies in classroom assessment decisions. She then applies the framework to examine the classroom assessment practice in an elementary-level English as an Additional Language (EAL) classroom. The evaluator observed, video- and audio-recorded, took field notes, and transcribed three assessments of classroom interaction (formal assessment, informal whole class assessment, informal small group work assessment). The evaluation was repeated in three school settings for one week each over three school terms. The evaluator also conducted semi-structured teacher interviews with two language support teachers, and one mainstream class teacher, before and after administering assessments. In addition, two learners in each of four classes were tracked in detail to reflect on students' assessment experiences. Analyses revealed three purposes of the assessment: (a) bureaucratic (providing information for external agency), (b) pedagogic (making instructional decisions based on learners' achievements), and (c) learning (developing learner awareness, understanding, and knowledge). Though in-depth evaluation of formative assessment practices may be difficult for teachers to practice on a daily basis, its cyclical use can raise awareness about the types and the roles of formative assessment.

**Rea-Dickins, P., & Germaine, K. (1992).** *Evaluation*. Oxford University Press.

*Keywords: teacher training; curriculum development; accountability; method; purpose; procedures; framework; participatory; principles*

Rea-Dickins and Germaine view evaluation as “the means by which both teaching and learning may function more efficiently and quality be assured” (p. xii). The perspective of evaluation for accountability, for curriculum development and innovation, and for professional development is consistent throughout the book. Section one (“Explanation”) provides principles of educational evaluation (innovation, management, and context), exploring the evaluation purpose, design, and framework. Section two is a collection of short summaries of 15 case studies with attention to the context, aim (purpose), design, and procedures. The examples range from evaluation of a project, an intensive ESL program, secondary schools, treatment of oral errors, materials, teachers, learner outcomes (process and product), to syllabus evaluation. The last section (“Exploring evaluation potential”) is devoted to the application of the previously mentioned frameworks, methodology, and other aspects of evaluation through tasks. Rea-Dickins and Germaine emphasize the involvement of teachers and stakeholders throughout the evaluation process

parallel to curriculum development. Rather than a theoretical argument, this book serves as a guide for teachers to clarify the principles and carry out evaluation in practice. The tasks (125 in total) provide opportunities for the practitioners to reflect and raise awareness to conduct evaluation in their own contexts. Although the book covers many aspects of evaluation process in a limited space, the managerial, political, and personal aspect of evaluation falls short (only mentioned briefly in four pages, section 1.4). One may want to look into Rea-Dickins and Germaine's (1998) other edited book, titled "Managing evaluation and innovation in language teaching: Bridging bridges."

**Rea-Dickins, P., & Germaine, K., P. (Eds.). (1998).** *Managing evaluation and innovation in language teaching: Building bridges*. Longman.

*Keywords: UK; Europe; ESL; EFL; innovation; management; implementation; teacher education; ethnography; culture*

The book is a collection of 11 chapters related to innovation and change in English language programs around the world, but primarily in European contexts. After the introductory chapter, an overview of trends in language program evaluation, the chapters are divided into three sections: (a) Evaluating innovation in language education (3 articles), (b) Managing evaluation and innovation (3 articles), and (c) views from the bridge (4 articles). The first two sections are reviewed in this annotation since they address different approaches to program evaluation and provide real-world evaluation examples in a variety of settings. Both sections seek bridges from other disciplines, thereby expanding potential methodologies and approaches in language program evaluation. The annotated book chapters are: "The price of everything and the value of nothing: Trends in language programme evaluation" (Rea-Dickins and Germaine, chapter 1); "Evaluating the implementation of educational innovations: Lessons from the past" (Karavas-Doukas, chapter 2); "Language and cultural issues in innovation: The European dimension" (Roberts, chapter 3); "Programme evaluation by teachers: Issues of policy and practice" (Kieley, chapter 4); "Using institutional self-evaluation to promote the quality of language and communication training programmes" (Mackay, Wellesley, Tasman, & Bazergan, chapter 5); "Managing developmental evaluation activities in teacher education: Empowering teachers in a new mode of learning" (Hedge, chapter 6); and "Managing and evaluating change: The case of teacher appraisal" (Anderson, chapter 7).

**Ricardo-Osorio, J. G. (2008).** A study of foreign language learning outcomes assessment in U.S. undergraduate education. *Foreign Language Annals*, 41(4), 590–610.  
<https://doi.org/10.1111/j.1944-9720.2008.tb03319.x>

*Keywords: US; university; survey; quantitative; performance-based; outcomes assessment; ACTFL; oral proficiency interview; foreign language*

This article reports on a survey of student learning outcomes assessment in university foreign language programs in the US. The study investigated which performance-based assessments are commonly used, how frequently the ACTFL guidelines and National Standards are used, and which obstacles impede the use of performance-based assessments. A Likert-style, web-based questionnaire was developed and quantitative data analysis was performed. The results indicated that faculty-designed multiple choice tests were the most common assessment method, followed by student papers and projects. Translation was more common than the oral proficiency interview or portfolios. ACTFL guidelines were often used for developing speaking assessments, but rarely for other purposes. Lack of time and lack of faculty knowledge were given as the main obstacles to using performance-based assessment. The article also provides a thorough literature review of the recent history and common types of performance-based assessment.

**Richards, J. (2001).** Approaches to evaluation. In J. Richards (Ed.), *Curriculum development in language teaching* (pp. 286–309). Cambridge University Press.

*Keywords: theoretical; formative; summative; illuminative; evaluation questions; method; development; accountability; stakeholder identification*

Richards situates evaluation as one of the key elements at stake throughout the curriculum development process, functioning as a “reflective analysis of the practices” (p. 286). This chapter briefly covers the three types of evaluation purposes (formative evaluation for ongoing development and improvement; illuminative evaluation for deeper understanding of the program; and summative evaluation for seeking program effectiveness). It then moves to issues in program evaluation (identification and involvement of the stakeholders, the use of quantitative and qualitative measurements, documentation of process information, and adequacy of the evaluation plan and implementation), and advantages and disadvantages of methodologies that can be used for data gathering. The evaluative questions Richards lists as examples for formative, illuminative, and summative evaluation are primarily focused on “what has happened” rather than “what we shall do from now.” Appendices include two examples of program evaluation (EFL courses in primary schools and language courses in a private language institute) with a focus on the audiences, methodology, and reporting of the evaluation.

**Rifkin, B. (2003).** Oral proficiency learning outcomes and curricular design. *Foreign Language Annals*, 36(4), 582–588. <https://doi.org/10.1111/j.1944-9720.2003.tb02148.x>

*Keywords: ACTFL proficiency guidelines; university; advanced proficiency*

Rifkin argues that the ACTFL Oral Proficiency Guidelines and associated assessments have not facilitated college foreign language (FL) instruction and curricula to advance students' oral proficiency towards stated outcomes. The argument derives from his premise that the Proficiency Guidelines should guide an overall curricular framework. Reviewing several studies on the amount of instruction and exposure necessary for FL oral proficiency development, he explains that the inability for college students to attain advanced levels of proficiency can be attributed to the limited time on task (i.e., instructional hours in language classrooms). Instructional suggestions for moving student learning towards advanced-level oral proficiency are provided.

For a response to Rifkin (2003) by Glisan and Donato (2004) and Rifkin's (2004) rebuttal to Glisan and Donato (2004), see the following references:

Glisan, E. W., & Donato, R. (2004). It's not "Just a matter of time on time": A response to Rifkin. *Foreign Language Annals*, 37 (3), 470-476.

Rifkin, B. (2004). A response to Glisan and Donato. *Foreign Language Annals*, 37 (3), 477-483. <https://doi.org/10.1111/j.1944-9720.2004.tb02705.x>

**Roberts, C. (1998).** Language and cultural issues in innovation: The European dimension. In P. Rea-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 51–77). Longman.

*Keywords: UK; Europe; university; overview; trends; methods; participatory evaluation; culture; ethnography*

Roberts argues for ethnographic methodologies in evaluating foreign language and culture learning in modern language degree programs at the tertiary level in the UK. In particular, she focuses on study abroad in the final year of the degree, the so-called 'Language Learners as Ethnographers' project.' The following tools were utilized within this ethnographic framework: interviews with students and lecturers, course diaries, end of course questionnaires, products of the ethnographic project (a written report), meetings with students abroad, classroom observations, joint assessment meetings on the projects and drafts of the project, staff-student discussions, observation of project supervision (field notes were taken during observation and meetings). She illustrates the value of eliciting rich and thick description of the program using an ethnographic approach. She also argues that, via ethnography, evaluation becomes a context-bound process of understanding (p. 75) for educational purposes rather than a set of facts

which can straightforwardly predict and replicate other successful projects (pp. 75-76) for accountability purposes.

**Rogan, F., & San Miguel, C. (2013).** Improving clinical communication of students with English as a second language using online technology: A small scale evaluation study. *Nurse Education in Practice*, 13(5), 400–406. <https://doi.org/10.1016/j.nepr.2012.12.003>

*Keywords: English as a second language instruction; computer assisted language learning; professional education; health care practitioners; English for special purposes*

Rogan and San Miguel describe and evaluate an innovation to assist ESL nursing students at an Australian university develop their clinical communication skills and practice readiness by providing online learning resources, using podcast and vodcast technology, that blend with classroom activities and facilitate flexible and independent learning. The innovation builds on an intensive clinical language workshop program called ‘Clinically Speaking’ which has evolved through a cyclical process of ongoing research to produce resources in response to students’ learning needs. Whilst uptake of the resources was modest, students of ESL as well as English speaking backgrounds (ESB) found the resources improved their clinical preparation and confidence by increasing their understanding of expectations, clinical language and communication skills. The innovation, developed with a modest budget, shows potential in developing ESL and ESB students’ readiness for clinical communication, enabling them to engage in clinical practice to develop competency standards required of nursing graduates and registration authorities.

**Ross, S. (1992).** Program-defining evaluation in a decade of eclecticism. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 167–195). Cambridge University Press.

*Keywords: Japan; university; EFL; quantitative; audio-lingual; functional-notional; grammar-based; self-access; task-based; materials; classroom observation; testing; checklist; teacher observer*

Ross demonstrates a program-formative evaluation approach applied at a Japanese junior college program, which he characterizes as a laissez-faire English Language Teaching environment. The goal of the study was to generalize the findings to other Japanese second language teaching contexts. He examined how the observed differences in methodological characteristics of five different teaching approaches (audio-lingual, functional-notional, grammar-based, self-access pair learning, task-based) as determined by the materials used in the English as a foreign language courses related to the product/outcome differences. The project utilized teachers as

participant observers to reduce the teachers' anxieties toward outsider observations. The researcher created a low-inference coding scheme to analyze classroom activity types (student activities, sources of input to students, student behavior, and the distribution of classroom time). Four observations were done by four different teachers. Each observed activities and behaviors in the four sections of the coding scheme. Their observations were tallied and summed for cluster analysis. Specific hypotheses were then created based on the observations and compared with the outcome measures (grammar test, listening cloze test, partial dictation test, a narrative discourse test, and a structured oral interview test) using analysis of covariance (the pre-test score, self-report of extra-curricular contact with native speakers, attendance rate as covariates). The link between process and product data was found for listening input and the development of listening skills in the post-test, but not for grammar input and pair-work. The quantitative data obtained through observation could only partially reveal methodological features of the instructional setting; thus, Ross notes that more affective and linguistic aspects of language learning are needed.

**Ross, S. (2003).** A diachronic coherence model for language programme evaluation. *Language Learning*, 53(1), 1–33. <https://doi.org/10.1111/1467-9922.00209>

*Keywords:* Japan; student mastery; achievement; proficiency; learning outcomes; testing; model

Ross warns that the use of norm-referenced testing can lead to incorrect inferences about program success and students' mastery. However, the use of syllabus-based assessments has been considered insufficient in terms of "hard evidence of generalized proficiency gains" (p. 6). This conflict has not been resolved and "no single approach has been able to assess achievement and proficiency simultaneously" (p. 7). The study analyzes the relationship between program-internal assessment (composite grades of portfolio self, peer, and teacher assessments, as well as syllabus content testing) and program-external assessment, from six cohorts of 1,820 undergraduates in an EAP/EFL program. Standardized proficiency testing (TOEFL) was administered at the beginning and at the end of the year, and achievement testing was undertaken twice per year (approximately every 80 classroom hours). The path analysis was undertaken to reveal the direct and indirect link between achievement and proficiency. The results revealed that listening skills developed independent of classroom instruction, while academic literacy appeared to be more program dependent, requiring greater learner effort. The result in the second instructional year showed an overall weak link between the pre- and post-reading proficiency measures. Also, the program-internal achievement tests had no direct impact on the post-reading proficiency test, but the note taking course did have an impact. These findings led to the reformation of the reading curriculum and better coordination of assessment criteria. The Diachronic Coherence Model, Ross proposes, reveals the strength of the relationships among the internal achievement tests, which are based on the learning outcomes, and between proficiency and achievement tests. This

model can respond to both external accountability and internal formative purposes. However, it is limited to a program evaluation based on student outcomes assessment.

**Ross, S. J. (2009).** Program evaluation. In M. Long & C. Doughty (Eds.), *Handbook of second and foreign language teaching* (pp. 756–778). Blackwell.

*Keywords: stakeholder involvement; logic model; program theory; evaluability; implementation evaluation; outcomes assessment; quantitative; program cohesion; value-added design*

The book chapter elucidates issues and recent developments in program evaluation. Ross explains that many evaluation studies find stakeholder involvement in evaluation process key to ensure relevancy of evaluation to local programs. In order to situate evaluation planning in local contexts, Ross considers it important to understand stakeholders' differential views and their needs in terms of programs and their evaluation. In order to establish a unified view of an often dynamic and complex program prior to evaluation, logic models can be developed to reveal initial assumptions about program inputs, processes, outputs, and outcomes. Ross also points out the necessity to monitor and evaluate program implementation. By tracking unforeseen factors that influence program delivery and theory, program improvements can be made as a program is being implemented. For outcomes assessment, Ross introduces efficacy and effectiveness evaluation and associated quantitative methodologies (e.g., propensity score analysis, regression discontinuity design). Other program evaluation topics introduced in the chapter include evaluability analysis, program cohesion analysis, and value-added evaluation design.

**Sanders, R. F. (2005).** Redesigning introductory Spanish: Increased enrollment, online management, cost reduction, and effects on student learning. *Foreign Language Annals*, 38(4), 523–532. <https://doi.org/10.1111/j.1944-9720.2005.tb02519.x>

*Keywords: USA; Spanish; university; outcomes assessment; computer-mediated instruction; OPI; WPT; WebCape; innovation*

Sanders describes a student learning outcomes (SLO) assessment study of a redesigned first-year Spanish program at Portland State University. The Spanish program implemented computer-mediated instruction featuring automated online exercises and synchronous discussion in order to respond to high enrollment demands, reduce program cost per student, and reduce student seat time and the number of students per class. The SLO assessment was conducted to examine whether students in the newly designed course could achieve proficiency outcomes comparable to students in the traditional face-to-face class (i.e., pre-innovation course). At the end of three quarters, students took the Brigham Young University WebCape tests (vocabulary, grammar, and reading tests), an ACTFL Oral Proficiency Interview (OPI), and an ACTFL



Writing Proficiency Test (WPT). Comparison between student performance prior to innovation and after two years of implementing the technology-enhanced program showed equal achievement levels in WebCape and OPI, but a decline in writing proficiency in the innovation group. Potential factors affecting the differences, like learner background and teacher experience, are discussed. Though not all proficiency outcomes were maintained, the program change solved immediate enrollment issues and reduced program cost per student.

**Schmitt, E. (2014).** Seat time vs. proficiency: Assessment of language development in undergraduate students. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 110–130). Cengage Learning.

*Keywords: longitudinal study; proficiency-driven liberal education program; proficiency requirement*

In a three-year longitudinal study, Elena Schmitt, the chair of the Department of World Languages and Literatures at Southern Connecticut State University, describes the implications of the transition from a seat-time requirement to a proficiency requirement for FL learning. Her chapter describes students' performance on the Standards-Based Measurement of Proficiency (STAMP) test as well as the influence of course scheduling on student learning following a university-wide movement toward skill-based and proficiency-driven education.

**Schneider, A. I. (2000).** Title VI funding for undergraduate international study programs: Long-term impact on language offerings. *ADFL Bulletin*, 32(1), 42–47.

*Keywords: US; university; undergraduate; international study; grant impact; accountability*

Schneider reports on an evaluation study of all U.S. Department of Education Undergraduate International Studies and Foreign Language Programs. The evaluation occurred in response to federal accountability requirements. The focus here is on the language instruction component of the overall evaluation study. A questionnaire was distributed to the 107 funded projects (75% return rate), and site visits were made to 51 of the respondents (64%), in order to find out about the impact of funding. Results indicated not only an overall strong and long-lasting impact on curriculum development and the campus environment, but also an indirect impact on student participation through the establishment of, or an increase in overseas study. It was also found that the programs lacked data management and needed help with systematic collection and analysis of grant impact (e.g., information on enrollment, revised/added courses, methods of instruction).

**Schulz, R. A. (2007).** The challenge of assessing cultural understanding in the context of foreign language instruction. *Foreign Language Annals*, 40(1), 9–26.  
<https://doi.org/10.1111/j.1944-9720.2007.tb02851.x>

*Keywords: intercultural competence; assessment; student learning outcomes; portfolio; German; university*

Schulz reviews the literature on intercultural competency outcomes and problematizes the inconsistencies of past operationalizations of intercultural learning and teaching in foreign language education. Based on her reviews, she proposes five fundamental objectives for cross-cultural awareness and understanding for pre-collegiate and college introductory level foreign language programs. The five objectives focus on the development of students' awareness of (a) factors that "impact...cultural perspectives, products, and practices" (p.16), (b) situational factors influencing interaction and behavior, (c) stereotypical views of the home and target culture, (d) culture-specific images and connotations of expressions, and (e) possible causative factors for cultural misunderstandings. To assess process- and product-oriented intercultural learning outcomes, she suggests using portfolio assessment that can be integrated into an existing course. The Appendix includes concrete instructions, tasks, and assessment criteria that are aligned with Schulz's five fundamental cultural objectives for introductory college-level German courses.

**Shaver, A. (2012).** Fostering integrative motivation among introductory-level German students through a language partners program. *Die Unterrichtspraxis/Teaching German*, 45, 67–73. <https://doi.org/10.1111/j.1756-1221.2012.00119.x>

*Keywords: German; FL; college; conversation partners; survey; qualitative*

This article describes a small Language Partners program used in a university German department, designed to improve German L2 students' motivation and oral communication skills. Program design, implementation, and evaluation are described. The program is evaluated with an annual survey of student participants and German Language Partners at the end of the academic session. The survey is qualitative and asks for the participants' perceptions on their experiences, the logistics, impact on L2 learning, and suggestions for improvement. After four years of data collection, findings indicate that the program enhances integrative motivation, learners' L2 confidence, and willingness to communicate.

**Seçer, Ş. Y. E., & Erişen, Y. (2020).** Evaluation of English language teaching undergraduate programs in terms of document analysis. *International Journal of Curriculum and Instruction, 12*(2), 671–703.

*Keywords: English as a foreign language (EFL) teacher education, program evaluation, document analysis, Turkey*

This qualitative study evaluates English Language Teaching (ELT) undergraduate programs in six Turkish universities through the method of document analysis. Based on the documents gathered from open-access official websites of programs, it comparatively analyzed the strengths and weaknesses of ELT undergraduate programs from the following four categories: students, academicians, practicum and educational opportunities. First, the authors evaluated the category of students by the number of students enrolled in programs and the perception level of students in terms of university entrance examination scores. The findings showed less-crowded classrooms and higher students' qualifications are conducive to quality education. Second, the category "academicians" examined the number and qualifications of academicians. The authors looked into the number of teaching staff in all positions, their educational background, working experiences, tenure terms and scientific publications. The importance of instruction and instructors in the higher-education quality was emphasized, which indicated that more academicians with higher qualifications is a strength for the ELT program. Third, the category "practicum" was analyzed. The study summarized three factors affecting the implementation of practicum practices stipulated by the Ministry of National Education, including the student-supervisor ratio, additional teaching practices provided by private universities, and divergence of supervisors' guidance, all of which are integral to the program quality. Finally, the article addressed the category "educational opportunities" by comparing the number of seminars, conferences, workshops and other social activities that provide professional development for students.

The findings revealed that public and private universities have their own merits and drawbacks with reference to four categories. Based on the findings, the authors suggested that an earlier start of practicum courses benefit ELT programs for both public and private universities.

**Shintani, N. (2011).** A comparative study of the effects of input-based and production-based instruction on vocabulary acquisition by young EFL learners. *Language Teaching Research, 15*(2), 137–158. <https://doi.org/10.1177/1362168810388692>

*Keywords: input, comprehension, production, vocabulary, acquisition, young learners, instructional processes, quasi-experimental, Japanese, EFL, listen-and-do, tasks, task-based, discrete-point, interaction, tokens*

In this study, Shintani investigated the effects of input-based and production-based instruction on vocabulary acquisition. Shintani wanted to know if participation in input-based as well as production-based tasks enabled young L2 learners to acquire new vocabulary, which instructional approach led to more newly acquired words, and whether or not instructional processes in the interactions of either differed. Shintani used a quasi-experimental, pre-test-post-test design. Participants in this study were 36 Japanese learners of English, aged 6-8, who were divided into one of three conditions: input-based, production-based, and control. The treatment consisted of six lessons over a six-week period (two lessons per week), and each lesson was 30-minutes long. Three listen-and-do tasks were conducted with the input-based group, and five activities were conducted with the production-based group. Shintani used Japanese to help explain tasks and activities (when needed) but otherwise spoke in English. Multiple assessments were used to gauge learners' target-vocabulary knowledge: two comprehension-based and two-production based measures, using both task-based and discrete-point items. Participants were administered a pre-test, a post-test, and a delayed-post test. All lessons were audio- and video-recorded. Analyses consisted of between and within-subject comparisons and, to analyze instructional processes, transcriptions of classroom interactions. Shintani found that there were similar effects for input- and production-based instructional conditions on vocabulary acquisition, that input-based tasks provided more opportunities for rich interaction than the production-based activities, and that each group acquired similar numbers of vocabulary tokens.

**Shintani, N. (2013).** The effect of focus on form and focus on forms instruction on the acquisition of productive knowledge of L2 vocabulary by young beginning-level learners. *TESOL Quarterly*, 47(1), 36–62. <https://doi.org/10.1002/tesq.54>

*Keywords: focus on forms, FonFs, focus on form, FonF, Japanese, EFL, PPP, nouns, adjectives, vocabulary, task-based teaching, TBT, discrete-point, audio-recording, transcription, tokens, input, negotiation of meaning, production, outcomes*

Shintani investigates the effects of focus on forms (FonFs) and focus on form (FonF) on complete-beginner L1 Japanese learners' acquisition of English nouns and adjectives. FonF was operationalized as a present-practice-produce (PPP) methodology while FonFs was operationalized as task-based teaching (TBT). Shintani wanted to know to what extent differences between FonFs and FonF were reflected in the process features of each instruction type; the effect that FonFs and FonF had on young, beginning-level learners' productive knowledge of nouns and adjectives; and whether or not there was a difference between the two. Participants were 45 Japanese children, aged six years-old, with no previous second language learning experience. Participants were divided into three groups—FonFs, FonF, and control—each further divided into two classes of six to nine participants. There were three different 30-minute lessons for each group, repeated nine times over five weeks. The study

included 24 nouns and 12 adjectives. In the FonFs group, nouns and adjectives were introduced linearly (following PPP) in three sets of 12 words (eight nouns and four adjectives). Words in the FonF group were introduced incidentally and in different ways. All 24 nouns appeared in every lesson but adjectives arose spontaneously. There were five activities in each FonFs lessons, which were repeated nine times. There were three listen-and-do tasks in FonF lessons. For both groups, activities were in English but procedures in Japanese. Productive knowledge was measured with discrete-point and task-based tests and involved pre-, post-, and delayed-post tests. Lessons were audio- and video-recorded and then transcribed. Transcriptions were analyzed for tokens in the input and output, and test scores were analyzed with descriptive statistics and for within- and between-group differences using PASW (/SPSS). Shintani found that both FonFs and FonF were effective for noun acquisition, but only FonF was effective for the adjective acquisition. Additionally, only FonF learners were able to use adjectives in free production. Contextualized input, negotiation of meaning, and student-initiated production were identified as the features of FonF instruction that led to differences in learning outcomes.

**Shneyderman, A., & Abella, R. (2009).** The effects of the extended foreign language programs on Spanish-language proficiency and academic achievement in English. *Bilingual Research Journal: The Journal of the National Association for Bilingual Education*, 32, 241–259. <https://doi.org/10.1080/15235880903372829>

*Keywords: Spanish; English; bilingual; two-way immersion; primary school; language assessment; achievement testing; quantitative*

This large-scale, longitudinal study compares two different program models for two-way immersion bilingual Spanish instruction. Under the “Extended Foreign Language” program approach, both program models provide instruction in both English and Spanish. However, in one program model students receive one hour of L2 language arts instruction, while in the other program model students receive one hour of L2 language arts instruction plus 30 minutes of science or social studies content instruction. Spanish language learning gains were measured using the reading comprehension section of the Aprenda assessment and the Language Assessment Scales-Oral (LAS-O). Academic achievement gains were measured in reading and math using the Florida Comprehensive Assessment Test. Comparison data included a control group of 418 students not enrolled in Extended Foreign Language programs and an intervention group of 418 students enrolled in one of the two Extended Foreign Language program models. Results of statistical analyses showed that greater gains are made in Spanish reading comprehension when students received the additional 30 minutes of content instruction.

**Slimani, A. (1992).** Evaluation of classroom interaction. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 197–221). Cambridge University Press.

*Keywords: conversation analysis; uptake; checklist; learner perspective*

Silimani investigates whether topicalization (learning opportunities) of English linguistic items in classroom interaction leads to ‘uptake’ by the learners. Here, ‘uptake’ was understood as what individual learners claim to have learned from the interactive classroom events which have just preceded (p. 199). Silimani operationalized ‘uptake’ by distributing an Uptake Recall Chart at the end of the observed lesson, and also an Uptake Identification Probe, which occurred three-hours later. Student interviews were also conducted but did not produce responses that were sufficiently precise to be interpreted in relation to what might account for their claims (p. 205). Findings indicated that learners’ perception of topicalization and uptake were more salient when initiated by the learner, suggesting also that learners’ perceptions are highly idiosyncratic. The study highlights the potential value of including the learners’ perspective regarding what they learned in classroom interaction. Tapping their perspectives through learning-focused measures (like the uptake charts) may provide an important supplement for the interpretation of learning processes and outcomes in evaluation studies.

**Snow, M. A., & Brinton, D. M. (1988).** Content-based language instruction: Investigating the effectiveness of the adjunct model. *TESOL Quarterly*, 22(4), 553–574.  
<https://doi.org/10.2307/3587256>

*Keywords: US; ESP; university; adjunct model; questionnaire; domain expert; content-based*

Snow and Brinton examine the effectiveness of the adjunct model (non-students concurrently taking a general education course and a sheltered language course, both linked by content) in the 7-week Freshman Summer Program (FSP) at UCLA. The language course was developed based on a needs analysis of the content discipline (instructors’ feedback, analysis of the language and content materials, review of previous curricula and assignments, and input from other specialists). It was also adjusted throughout the instructional period by holding weekly staff meetings. 79 out of 224 former FSP students responded to a retrospective survey which requested demographic information, global benefits of FSP courses, and the usefulness of specific activities and skills students learned in FSP. Open-ended questions revealed three types of positive attitudes towards FSP (The ease of adjustment, self-confidence, and learning to get help) and constructive comments (follow-up support is needed after FSP, and the program focuses less on natural science but more on social science and humanities). A follow-up study featured interviews with the new graduates of FSP, reflecting on the beneficial effects of FSP in equipping students for the academic demands they face. The authors also compared FSP students and non-FSP students in their performances on the placement exam and an end-of-semester simulated academic task. Despite the FSP students’ lower and broader distribution on the

placement test, they were able to perform similarly to ESL students in the simulated final examination. However, the incompatibility of the initial and final tests and the difference in timing of the placement test make the conclusion difficult to interpret. Although sampling may be difficult, the study shows how former students can be one important source on program effectiveness.

**Spaventa, L., J., & Williamson, J., S. (1991).** Participatory placement: A case study. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 75–97). NAFSA.

*Keywords: ESL; university; process-oriented; placement test; testing*

Spaventa and Williamson illustrate the process of how a new placement test and associated procedures were created for a 10-week ESL program. The program had encountered several problematic issues in placing students, including: (a) no reflection of students' oral competence by the existing test (Michigan English Language Placement Test, MELPT); (b) a "lack of standardization of evaluative criteria" (p. 80) for oral testing; (c) excessive time and energy for administering and scoring of tests; and (d) lack of teacher involvement in student placement. To address the issues of scorability, economy, and administrability, a C-test (a text with deletion of one or more letters after the first letter for every other word), an oral placement (a combination of students' self assessment and teacher-student discussion), and a 10-minute writing test were developed for placement decisions. Where needed, the MELPT, which correlated relatively highly with the C-test, was administered for informing potential placement level switches. Spaventa and Williamson summarize their process model of participatory testing, which includes students' and teachers' voice in the placement decision making process.

**Su, Y. (2011).** The effects of the cultural portfolio project on cultural and EFL learning in Taiwan's EFL college classes. *Language Teaching Research*, 15(2), 230–252.  
<https://doi.org/10.1177/1362168810388721>

*Keywords: Taiwan; English; FL; college; questionnaires; observation; interview; portfolio; qualitative; culture*

This article describes the implementation and evaluation of a semester-long cultural portfolio project completed by 38 undergraduate English as a foreign language students in Taiwan. The project was designed to increase students' knowledge of English-speaking cultures, identify how a cultural portfolio can change students' stereotypes and misconceptions of other cultures, and measure changes in cultural knowledge. Students completed open-ended pre- and post-project questionnaires about experiences and beliefs related to English-speaking cultures and the impact

of the portfolio project on their cultural knowledge. Classroom observations, student interviews, journal reflections, written project reports, and oral project reports were collected and analyzed, using the qualitative method of constant comparison, to identify major themes in students' cultural learning, perceptions, and change. Results suggested the strong, positive impact that the portfolio project had on students' cultural awareness, on strategies for constructing and internalizing cultural knowledge, and on the role of active engagement in confronting cultural stereotypes. Instruments are included in the appendices.

**Sullivan, J. H. (2006).** The importance of program evaluation in collegiate foreign language programs. *Modern Language Journal*, 90(4), 590–593. [https://doi.org/10.1111/j.1540-4781.2006.00466\\_6.x](https://doi.org/10.1111/j.1540-4781.2006.00466_6.x)

*Keywords: assessment; teacher certification; model; NCATE; university*

As a member of the National Council for the Accreditation of Teacher Education (NCATE) Board of Examiners, Sullivan stresses the importance of “collaboration and collegiality” (p. 592) in conducting and sustaining effective program evaluation. As outlined in the NCATE/ACTFL guidelines for accreditation, programs are expected to create a professional learning community and implement a locally contextualized program evaluation. Sullivan introduces and exemplifies a template for the NCATE/ACTFL six-step approach to evidence gathering on teacher-candidates' performance. Sullivan also stresses the autonomy and willingness of faculty to take control of departmental self-study as keys to supporting claims about the value of faculty members' educational efforts.

**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). Heinle Cengage.

*Keywords: German; FL; college; course evaluation; teacher evaluation; quantitative; questionnaire*

This book chapter compares the effectiveness of two different forms of student L2 course evaluations. A large-scale study was conducted with course evaluation data from 80 German courses provided by 1,424 students regarding the German L2 instruction provided by 30 graduate student teaching assistants (TAs), attempting to analyze the connection between students' perceptions of a course and their attitudes regarding the instructor's effectiveness. The chapter begins with a discussion of the validity and multidimensionality of student evaluations of teacher effectiveness (SETEs), then describes two versions of a SETE used in the German program. The



original SETE used Likert-scale items to measure the students' overall perceptions of the course and teacher effectiveness, which led to some conflation of course effectiveness with teacher effectiveness. The follow-up survey explicitly separated out items on course effectiveness from teacher effectiveness. Results from both surveys showed that students generally rated the teachers higher than the course, but there was a stronger correlation between teacher and course effectiveness in the original survey than in the follow-up survey. The authors found that separating out the items on the course from the teacher in the follow-up survey allowed students to evaluate more of the aspects of the teacher quality without conflating their attitudes about the course itself. Recommendations for course evaluations include having specific questions about different course elements, avoiding global evaluations of teacher effectiveness, and supplementing SETEs with other types of evaluation information. The follow-up SETE survey is included in the appendix.

**Towell, R., & Tomlinson, P. (1999).** Language curriculum development research at university level. *Language Teaching Research*, 3(1), 1–32.  
<https://doi.org/10.1177/136216889900300102>

*Keywords: curriculum development, French, United Kingdom, second language acquisition, communicative language teaching, text typology, methodology, interpersonal skills, task, curriculum design, framework, implementation, evaluation, survey questionnaires, diaries, innovation, outcomes*

annotation: In this article, Towell and Tomlinson report on a largely formative ten-year curriculum-development project that took place in the French section of the Department of Modern Languages at the University of Salford in the United Kingdom from 1988-1998. The authors first outline a theoretical framework for the development of a model to be used in task-based curriculum design. Using this model, a task-based curriculum was implemented and evaluated during this ten-year period in two distinct phases. In the first phase, which consisted of several units, the task of one unit involved students working in teams to plan an election campaign. The two outcomes for the task included a typed report as well as an oral justification and presentation of ideas put forward. Teams were assessed by a panel of staff on five weighted questions. Evaluation of the first phase (1989-1990) included student diaries, a survey questionnaire, and an examination similar to the above assessment format. Data from student survey questionnaires and diaries in the first phase helped the authors identify areas that could be improved in the curriculum. These improvements were then incorporated into the second phase of the task-based curriculum. There were seven units covered, each unit now shorter (in response to student feedback) than units in phase one, with each unit consisting of its own task and various intermediate objectives (sometimes with objectives for each class hour), outcomes, and built-in text types. At the end of the second phase, the authors administered survey questionnaires to students to help them understand to what extent the course appeared to meet course objectives

and students' perceived linguistic ability (to comprehend, speak, and write). The authors use this ten-year experience to chronicle the possibility of designing and implementing a task-based curriculum in keeping with their predetermined theoretical orientations and to highlight the role of evaluation in supporting innovation and ongoing improvement.

**Tribble, C. (2000).** Designing evaluation into educational change processes. *ELT Journal*, 54 (4), 319–327. <https://doi.org/10.1093/elt/54.4.319>

*Keywords: EFL; Central Europe; China; project evaluation; exit exam system; teacher training; insider/stakeholder involvement; program development; baseline evaluation; logic model*

Tribble argues the value of an insider-informed baseline evaluation study in order to ensure that program innovation is appropriately contextualized. The author showcases two innovation projects, one with a baseline evaluation study and one without. The first case, Year 12 Project, took place in Central Europe and aimed at creating a new English examination system for school leavers. The insider project teams conducted a baseline evaluation study in addition to the one done by the British Council consultants. The insider-managed baseline evaluation benefited the local teams in many ways, for example, they were able to gain familiarity with research methods and have realistic understanding of the project. The second innovation project was a project to train English teacher trainers for middle schools in China. The project was planned on a false assumption that trained trainers are in position to mentor the local teachers. The baseline evaluation could have redirected the project planning to overcome the local issues, if time and resources were put in for such an evaluation study. Learning from the two cases, insider-informed baseline evaluation can contribute to sustain project benefits, clarify the political climate of the project, and ensure that educational change processes are embedded in local contexts.

**Tsou, W., & Chen, F. (2014).** ESP program evaluation framework: Description and application to a Taiwanese university ESP program. *English for Specific Purposes*, 33, 39–53. <https://doi.org/10.1016/j.esp.2013.07.008>

*Keywords: ESP; program evaluation; higher education; authenticity; learner autonomy; learning transfer*

Tsou and Chen describe a combined, updated model for ESP program evaluation and report on how the model worked when it was applied to evaluate an ESP program in a university in Taiwan. To update the model for higher education ESP program evaluation, they combined Hutchinson and Waters' 1987 model and the comprehensive framework for foreign language program evaluation developed by Watanabe, Norris, and Gonzalez-Lloret (2009), while also

incorporating recent findings from emerging research on ESP learning and teaching that include topics such as authenticity, learner autonomy and learning transfer. When the model was applied to evaluate a university ESP program, the findings enabled the researchers to identify strengths and weaknesses of the updated model. Tsou and Chen discuss both the updated model and the evaluation process which provides valuable insights into the design and implementation of ESP program evaluation.

**Tucker, G. R., & Cziko, G. A. (1978).** The role of evaluation in bilingual education. In J. E. Alatis (Ed.), *International dimensions of bilingual education* (pp. 111–124). Georgetown University Press.

*Keywords: bilingual education; Canada; Nigeria; Philippines; experimental*

The authors highlight three bilingual education (BE) programs in Canada, Nigeria, and the Philippines as examples of evaluation applied to BE. Many bilingual education programs adopted an experimental, comparative approach for evaluation, which may be susceptible to problems with random assignment, teacher effect, and/or uncontrolled/unmeasured factors. In addition, many programs had not articulated locally agreed consensual goals for education in terms of “affective, cognitive, linguistic or social development,” which affected the choice of evaluation strategy (p. 432). Instead of making judgmental decisions about the programs, the authors suggest that evaluators “evaluate the relative strengths and weaknesses of a variety of program alternatives and to specify the conditions under which each might be more or less successful” for knowledge formulation purposes (p. 433). Conducting a formal evaluation will lead program stakeholders to collaboratively specify, operationalize, and implement program goals and objectives. Thus, evaluation can also be seen as knowledge formation for formative and summative purposes. The authors introduce the notion of contextually tailored testing, which can be adapted to different contexts and updated as teaching and learning situations evolve.

**Uzun, L. (2016).** Evaluation of the latest English language teacher training programme in Turkey: Teacher trainees' perspective. *Cogent Education*, 3(1).  
<https://doi.org/10.1080/2331186X.2016.1147115>

*Keywords: English language teacher training program, teacher trainees, program evaluation, Turkey*

The study evaluates an English Language Teacher Training Program (ELTTP) at one university in Turkey. The purpose was to reveal the extent to which the program satisfied the needs of the students and to discover the degree to which the trainees benefited from the professional training. The major questions posed were (a) how were the ELTTP courses regarded, (b) to what extent

did the program help students develop, and (c) were the program's components regarded as effective for improving student development.

Data was collected from 90 students using questionnaires that asked whether there was an improvement in the following three components: personal development, professional development, and theoretical and practical knowledge. Additionally, interviews were implemented to gather in-depth opinions from the students. Findings showed that the program did not provide enough knowledge and skills that satisfy teacher trainees' needs. In general, students seemed to be more satisfied with the instructors rather than the contents of the course themselves. The majority of students noted that there were not ample opportunities for them to practice what they have learned inside the classroom. This result was consistent across the gender and proficiency levels of the students. Based on the findings, the study suggested including more courses that can enhance learners' creativity and knowledge through actual practice.

**Van den Branden, K. (2006b).** Training teachers: Task-based as well? In K. Van den Branden (Ed.), *Task-based language teaching in practice* (pp. 217–273). Cambridge University Press.

*Keywords: teacher training, perceptions, TBLT, task-based language teaching, Flanders, Belgium, cognition, in-service, implementation, Brussels, Dutch as a second language, support*

Van den Branden discusses the relationship between teacher training and the introduction and implementation of task-based language teaching (TBLT) in Flanders. The report is divided into six sections. In the first, Van den Branden discusses the relationship between teacher cognitions—that is, beliefs, knowledge, and thoughts—and their actions in the classroom, which may be inconsistent due to contextual constraints, conflicting beliefs, or disconnect between beliefs and skills. Van den Branden remarks that research has not concluded whether the failure of teacher training to impact practice has been a result of the program itself or its particular features. In section two, Van den Branden underscores the need to ‘teach what you preach.’ In the third section, Van den Branden reviews several studies that report on the impact of in-service teacher training programs in Flanders: an early study by Linsen (1994) and later studies by Timmermans (2005), Luyten and Houben (2002), and D’hont (2004). Linsen identified a number of reported differences between TBLT and teachers’ other approaches as well as four problems teachers in EPP schools encountered: task complexity, control, learner differences, and group work. Later studies by Timmermans, Luyten and Houben, and D’hont corroborate these findings and note ways that teachers would adapt or simplify tasks to accommodate their own teaching and learning views. Van den Branden concludes this section with a number of positive reactions to task-based syllabuses, too. In the fourth section, Van den Branden reports on a longitudinal study (by Devlieger et al., 2003) into the implementation of TBLT in Dutch-medium primary

schools in Brussels. In section five, Van den Branden discusses the importance of contextual factors on implementation (and teachers' perceptions), outlining four sub-categories of contextual conditions, and underscoring the importance of head teachers in implementation activities. In the last section, Van den Branden reports on two evaluations of task-based implementation: one by Hillewaere (2000), who looked at the extent to which educational quality improved as a consequence of the Flemish government's support, and another by Devlieger and Goossens (2004). Hillewaere, using observations, interviews with teachers and head teachers in 20 schools, and scores on language proficiency and arithmetic tests, concluded that language education showed strongest effects in implementing educational innovations. Devlieger and Goossens compared teachers' actions at the start of implementation with their actions at the end of the three-year implementation program. They found the following changes: stronger orientation towards functional language goals, higher quality of the teacher's input, higher level of involvement and motivation among pupils, lack of EPP support transfer to rest of curriculum, a more functional arrangement of classroom, a need for guided time to adapt to new syllabus, lack of effect on methodological formats, need for control, and effects of needs-based coaching. Van den Branden concludes by emphasizing the importance of sustained (teacher) support during task-based implementation and by referring to the larger impacts of TBLT on learner language performance and participation in education and society.

**Van Viegen, S., & Russell, B. (2019).** More than language—Evaluating a Canadian university EAP bridging program. *TESL Canada Journal*, 36(1), 97–120.  
<https://doi.org/10.18806/tesl.v36i1.1304>

*Keywords: English for Academic Purposes (EAP), international students, students' needs, bridging program, Canada, university*

The study examines the English for Academic Purpose Bridging Program (EBP) designed to support international students at a Canadian university. The focus of the program is on developing students' communication competence which is needed in the academic context. The purpose of the evaluation was to bridge the gap between the program policy and implementation, by examining (a) the relationship between students' performance in the bridging program and their academic performance in their subsequent degree programs, and (b) the linguistic and academic competencies that are most needed for the successful transition to the degree programs. Quantitative data was collected from five cohorts of students in the liberal arts stream and qualitative data was collected through open-ended interviews with broader university administrators and focus groups with EBP faculty, past and current students ( $n=25$ ), and tutorial assistants ( $n=13$ ). Based on the participatory design, researchers informed stakeholders (e.g., administrators, faculty, and program leaders) of students' needs and the way to meet them.

The findings revealed that the bridging program played a considerable role in facilitating learners' performance in academic courses. Additionally, the study found that the language proficiency assessments were only one of the factors that impact students' future academic performance, indicating the limitation of standardized language proficiency as a predictor of students' success in subsequent degree programs. Interview data from students and focus groups suggested that preparing students for a successful transition to the degree programs is not merely teaching them language skills, but also involves a dynamic process of academic discourse socialization.

In general, the study suggested that the program help learners build both linguistic and academic skills, thereby suiting both general and discipline-specific needs. The authors concluded by broaching the need for a disciplinary literacy approach that can move the program from a general to a more discipline-specific academic program.

**Wang, P. (2013).** Can automated writing evaluation programs help students improve their English writing? *International Journal Of Applied Linguistics & English Literature*, 2(1), 6–11. <https://doi.org/10.7575/ijalel.v.2n.1p.6>

*Keywords: automated writing evaluation, criterion, writing*

This study explores the effect of the automated writing evaluation (AWE) on Taiwanese students writing, and whether student improvement and their perception of the program are related. Instruments included a questionnaire, 735 essays analyzed in Criterion, and a pre/post essay. Two classes of 53 college students participated in the study. Descriptive statistics, paired-samples t-tests, Pearson correlation, effect size, and regression were used to analyze the data. Results showed that students improved significantly in terms of the length of the essay and the scores awarded by the machine and the human raters. However, among the five essays, the first essay is the only one showing a significant level of consistency between student improvement and student attitude, and the correlation declined dramatically after the first essay. This study may be of importance in confirming the usefulness of the AWE functions such as recursive revising and instant scoring, as well as in providing teachers with a better understanding of how student beliefs about the Criterion program might relate to their writing performance.

**Warford, M. K. (2006).** Assessing target cultural literacy: The Buffalo state experience. *ADFL Bulletin*, 37(2–3), 47–57.

*Keywords: USA; university; foreign languages; intercultural competence; outcomes assessment; teaching culture*

Warford discusses challenges in assessing cultural competency in foreign language (FL) programs and details a cultural literacy assessment initiative undertaken by the Modern and Classical Languages (MCL) Department at Buffalo State College. According to the author, two main issues of cultural competency assessment exist in the field of FL education. First, defining and assessing cultural competency learning have been subsidiary to language proficiency assessment and are often elusive targets at best (“difficult to measure objectively,” p. 49). Second, a “positivistic” approach to language teaching and testing has limited cultural competency assessment to discrete-point paper-and-pencil tests. At Buffalo State College, the MCL Department engaged in a year-long departmental discussion on cultural competency outcomes, level progressions, and assessment rubrics. The discussion was based on available literature on cultural competency assessment as well as syllabi from across its curriculum. In order to state the outcomes and create a rubric, the department reviewed the National Standards for Foreign Language Learning published in 1996, the American Association for Teachers of French framework of cultural competence, and Seelye’s (1993) five-stage cultural competency attainment approach. The associated assessment focused on cognitive and sociolinguistic outcomes within the B.A. program. Samples of final exam essays and term papers from upper-level courses underwent double-blind rating by two trained faculty members, and findings indicated that no curricular change was needed. Of interest are the processes of rubric development and continual adjustment during and after collecting assessment data. Appendixes include target cultural literacy assessment rubrics for entry-level, intermediate, and major/concentration programs.

**Weir, C., & Roberts, J. (1994).** *Evaluation in ELT*. Blackwell.

*Keywords: overview; theoretical; project evaluation, program evaluation, accountability-oriented, development-oriented, formative, summative, external evaluator, internal evaluator, ODA, case study, method, self-report, observation, political dimension, personal dimension; methods; English*

Weir and Roberts provide a good overview of the development of theory and practice in language program evaluation, based primarily on their experiences as internal and external (Overseas Development Administration, ODA) evaluators. Both accountability- and development-oriented dimensions are discussed throughout the book. The book consists of four parts and includes extensive appendices.

Part I (chapter 1 and 2) provides a comprehensive overview of approaches for evaluating second language projects and programs. In chapter 1, the authors review the accountability- and developmental-orientations to formative and summative evaluation. They suggest an integrated approach to these two supposedly conflicting purposes of program evaluation, viewing “evaluation as contributing to understanding and thereby to general professional accountability

and development as well as satisfying any contractual accountability requirements” (p. 10). During the planning stage, the purpose (why), the focus (what), the people (who), the duration (how long), timing (when), and the method (how) of evaluation all have to be aligned. The authors also review two sets of standards for educational evaluation by the Joint Committee (1981) and Harlen and Elliott (1982). Chapter 2 focuses on the issue of base-line evaluation design (data gathering at the appraisal stage to determine appropriateness, viability, and sustainability of the project, and at the implementation stage to later determine the effectiveness of the project based on the pre-determined outcomes), with a particular focus on project evaluation (contractual accountability-driven projects funded by ODA).

Part II is a collection of three case studies. Authors reflect on their experiences as external (chapter 3 and 5) evaluators and as insider evaluation recipients (chapter 4). The first case (chapter 3) is an evaluation of a four-week ODA-funded EFL in-service teacher training project in Nepal, following a summative and external paradigm with a ‘cost-benefit’ approach. The purpose of the evaluation was to examine whether teacher training, which was supposed to affect pedagogical practice, made a difference in students’ performance in the School Leaving Certificate English examination. Chapter 4 (the second case study) describes a formative evaluation of a twelve-week pre-sessional EAP program for “understanding, action and improvement” (p. 84). The authors, as insiders carrying out inspection of the program, experienced the tension and differing priorities between developmental and accountability-oriented perspectives to evaluation. They address the need for a synthesized approach, in the form of a utilization-focused evaluation. The third case study (chapter 5) is an evaluation of a two-year ODA-funded professional development program for secondary school teachers at a tertiary institution in Latin America. This is a case of an external evaluator playing a critical role in helping insiders to set up feedback and evaluation structures to inform self-evaluation study. A framework for integrating internal and external perspectives for carrying out development-oriented evaluation is proposed.

Part III (chapter 6-7) explores various discrete evaluation methods, including self-report (interviews and questionnaires) and classroom observation. The authors discuss advantages and limitations of each method, step-by-step guides, and examples. Rather than adhering to one paradigm of enquiry, they caution would-be evaluators to determine methodology by first clarifying: (a) the purpose of evaluation; (b) what information is required for fulfilling the purpose; (c) the availability and access to resources; and (d) “the characteristics of informants” (p. 132).

Part IV, the final chapter, discusses the political and personal dimensions of evaluation, including issues of power, decision making, fairness, ownership, and climate. The extensive Appendices are useful for understanding criteria and instruments used in the three case studies.



**Wenzhong, Z., & Cheng, Z. (2013).** An Empirical Study on the Curriculum Construction of Business English for International Trade Based on the Case of GDUFS. *English Language Teaching*, 6(4), 19–28. <https://doi.org/10.5539/elt.v6n4p19>

*Keywords: needs analysis, business English, curriculum construction*

Through review and reflection into the findings of the practice of business English in Guangdong University of Foreign Studies (GDUFS) and other instructions, Wenzhong and Cheng aimed at analyzing the situation of the present curriculum, identifying existing problems such as the imbalance between the English courses and Business courses based on the theory of Needs Analysis, and then make some suggestions to increase the percentage of the Business courses and Business practices. In this way, this paper can serve as a reference for the curriculum construction of Business English for International trade in GDUFS.

**Wherritt, I., & Cleary, T. A. (1990).** A national survey of Spanish language testing for placement or outcome assessment at B.A.-granting institutions in the United States. *Foreign Language Annals*, 23(2), 157–165.

*Keywords: Spanish; university; outcomes assessment; placement test; test use; survey*

The nation-wide survey study was conducted as part of the University of Iowa's Foreign Language Assessment Project, which aimed to develop a placement test to improve high school-university articulation and outcomes assessment instruments for various curricular purposes (e.g., exit requirement, teacher certification). The survey investigated seven major areas: (a) special first-year course offerings (b) test use purposes, (c) types of tests used for placement, (d) skill areas often tested on placement tests, (e) incentives and penalties for freshmen to follow placement decisions, (f) instructional activities, and (g) class size. For the survey, Spanish was chosen as a target language, because many B.A.-granting institutions have a Spanish language program and because Spanish programs often face placement issues. A mail-in survey was sent to 223 systematically sampled college Spanish departments listed in the Modern Language Association directory, and also to 90 additional Spanish programs primarily from large research institutions known for their placement systems. A total of 126 respondents returned the survey (79 out of the 223 and 47 out of the 90). Seventy-two percent of the respondents reported that they have a placement test, but great variability existed in terms of the tests used, ranging from in-house tests to the College Board Achievement Test. For locally-constructed placement tests, discrete-point grammar and reading comprehension were the top two skill areas often tested. Approximately half of the respondents reported rewarding incoming students for college credits when they are placed above the entrance requirement. The authors suggest that placement "policies must include a simultaneous penalty and credit system in order to alleviate the problem of false beginners" (p. 161).

**Williams, E. (2007).** Extensive reading in Malawi: Inadequate implementation or inappropriate innovation? *Journal of Research in Reading*, 30(1), 59–79.  
<https://doi.org/10.1111/j.1467-9817.2006.00328.x>

*Keywords: Malawi; English elementary school; extensive reading; summative; testing; observation; interview*

The author, who was an external evaluator, reports on a summative evaluation of an extensive reading (ER) program for fourth- and fifth-year primary school students implemented across all public schools in Malawi. This program was funded by the UK Department for International Development and was implemented by the Malawian Ministry of Education. In order to assess students' English reading outcomes in the ER program, a cloze test was chosen as the main information source (note that, in addition, informal reading assessment was conducted). The author adopted a time lapse design comparing sixth-year students in the second year of the project with no ER experience (control group) and sixth-year students in the fourth year of the project who went through the ER program for two years (experimental group). The findings from the cloze test were not favorable to the ER program; however, the author cautions that the findings needed to be interpreted with a grain of salt, since the ER program was not implemented as intended. From the program implementation data gathered via classroom observations, interviews, and feedback from teachers and students, there was great variability in ER implementation. Although speculative, the author outlines potential extrinsic factors (e.g., low teacher morale, change in student profiles due to growth of private schools) and intrinsic factors (e.g., insufficient teacher training) that may explain the results. Besides logistical challenges in conducting evaluation, this evaluation case study raises various issues in conducting summative evaluation when program innovation is not appropriately adapted to local educational contexts.

**Windham, S. (2008).** Redesigning lower-level curricula for learning outcomes: A case study. *ADFL Bulletin*, 39(2&3), 31–35. <https://doi.org/10.1632/adfl.39.2.31>

*Keywords: US; university; German; outcomes; assessment; ACTFL; proficiency; standards; program improvement; case study*

This case study describes the modifications Elon University's German department made to its curriculum in order to develop a program with more clearly defined and consistent outcomes. The program changes were motivated by an external review which revealed that, although individual courses had specific objectives, the program lacked articulation across and between levels. The program used the ACTFL proficiency guidelines as a basis for the development of learning outcomes for its different levels. The adoption of new student learning outcomes also led to the development of new evaluation criteria, assessment tools, course content, and pedagogical practices. The result has been clear improvement in student performance on oral proficiency assessments. The author emphasizes that the ACTFL guidelines provided an overall

framework, but the German department made changes and supplemented the guidelines to meet its local needs. Thus, the ACTFL guidelines were used to inform the discussion on outcomes, but not as the final measure of outcomes.

**Winskowski-Jackson, C. (1991).** Evaluation of culture components in ESL programs. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 98–134). NAFSA.

*Keywords: ESL; culture; acculturation; checklist*

Cultural components are often intertwined with language learning in ESL classes. Winskowski-Jackson provides a variety of checklists (staff, curriculum, and student) for identifying the cultural components in any program. Proposed methods include a survey of the program staff's cultural competencies, identification of the curriculum infrastructure in terms of cultural content, and assessment of students' cultural competence (measuring as well their cognitive knowledge, affective development, and psychomotor ability). Although cultural competencies may vary by individual history, the author specifies several that can be evaluated over critical periods of development (at initial, three month, six month, one year, and two year stages). The evaluation of effectiveness for a cultural training program can be complex since the degree of acculturation may differ individually.

**Wright, B. D. (2006).** Learning languages and the language of learning. *Modern Language Journal*, 90(4), 593–597. [https://doi.org/10.1111/j.1540-4781.2006.00466\\_7.x](https://doi.org/10.1111/j.1540-4781.2006.00466_7.x)

*Keywords: assessment; model; accreditation; university*

Wright reconceptualizes assessment as a four-step inquiry and improvement process that “goes far beyond mere measurement” (p. 594). In step one, faculty define and clarify student learning outcomes and address questions about learning. In particular, Wright warns that learning outcomes must be specific to the program goals and not bound by assessment practices. In step two, faculty gather evidence on student learning. Evidence gathering decisions must respond to questions raised regarding the specific goals of student learning in step one. In step three, faculty analyze and interpret gathered information. Lastly, in step four, faculty use evidence of learning to improve student learning. Wright emphasizes both the use of findings and the commitment to taking action through assessment, unlike traditional input and output assessment models. Accreditation bodies and policy makers now require information on how well students reach the outcomes and not just what they learn. Finally, the author proposes that faculty actively engage in self-improvement and utilize assessment for meaningful purposes rather than passively going through the routine solely for meeting accreditation standards.

**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. <https://doi.org/10.1111/j.1944-9720.2013.12005.x>

*Keywords: Chinese; FL; high school; proficiency assessment; observation; interview; survey; quantitative; qualitative*

The authors examined the effectiveness of an intensive (25 hours/week) STARTALK Chinese Language and Culture Program, with a focus on describing and recommending promising practices for intensive FL instruction. Program components included theme-based classroom instruction using a communicative language teaching approach, computer lab activities, field trips, and community-based internships, all designed to develop linguistic skills as well as cultural awareness. Program effectiveness and participant satisfaction were evaluated quantitatively and qualitatively with (a) pre- and post-test scores from the Standards-based Assessment and Measurement of Proficiency (STAMP) and the Simulated Oral Proficiency Interview (SOPI) in Mandarin; (b) classroom observations; (c) student and teacher interviews; (d) student journals; and (e) post-program surveys. Participants included a total of 93 intermediate-level students and two co-teachers. Posttest scores showed that students improved significantly in their language skills regardless of incoming proficiency level. Qualitative data from interviews and surveys indicate that the co-teaching model, the communicative language teaching approach using a variety of activities, the use of technology, field trips, and internships were perceived as being strong contributors to the program's success. No participant comparison group was included. No specific evaluation framework or approach was used.

**Yamada, M. (2018).** Evaluation of an EFL teacher training program in Japan. *Journal of Applied Social Science*, 12(1), 25–45. <https://www.jstor.org/stable/26370954>

*Keywords: English as a Foreign Language (EFL), teacher training program, program evaluation, teacher trainees, multilingualism, multiculturalism, Japan*

The study investigates an English as a Foreign Language (EFL) teacher training program in Japan to determine (a) how Japanese teacher trainees perceived their English learning, (b) what knowledge and skills the trainees considered important for effective interactions in English, and (c) how the trainees could structure their English teaching within Japanese society. To address these questions, 19 teacher trainees were first interviewed to get insight into potential components of the curricula that needed a new direction. Based on the interviews, surveys were then implemented to identify the teacher trainees' perceptions of issues related to linguistic and cultural diversity in Japan. The interview transcripts were analyzed to identify major themes while the survey data was analyzed using descriptive statistics.

One of the emerging themes from the interviews was the role of local knowledge in fostering multilingual and multicultural attitudes toward Japan's ethnic, racial, and gender minorities. The findings also showed the trainees' mixed responses to Japan's cultural diversity. Their understanding of Japan's diversity was subject to individual experiences through which they built their perspectives. Throughout the interviews, the challenges of teaching intercultural communication skills in EFL situations were also addressed.

Most importantly, the author emphasized the significance of fostering the trainees' language awareness as it can provide them with insights into cultural diversity and enhance their linguistic ability. Based on these findings, the study made recommendations that could move the EFL training program classroom toward multiculturalism and multilingualism.

**Yang, W. (2009).** Evaluation of teacher induction practices in a US university English language program: Towards useful evaluation. *Language Teaching Research*, 13(1), 77–98.  
<https://doi.org/10.1177/1362168808095524>

*Keywords: teacher induction; utilization-focused; useful; program improvement; formative; internal; university; case study; US; EAP*

This case study provides an example of how a utilization-focused approach can help teachers conduct internal evaluations that lead to practical program improvements. Yang describes an internal, formative evaluation of the new teacher induction program at a US university's EAP program. The evaluation was conducted by Yang and a new teacher, both working within the program, in an effort to improve the induction process for new teachers. Following Patton's utilization-focused approach, Yang identified the program administrators as the primary-intended users and worked closely with them to design and implement an evaluation that would meet their needs. She explains the utilization-focused evaluation approach and then describes the evaluation process in detail. She illuminates the preliminary negotiations with the primary intended users that led to development of research questions and evaluation design, reports on findings about teacher induction practices, and then discusses how the findings and recommendations were eventually used for the benefit of teachers and the program.

**Yavuz, A., & Zehir Topkaya, E. (2013).** Teacher educators' evaluation of the English language teaching program: A Turkish case. *Novitas ROYAL (Research on Youth and Language)*, 7(1), 64–83.

*Keywords: Second Language Teachers; Teacher Education; Turkey; TESOL; Student Teachers*

Yavuz and Zehir Topkaya explored the perceptions of teacher educators regarding the changes in the English Language Teacher Education Program introduced by the Turkish Higher Education

Council in 2006. Employing a qualitative design, open-ended questionnaires were administered to 18 lecturers working at five different state universities. The analysis of the data yielded that while teacher educators found some of the changes appropriate, such as the addition of some courses, they raised far more serious concerns with the new program regarding the sequence, content, structure, procedure and removal of courses. In addition, the top-down and centralized program restructuring movement, disregarding the opinions, experiences and the practices of the end users of the program, such as teacher educators, teachers and teacher trainees, was also criticized heavily by the participants.

**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(1), 160–181.

*Keywords: Spanish; USA; university; CALL; blended mode; face-to-face; online; SOPI; achievement test; proficiency test; pre-post; reduced contact time*

This article responds to a question: Can online format reduce classroom contact hours and yet have an equivalent level of outcomes as traditional face-to-face classroom learning? Young examined differences in student learning outcomes between two instructional delivery systems, face-to-face versus blended-learning (online and face-to-face), in Spanish 150 (third-semester) courses. The blended-learning sections met two class hours per week for interactive tasks, while face-to-face sections met three class hours per week. Ten sections of the Spanish 150 course were divided into the two delivery systems. Student performances on the following tests were gathered: a set of three pre-post tests (Spanish placement test, listening and reading sections of the Minnesota Language Proficiency Assessment), midterm and final exams (as outcomes achievement measures), and a Spanish Simulated Oral Proficiency Interview (SOPI). Other data included student background, student perceptions of their learning experience, instructor perceptions of the course format, observed classroom practices, and online student behaviors. The traditional face-to-face group outperformed the blended-learning group in the midterm exam, but a reverse pattern was found for the SOPI test. No difference was found for the proficiency pre-post tests and the final exam. The student and instructor perception data indicated that the blended-learning students were engaged in interactive activities in class and attended the class more prepared than the students in the face-to-face sections. A follow-up study was conducted to examine whether the effectiveness of the blended-learning format was due to instructional hours (two-days per week versus three-days per week).

**Yu-Chih, D. S., & Yun-Pi, Y. (2019).** Evaluating an English elite program in Taiwan using the CIPP model. *Journal of Asia TEFL*, 16(1), 1738–3102.  
<http://dx.doi.org/10.18823/asiatefl.2019.16.1.13.200>

*Keywords: English elite program, CIPP model, blended learning, distance learning, Taiwan*

The study uses the CIPP (Context, Input, Process, and Product) model to evaluate the English Elite Program at Fu Jen Catholic University. The English program used both blended learning (a mix of online and face-to face instruction) and distance learning (a mix of synchronous and asynchronous instruction) and focuses on listening, speaking, reading, and writing.

The research sought to investigate (a) the outcomes of the English Elite Program evaluation using the CIPP Model, and (b) the students' perceptions of taking English language courses via distance and blended modes.

Using the CIPP model, the researchers came up with questions according to each keyword: context, input, process, and product. The participants were 55 graduates of the program, teachers, TAs, drop-out students, and administrative assistants. The study used two questionnaires (*Questionnaire of Program Evaluation* (Kiely and Rea-Dickins, 2005) and *Questionnaire of Program Drop-Out* designed by the researchers), classroom observation (with audio recordings), interviews, email analysis, and average score analysis of the students by section and overall.

Results showed that the program faced challenges such as a high dropout rate, undergraduate students' unfamiliarity with administrative work, funding acquisition for the third year of the program, and unfamiliarity with the online environment.

On the positive side, the course had advantages like an all-English policy, multidimensional courses, strong content material, and effective Professional Dialogue training. Suggestions for improvement included minimizing reliance on undergraduate assistants, updating technology, offering more offline classes, increasing the credit of Professional Dialogue courses, and incorporating stronger English language training courses.

**Zannirato, A. (2014).** Promoting Change in “Two-tiered” departments: An exploratory evaluation of conflict and empowerment among language and literature faculty. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 154–182). Cengage Learning.

*Keywords: exploratory evaluation; national, departmental, and individual levels; two-tiered departments; structural factors*

Alessandro Zannirato explicitly adopts empowerment evaluation in his exploration of the organizational dynamics among language and literature faculty in two-tiered departments. As he explains, empowerment evaluation sets out to give voice to diverse stakeholders who may typically be disenfranchised in existing political or power structures of organizations. His survey research findings suggest that conflict and disempowerment are key factors that influence interactions among those in first-tier and second-tier faculty positions. Zannirato, himself an LPD, concludes by recommending that departments identify structural factors that are conflict generating and disempowering and then collectively implement specific plans of action that reduce inequities among language and literature faculty.

**Zapata, G. (2011).** The effects of community service learning projects on L2 learners' cultural understanding. *Hispania*, 94, 86–102. <https://doi.org/10.1353/hpn.2011.0024>

*Keywords: Spanish; FL; college; community service learning; culture; questionnaire; quantitative; qualitative; project evaluation*

This small-scale study compared the impact of community service learning (CSL) projects and cultural presentations on low-intermediate and high-intermediate Spanish L2 university learners' attitudes and L2 use. The 52 undergraduate student participants completed Likert-scale pre- and post-class questionnaires that asked them to rate their L2 proficiency, use, and perceptions of usefulness, and to evaluate their own work in either the CSL project or the cultural presentation. The post-questionnaire also asked open-ended questions about the knowledge they acquired about the target culture and their own learning through participation. Results indicated that high-intermediate learners gained more awareness of the target culture and more confidence as L2 speakers than the low-intermediate learners, and that the CLS projects had a greater impact on attitudes toward the L2 and culture than did the cultural presentations. Recommendations are given regarding the implementation of CSL projects in L2 programs.

**Zareva, A., & Fomina, A. (2013).** Strategy use of Russian pre-service TEFL university students: Using a strategy inventory for program effectiveness evaluation. *International journal of english studies (IJES)*, 13(1), 69–88. <https://doi.org/10.6018/ijes/2013/1/148711>

*Keywords: language learning strategies; ESL/EFL strategy use; pre-service teachers; TEFL program; SILL*

In their study, Zareva and Fomina identify categories of learning strategies that are mostly used by Russian university students in an English Linguistics Program with a TEFL concentration. The more specific goal of the study is to offer a model of evaluation of the effectiveness of TEFL-oriented programs in terms of the language learning strategies their students use and



recognize as pedagogically applicable to their EFL environment. To this end, two groups of students were compared on their self-reported frequency of strategy use —1st year students (n = 23), who had just entered the program, and 4th year students (n = 38), who were close to graduating from the program and entering the teaching profession. The main instrument used in the study was a version of the Strategy Inventory for Language Learning (SILL), designed by Oxford (1990). Overall, both groups showed high to medium frequency of use of all strategy categories; however, the 4th year students revealed a more finely-grained scale of strategy use priorities. The findings of the study can help curriculum designers and instructors refine the focus of their TEFL-track programs and make informed decisions about emphases and de-emphases in their students' training.

## References

- ADFL. (2009). ADE guide for external reviewers. *ADFL Bulletin*, 40(2-3), 138–143.
- ADFL. (2009). Checklist for self-study for departments of foreign languages and literatures. *ADFL Bulletin*, 40(2-3), 144–153.
- 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(1), 75–93. <https://doi.org/10.1080/13803610500392160>
- Agustina, N. Q., & Mukhtaruddin, F. (2019). The CIPP model-based evaluation on Integrated English Learning (IEL) program at language center. *English Language Teaching Educational Journal*, 2(1), 22–31. <https://doi.org/10.12928/eltej.v2i1.1043>
- Alderson, J. C. (1992). Guidelines for the evaluation of language education. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 274–304). Cambridge University Press.
- Alderson, J. C., & Beretta, A. (Eds.). (1992). *Evaluating second language education*. Cambridge University Press.
- Alderson, J. C., & Scott, M. (1992). Insiders, outsiders and participatory evaluation. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 25-57). Cambridge University Press.
- Allen, L. Q. (2010). The impact of study abroad on the professional lives of world language teachers. *Foreign Language Annals*, 43, 93–104. <https://doi.org/10.1111/j.1944-9720.2010.01062.x>

- Altstaedter, L. L., & Jones, B. (2009). Motivating students' foreign language and culture acquisition through web-based inquiry. *Foreign Language Annals*, 42, 640–657.  
<https://doi.org/10.1111/j.1944-9720.2009.01047.x>
- Anderson, J. (1998). Managing and evaluating change: The case of teacher appraisal. In P. Rea-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 159–186). Longman.
- Arias, C. I., Maturana, L. M., & Restrepo, M. I. (2012). Evaluation in foreign language learning: Towards fair and democratic practices. *Lenguaje*, 40(1), 99–126.
- Arnold, N. (2009). Online extensive reading for advanced foreign language learners: An evaluation study. *Foreign Language Annals*, 42(2), 340–366.  
<https://doi.org/10.1111/j.1944-9720.2009.01024.x>
- Atar, C., Kır-Cullen, E., & Denkci-Akkaş, F. (2020). An evaluation of the English preparatory program at İstanbul Medeniyet University. *Language Teaching and Educational Research (LATER)*, 3(1), 94–115. <https://doi.org/10.35207/later.714457>
- Ataş, U., & Balıkçı, G. (2021). Linking theory and practice in ELT program: An evaluation of materials, testing and practicum courses. *Journal of Narrative and Language Studies*, 9(17), 211–226. <https://nalans.com/index.php/nalans/article/view/345>
- Bachman, L. F. (1989). The development and use of criterion-referenced tests of language ability in language program evaluation. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 242–258). Cambridge University Press.
- 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.

<https://doi.org/10.1111/j.1756-1221.2009.00034.x>

Barr, D., Leakey, J., & Ranchoux, A. (2005). Told like it is! An evaluation of an integrated oral development pilot project. *Language Learning & Technology*, 9(3), 55–78.

<https://www.learntechlib.org/p/74478/>

Barreneche, G. I. (2011). Language learners as teachers: Integrating service-learning and the advanced language course. *Hispania*, 94, 103–120.

<https://doi.org/10.1353/hpn.2011.0028>

Bayram, I., & Canaran, Ö. (2019). Evaluation of an English preparatory program at a Turkish foundation university. *Journal of Language and Linguistic Studies*, 15(1), 48–69.

<https://www.researchgate.net/publication/332136972>

Beretta, A. (1986a). A case for field-experimentation in program evaluation. *Language Learning*, 36(3), 295–309. <https://doi.org/10.1111/j.1467-1770.1986.tb00557.x>

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(1), 144–155. <https://doi.org/10.2307/3586398>

Beretta, A. (1990a). Implementation of the Bangalore Project. *Applied Linguistics*, 11(4), 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 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 University Press.
- Beretta, A., & Davies, A. (1985). Evaluation of the Bangalore Project. *ELT Journal*, 39(2), 121–127. <https://doi.org/10.1093/elt/39.2.121>
- Bernhardt, E. B. (2006). Student learning outcomes as professional development and public relations. *Modern Language Journal*, 90(4), 588-590.  
[https://doi.org/10.1111/j.1540-4781.2006.00466\\_5.x](https://doi.org/10.1111/j.1540-4781.2006.00466_5.x)
- Bernhardt, E. B. (2008). Assessment as a keystone for language and literature programs. *ADFL Bulletin*, 40(1), 14–19.
- Bernhardt, E., & Brillantes, M. (2014). The development, management, and costs of a large-scale foreign language assessment program. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 41–61). Cengage Learning.
- Birckbichler, D. W. (Ed.). (2006). *Evaluating foreign language programs: Content, context, change*. Foreign Language Center, the Ohio State University.
- Brantmeier, C., Vanderplank, R., & Strube, M. (2012). What about me?: Individual self-assessment by skill and level of language instruction. *System*, 40(1), 144–160.  
<https://doi.org/10.1016/j.system.2012.01.003>
- Brindley, G. (1998). Outcomes-based assessment and reporting in language programs: A review of the issues. *Language Testing*, 15(1), 45–85.  
<https://doi.org/10.1177/026553229801500103>

- Brindley, G. (2001). Outcomes-based assessment in practice: Some examples and emerging insights. *Language Testing*, 18(4), 393–407.  
<https://doi.org/10.1177/026553220101800405>
- Brown, A. (2014). Foreign language course grades as pre-requisites and programmatic gatekeepers. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 183–207). Cengage Learning.
- Brown, J. D. (1989). Language program evaluation: A synthesis of existing possibilities. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 222–241). Cambridge University Press.
- Brown, J. D. (1995a). *The elements of language curriculum: A systematic approach to program development*. Heinle & Heinle.
- Brown, J. D. (1995b). Language program evaluation: Decisions, problems and solutions. *Annual Review of Applied Linguistics*, 15, 227–248.
- Brown, J. D. (2008). Testing-context analysis: Assessment is just another part of language curriculum development. *Language Assessment Quarterly*, 5(4), 275–312.  
<https://doi.org/10.1080/15434300802457455>
- Brown, J. D., & Pennington, M. C. (1991). Developing effective evaluation systems for language programs. In M. C. Pennington (Ed.), *Building better English language Programs: Perspectives on evaluation in ESL* (pp. 3–18). NAFSA.
- 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. <https://doi.org/10.1016/j.tate.2007.11.012>

- Burden, R., & Williams, M. (1996). Evaluation as an aid to innovation in foreign language teaching: The “SPARE” wheel model. *Language Learning Journal*, 13, 51–54.
- Burke, B. M. (2012). Experiential professional development: Improving world language pedagogy inside Spanish classrooms. *Hispania*, 95, 714–733.
- Byrd, P., & Constantinides, J., C. (1991). Self-study and self regulation for ESL programs: Issues arising from the associational approach. In M. C. Pennington (Ed.), *Building better English language programs* (pp. 19–35). NAFSA.
- Byrnes, H. (2008). Owning up to ownership of foreign language program outcomes assessment. *ADFL Bulletin*, 39(2&3), 28–30.
- Cachey, T. (2014). Reframing Assessment: Innovation and accountability between the global and the local. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 230–244). Cengage Learning.
- 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. <https://doi.org/10.1111/j.1944-9720.2012.01204.x>
- Cañado, M. L. P. (2017). CLIL teacher education: Where do we stand and where do we need to go? In M. E. Gómez Parra, R. Johnstone (Eds.), *Bilingual education: Educational trends and key competences* (pp. 129–144). Ministerio de Educación, Cultura y Deporte. <https://www.researchgate.net/publication/320298606>
- Carsten-Wickham, B (2008). Assessment and foreign languages: A chair’s perspective. *ADFL Bulletin*, 39(2&3), 36–43.

- Celen, K. M., & Akcan, S. (2017). Evaluation of an ELT practicum programme from the perspectives of supervisors, student teachers and graduates. *Journal of Teacher Education and Educators*, 6(3), 251–274. <https://eric.ed.gov/?id=EJ1207287>
- Cellante, D., & Donne, V. (2013). A program evaluation process to meet the needs of English language learners. *Education*, 134(1), 1–8.
- Chanda, C. (2008). *Teaching and learning of English in secondary schools: A Zambian case study in improving quality*. Commonwealth Secretariat.
- Chang, J.-Y., Kim, W., & Lee, H. (2017). A language support program for English-medium instruction courses: Its development and evaluation in an EFL setting. *International Journal of Bilingual Education and Bilingualism*, 20(5), 510–528. <https://doi.org/10.1080/13670050.2015.1080658>
- Chase, G. (2006). Focusing on learning: Reframing our roles. *Modern Language Journal*, 90(4), 583–588. [https://doi.org/10.1111/j.1540-4781.2006.00466\\_3.x](https://doi.org/10.1111/j.1540-4781.2006.00466_3.x)
- 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 University Press.
- Coleman, H. (1992). Moving the goalposts: Project evaluation in practice. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 222–246). Cambridge University Press.
- Coombe, C., Al-Hamly, M., Davidson, P., & Troudi, S. (Eds.). (2007). *Evaluating teacher effectiveness in ESL/EFL contexts*. The University of Michigan Press.



- Dassier, J. P., & Powell, W. (2001). *Formative foreign language program evaluation: Dare to find out how good you really are. Dimension 2001: The odyssey continues*. Selected proceedings of the 2001 Conference of the Southern Conference on Language Teaching, Birmingham, AL, 15-30.
- Davis, J. M., & McKay, T. H. (2018). *A guide to useful evaluation of language programs*. Georgetown University Press. <https://doi.org/10.2307/j.ctvvngs5>
- Ebtesam, E., & Foster, S. (2019). Implementation of CIPP model for quality evaluation at Zawia University. *International Journal of Applied Linguistics and English Literature*, 8(5), 106–115. <https://doi.org/10.7575/aiac.ijalel.v.8n.5p.106>
- Ecke, P., & Ganz, A. (2014). Student analytics and the longitudinal evaluation of language programs. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 62–82). Cengage Learning.
- Efeoglu, G., Ilerten, F., & Basal, A. (2018). A utilization focused evaluation of the preparatory school of an ELT program. *International Online Journal of Educational Sciences*, 10(4), 1309–2707. <https://doi.org/10.15345/iojes.2018.04.008>
- Ekkens, K., & Winke, P. (2009). Evaluating workplace English language programs. *Language Assessment Quarterly*, 6, 265–287. <https://doi.org/10.1080/15434300903063038>
- Elder, C. (2005). Evaluating the effectiveness of heritage language education: What role for testing? *The International Journal of Bilingual Education and Bilingualism*, 8(2&3), 196–212.

- Elder, C. (2009). Reconciling accountability and development needs in heritage language education: A communication challenge for the evaluation consultant. *Language Teaching Research*, 13(1), 15–33. <https://doi.org/10.1177/1362168808095521>
- Elley, W. B. (1989). Tailoring the evaluation to fit the context. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 270–285). Cambridge University Press.
- Enkin, E., & Correa, M. (2018). Evaluating learner and teacher perceptions of program outcomes in the foreign language major. *Electronic Journal of Foreign Language Teaching*, 15(1), 66–80. <https://e-flt.nus.edu.sg/wp-content/uploads/2020/09/enkin.pdf>
- Ennemoser, M., Kuhl, J., & Pepouna, S. (2013). Evaluation of a dialogic reading program to improve language proficiency in children with a migrant background. *Zeitschrift Für Pädagogische Psychologie*, 27(4), 229–239.
- Eskey, D. E., Lacy, R., & Kraft, C. A. (1991). A novel approach to ESL program evaluation. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 36–53). NAFSA.
- Fall, T., Adair-Hauck, G., Glisan, E. (2007). Assessing students' oral proficiency: A case for online testing. *Foreign Language Annals*, 40(3), 377–406. <https://doi.org/10.1111/j.1944-9720.2007.tb02865.x>
- Fenton-Smith, B., & Torpey, M. J. (2013). Orienting EFL teachers: Principles arising from an evaluation of an induction program in a Japanese university. *Language Teaching Research*, 17(2), 228–250. <https://doi.org/10.1177/1362168813475946>
- Fox, R. P. (1991). Evaluating the ESL program director. In M. C. Pennington (Ed.), *Building better English language programs* (pp. 228–240). NAFSA.

- Gattullo, F. (2000). Formative assessment in primary (elementary) ELT classes: An Italian case study. *Language Testing*, 17(2), 278–288. <https://doi.org/10.1177/026553220001700210>
- Gieve, S., & Cunico, S. (2012). Language and content in the modern foreign languages degree: A students' perspective. *Language Learning Journal*, 40(3), 273–291. <https://doi.org/10.1080/09571736.2011.639459>
- 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(3), 505–540. <https://doi.org/10.1111/j.1944-9720.2009.01034.x>
- Gottlieb, M., & Nguyen, D. (2007). *Assessment and accountability in language education programs: A guide for administrators and teachers*. Caslon Publishing.
- Grim, F. (2010). Giving authentic opportunities to second language learners: A look at a French service-learning project. *Foreign Language Annals*, 43, 605–623. <https://doi.org/10.1111/j.1944-9720.2010.01104.x>
- Griva, E., & Sivropoulou, R. (2009). Implementation and evaluation of an early foreign language learning project in kindergarten. *Early Childhood Education Journal*, 37(1), 79–87. <https://doi.org/10.1007/s10643-009-0314-3jj>
- Grosse, C. U. (2004). The competitive advantage of foreign languages and cultural knowledge. *Modern Language Journal*, 88(3), 351–373. <http://www.jstor.org/stable/3588783>
- Gruba, P., Cárdenas-Claros, M. S., Suvorov, R., & Rick, K. X. (2016d). *Blended language program evaluation*. Springer. <https://doi.org/10.1057/9781137514370>
- Gruba, P., & Chau Nguyen, N. B. (2019). Evaluating technology integration in a Vietnamese university language program. *Computer Assisted Language Learning*, 32(5-6), 619–637. <https://doi.org/10.1080/09588221.2018.1527365>

- Hajjaj, A., & Al-Najjar, B. (1989). ESL program evaluation: Realities and perspectives. In J. E. Alatis (Ed.), *Georgetown University round table on languages and linguistics, 1989* (pp. 133–141). Georgetown University Press.
- Hampel, R., & de los Arcos, I. A. (2013). Interacting at a distance: A critical review of the role of ICT in developing the learner-context interface in a university language programme. *Innovation in Language Learning and Teaching*, 7(2), 158–178.  
<https://doi.org/10.1080/17501229.2013.776051>
- Hargreaves, P. (1989). DES-IMPL-EVALU-IGN: An evaluator's checklist. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 35–47). Cambridge University Press.
- Harklau, L., Norwood, R. (2005). Negotiating researcher roles in ethnographic program evaluation: A postmodern lens. *Anthropology and Education Quarterly*, 36(3), 278–288.  
<https://www.jstor.org/stable/3651396>
- Harris, J. (1990). The second language programme-evaluation literature: Accommodating experimental and multifaceted approaches. *Language, Culture, and Curriculum*, 3(1), 83–92.
- Harris, J. (2009). Late-stage refocusing of Irish-language programme evaluation: Maximizing the potential for productive debate and remediation. *Language Teaching Research*, 13(1), 55–76. <https://doi.org/10.1177/1362168808095523>
- Hedge, T. (1998). Managing developmental evaluation activities in teacher education: Empowering teachers in a new mode of learning. In P. Rea-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 132–158). Longman.

- Heining-Boynton, A. L. (1990). The development and testing of the FLES program evaluation inventory. *Modern Language Journal*, 74(4), 432–439. <https://doi.org/10.2307/328517>
- Herzog, M. (2003). Impact of proficiency scale and the oral proficiency interview on the foreign language program at the Defense Language Institute Foreign Language Center. *Foreign Language Annals*, 36(4), 566–571. <https://doi.org/10.1111/j.1944-9720.2003.tb02146.x>
- 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(5), 37–54.
- Horwitz, E. K. (1985). Formative evaluation of an experimental foreign-language class. *Canadian Modern Language Review*, 42(1), 83–90.
- Houston, T. (2005). Outcomes assessment for beginning and intermediate Spanish: One program's process and results. *Foreign Language Annals*, 38(3), 366–376. <https://doi.org/10.1111/j.1944-9720.2005.tb02223.x>
- Hudson, T. D. (1989). Mastery decisions in program evaluation. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 259–269). Cambridge University Press.
- Imani, S. I. (2013). Evaluation of modular EFL educational program (Audio-visual materials translation & translation of deeds & documents). *English Language Teaching*, 6(4), 8–17. <http://dx.doi.org/10.5539/elt.v6n4p8>
- Issa, B. I., Koronkiewicz, B., & Faretta-Stutenberg, M. (2022). Second-language writing in university-level basic language programs: A survey of student and instructor beliefs. *Foreign Language Annals*, 55(2), 383–407. <https://doi.org/https://doi.org/10.1111/flan.12609>

- 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)* [PDF document]. University of Hawai'i, Second Language Teaching & Curriculum Center.
- Jacobson, P. L. H. (1982). Using evaluation to improve foreign language education. *Modern Language Journal*, 66, 284–291.
- Janssen, G., & Restrepo, S. (2019). A utilization-focused program evaluation of an ERPP tutoring service at one Colombian university. In J. N. Corcoran, K. Englander, & L. Muresan (Eds.), *Pedagogies and policies for publishing research in English* (pp. 19–35). Routledge. <https://doi.org/10.4324/9781315151229>
- Jenks, F. L. (1991). Designing and assessing the efficacy of ESL promotional materials. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 172–188). NAFSA.
- Johnson, R. K. (Ed.) (1989). *The second language curriculum*. Cambridge University Press.
- Karava-Doukas, K. (1998). Evaluating the implementation of educational innovations: Lessons from the past. In P. Rea-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 25–50). Longman.
- Kennedy, C. (1988). Evaluation of the management of change in ELT projects. *Applied Linguistics*, 9, 329–342.
- Kennedy, C., & Miceli, T. (2013). In Piazza online: Exploring the use of wikis with beginner foreign language learners. *Computer Assisted Language Learning*, 26(5), 389–411. <https://doi.org/10.1080/09588221.2013.770035>

- Kiely, R. (1998). Programme evaluation by teachers: Issues of policy and practice. In P. Rea-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 78–104). Longman.
- Kiely, R. (2006). Evaluation, innovation, and ownership in language programs. *Modern Language Journal*, 90(4), 597–601. [https://doi.org/10.1111/j.1540-4781.2006.00466\\_8.x](https://doi.org/10.1111/j.1540-4781.2006.00466_8.x)
- Kiely, R. (2009). Small answers to the big question: Learning from language programme evaluation. *Language Teaching Research*, 13(1), 99–116. <https://doi.org/10.1177/1362168808095525>
- Kiely, R., & Rea-Dickens, P. (2005). *Program evaluation in language education*. Palgrave Macmillan.
- Kiely, R., & Rea-Dickens, P. (2009). Evaluation and learning in language programmes. In K. Knapp & B. Seidlhofer (Eds.), *Handbook of foreign language communication and learning* (pp. 663–694). Mouton de Gruyter.
- Kim, Y. (2012). Implementing ability grouping in EFL contexts: Perceptions of teachers and students. *Language Teaching Research*, 16, 289–315. <https://doi.org/10.1177/1362168812436894>
- Klee, C., Melin, C., Sonesson, D., & Tarone, E. (2014). From frameworks to oversight: Components to improving foreign language program efficacy. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 131–153). Cengage Learning.
- Kondo-Brown, K., Davis, J. McE., & Watanabe, W. (2014). Evaluation capacity building in college language programs: Developing and sustaining a student exit survey project. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program*

- direction: Innovation and accountability in language program evaluation* (pp. 15–40). Cengage Learning.
- Koyalan, A. (2009). The evaluation of a self-access centre: A useful addition to class-based teaching? *System*, 37, 731–740. <https://doi.org/10.1016/j.system.2009.09.016>
- Lai, C., Zhao, Y., & Wong, J. (2011). Task-based language teaching in online ab initio foreign language classrooms. *Modern Language Journal*, 95(s), 81–103. <https://doi.org/10.2307/41413421>
- Lee, J. (2009). ESL student teachers' perceptions of a short-term overseas immersion programme. *Teaching and Teacher Education*, 25, 1095–1104. <https://doi.org/10.1016/j.tate.2009.03.004>
- Lett, J. A. (2005). Foreign language needs assessment in the US military. In M. H. Long (Ed.), *Second language needs analysis* (pp. 105–124). Cambridge University Press.
- Li, B., & Tin, T. B. (2013). Exploring the expectations and perceptions of non-native English speaking students in masters level TESOL programs. *New Zealand Studies in Applied Linguistics*, 19(2), 21–35.
- Li, L., & Walsh, S. (2011). Technology uptake in Chinese EFL classes. *Language Teaching Research*, 15, 99–125. <https://doi.org/10.1177/1362168810383347>
- Liskin-Gasparro, J. E. (1995). Practical approaches to outcomes assessment: The undergraduate major in foreign languages and literatures. *ADFL Bulletin*, 26(2), 21–27.
- Liskin-Gasparro, J., & Vasseur, R. (2014). Designing an embedded outcomes assessment for Spanish majors: Literary interpretation and analysis. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 83–110). 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(1), 35–54. <https://doi.org/10.1177/1362168808095522>
- Lo, Y. (2010). Implementing reflective portfolios for promoting autonomous learning among EFL college students in Taiwan. *Language Teaching Research*, 14, 77–95. <https://doi.org/10.1177/1362168809346509>
- Long, M. H. (1984). Process and product in ESL program evaluation. *TESOL Quarterly*, 18, 409–425. <https://doi.org/10.2307/3586712>
- Long, M. H. (2005a). Methodological issues in learner needs analysis. In M. H. Long (Ed.), *Second language needs analysis* (pp. 19–76). Cambridge University Press.
- Long, M. H. (Ed.) (2005b). *Second language needs analysis*. Cambridge University Press.
- Loughrin-Sacco, S. J., Matthews, S. A., Sweet, W. M., & Miner, J. A. (1990). Reviving language skills: A description and evaluation of Michigan Tech’s summer intensive French course. *ADFL Bulletin*, 21(2), 34–40.
- Lynch, B. K. (1990). A context-adaptive model for program evaluation. *TESOL Quarterly*, 24(1), 23–42. <https://doi.org/10.2307/3586850>
- Lynch, B. K. (1992). Evaluating a program inside and out. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 61–99). Cambridge University Press.
- Lynch, B. K. (1996). *Language program evaluation: Theory and practice*. Cambridge University Press.
- Lynch, B. K. (2000). Evaluating a project-oriented CALL innovation. *Computer Assisted Language Learning*, 13(4-5), 417–440. [https://doi.org/10.1076/0958-8221\(200012\)13:4-5;1-E:FT417](https://doi.org/10.1076/0958-8221(200012)13:4-5;1-E:FT417)

- Lynch, B. K. (2003). *Language assessment and programme evaluation*. Edinburgh University Press.
- Ma, Q. (2008). Empirical CALL evaluation: The relationship between learning process and learning outcome. *CALICO Journal*, 26(1), 108–122.  
<https://doi.org/10.1558/cj.v26i1.108-122>
- Mackay, R. (1988). Position paper: Program evaluation and quality control. *TESL Canada Journal*, 5(2), 33–42.
- Mackay, R. (1994). Undertaking ESL/EFL programme review for accountability and improvement. *ELT Journal*, 48(2), 142–149.
- Mackay, R., & Bosquet, M. (1981). LSP curriculum development: From policy to practice. In R. Mackay & J. D. Palmer (Eds.), *Language for specific purposes: Program design and evaluation* (pp. 1–28). Newbury House.
- Mackay, R., Wellesley, S., & Bazergan, E. (1995). Participatory evaluation. *ELT Journal*, 49(4), 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-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 111–131). Longman.
- Markee, N. (1996). *Managing curricular innovation*. Cambridge University Press.
- Martinez Agudo, J. d. D. (2017). What EFL student teachers think about their professional preparation: Evaluation of an English language teacher education programme in Spain. *Australian Journal of Teacher Education*, 42(8), 62–76.  
<https://doi.org/10.14221/ajte.2017v42n8.5>

- 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). Heinle Cengage.
- Martinez-Lirola, M., & Rubio, F. (2009). Students' beliefs about portfolio evaluation and its influence on their learning outcomes to develop EFL in a Spanish context. *International Journal of English Studies*, 9(1), 91–111. <https://doi.org/10.6018/ijes.9.1.90661>
- Mathews, T., J., & Hansen, C., M. (2004). Ongoing assessment of a university foreign language program. *Foreign Language Annals*, 37(4), 630–640. <https://doi.org/10.1111/j.1944-9720.2004.tb02429.x>
- Matthies, B., F. (1991). Administrative evaluation in ESL programs: “How’m I doin’?” In M. C. Pennington (Ed.), *Building better English language programs* (pp. 241–256). NAFSA.
- McAlpine, D., & Dhonau, S. (2007). Creating a culture for the preparation of an ACTFL/NCATE program review. *Foreign Language Annals*, 40(2), 247–259. <https://doi.org/10.1111/j.1944-9720.2007.tb03200.x>
- McDonough, K., & Chaikitmongkol, W. (2007). Teachers' and learners' reactions to a task-based EFL course in Thailand. *TESOL Quarterly*, 41(1), 107–132. <https://doi.org/10.1002/j.1545-7249.2007.tb00042.x>
- Middlebrook, G. C. (1991). Evaluation of student services in ESL programs. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 135–154). NAFSA.
- Milleret, M. (2008). The trials and tribulations of comprehensive program evaluation. *ADFL Bulletin*, 39(2&3), 44–48.

- Miranda, J. L. W., Wells, J. C., & Jenkins, A. (2019). Preparing special education teacher candidates to teach English language learners with disabilities: How well are we doing? *Language Teaching Research*, 23(3), 330–351.  
<https://doi.org/10.1177/1362168817730665>
- Mitchell, R. (1989). Second language learning: Investigating the classroom context. *System*, 17(2), 195–210. [https://doi.org/10.1016/0346-251X\(89\)90033-X](https://doi.org/10.1016/0346-251X(89)90033-X)
- Mitchell, R. (1990). Evaluation of second language teaching projects and programmes. *Language, Culture, and Curriculum*, 3(1), 3–15.
- Mitchell, R. (1992). The “independent” evaluation of bilingual primary education: A narrative account. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 100-140). Cambridge University Press.
- Moreno-Lopez, I., Saenz-de-Tejada, C., & Tami, K. S. (2008). Language and study abroad across the curriculum: An analysis of course development. *Foreign Language Annals*, 41, 674–686.
- Morris, M. (2006). Addressing the challenges of program evaluation: One department’s experience after two years. *Modern Language Journal*, 90(4), 585–588.
- Murdoch, G. (2000). Introducing a teacher-supportive evaluation system. *ELT Journal*, 54(1), 54–64. <https://doi.org/10.1093/elt/54.1.54>
- Myers, L., & Lindsay, N. (2014). Do we speak the same language?: The iterative development of an institutionally mandated foreign language assessment program. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 208–229). Cengage Learning.

- National Research Council of the National Academies. (2007). *International education and foreign languages: Keys to securing America's future*. The National Academies Press.
- 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. <https://doi.org/10.1016/j.tate.2011.07.001>
- Norris, J. M. (2006). The why (and how) of assessing student learning outcomes in college foreign language programs. *Modern Language Journal*, 90(4), 576–583. [https://doi.org/10.1111/j.1540-4781.2006.00466\\_2.x](https://doi.org/10.1111/j.1540-4781.2006.00466_2.x)
- Norris, J. M. (2008). *Validity evaluation in language assessment*. Peter Lang.
- Norris, J. M. (2009). Understanding and improving language education through program evaluation: Introduction to the special issue. *Language Teaching Research*, 13(1), 7–13. <https://doi.org/10.1177/1362168808095520>
- Norris, J. M. (2016). Language program evaluation. *The Modern Language Journal*, 100, 169–189. <http://www.jstor.org/stable/44135002>
- Norris, J. M. (2015). Thinking and acting programmatically in task-based language teaching: Essential roles for program evaluation. In M. Bygate (Ed.), *Domains and directions in the development of TBLT: A decade of plenaries from the international conference*. John Benjamins.
- Norris, J. M., Davis, McE. J., Sinicrope, C., & Watanabe, Y. (Eds.). (2009). *Toward useful program evaluation in college foreign language education*. National Foreign Language Resource Center.

- Norris, J. M., & Mills, N. (2014). Innovation and accountability in foreign language program evaluation. *AAUSC series issues in language program direction*. Heinle.
- Obando Guerrero, G. V., & Sánchez Solarte, A. C. (2018). Learners' satisfaction in two foreign language teacher education programs: Are we doing our homework? *HOW Journal*, 25(1), 135–155. <https://doi.org/10.19183/how.25.1.436>
- Oliver, K., Kellogg, S., & Patel, R. (2012). An investigation into reported differences between online foreign language instruction and other subject areas in a virtual school. *CALICO Journal*, 29, 269–296.
- Palmer, A. (1992). Issues in evaluating input-based language teaching programs. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 144–166). Cambridge University Press.
- Pardo-Ballester, C. (2012). CALL evaluation: Students' perception and use of LoMasTv. *CALICO Journal*, 29, 532–547. <https://doi.org/10.11139/cj.29.3.532-547>
- Pawan, F., & Thomalla, T. G. (2006). Making the invisible visible: A responsive evaluation study of ESL and Spanish language services for immigrants in a small rural county in Indiana. *TESOL Quarterly*, 39(4), 683–705. <https://doi.org/10.2307/3588527>
- Patton, M. Q. (2022). *Utilization-focused evaluation* (5th ed.). Sage publications.
- Peacock, M. (2009). The evaluation of foreign-language-teacher education programmes. *Language Teaching Research*, 13(3), 259–278. <https://doi.org/10.1177/1362168809104698>
- Pellettieri, J. (2011). Measuring language-related outcomes of community-based learning in intermediate Spanish courses. *Hispania*, 94, 285–302. <https://doi.org/10.1353/hpn.2011.0041>

- Pennington, M. C. (1991). *Building better English language programs: Perspectives on evaluation in ESL*. NAFSA.
- Pennington, M. C., & Brown, J. D. (1991). Unifying curriculum process and curriculum outcomes: The key to excellence in language education. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 57–74). NAFSA.
- Pennington, M. C., & Young, A. L. (1991). Procedures and instruments for faculty evaluation in ESL. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 191–227). NAFSA.
- Plakans, L., & Burke, M. (2013). The decision-making process in language program placement: Test and nontest factors interacting in context. *Language Assessment Quarterly*, *10*(2), 115–134. <https://doi.org/10.1080/15434303.2011.627598>
- Plass J. L. (1998). Design and evaluation of the user interface of foreign language multimedia software: A cognitive approach. *Language Learning & Technology*, *2*(1), 35–45. <http://dx.doi.org/10125/25031>
- Ponder, R., & Powell, B. (1991). Creating and operating a statistical database for evaluation in an English language program. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 155–171). NAFSA.
- Prabhu, N. S. (1987). *Second language pedagogy*. Oxford University Press.
- Rarick, D. (2009). Resuscitating university language programs in the global age: The international engineering program at the university of Rhode Island. *ADFL Bulletin*, *41*, 113–124. <https://doi.org/10.1632/adfl.41.1.113>

- Rea-Dickens, P., & Germaine, K. P. (1998). The price of everything and value of nothing: Trends in language program evaluation. In P. Rea-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 3–19). Longman.
- Rea-Dickens, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing*, 18(4), 429–462.  
<https://doi.org/10.1177/026553220101800407>
- Rea-Dickens, P., & Germaine, K. (1992). *Evaluation*. Oxford University Press.
- Rea-Dickens, P., & Germaine, K., P. (Eds.). (1998). *Managing evaluation and innovation in language teaching: Building bridges*. Longman.
- Ricardo-Osorio, J. G. (2008). A study of foreign language learning outcomes assessment in U.S. undergraduate education. *Foreign Language Annals*, 41(4), 590–610.  
<https://doi.org/10.1111/j.1944-9720.2008.tb03319.x>
- Richards, J. (2001). Approaches to evaluation. In J. Richards (Ed.), *Curriculum development in language teaching* (pp. 286–309). Cambridge University Press.
- Rifkin, B. (2003). Oral proficiency learning outcomes and curricular design. *Foreign Language Annals*, 36(4), 582–588. <https://doi.org/10.1111/j.1944-9720.2003.tb02148.x>
- Roberts, C. (1998). Language and cultural issues in innovation: The European dimension. In P. Rea-Dickens & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 51–77). Longman.
- Rogan, F., & San Miguel, C. (2013). Improving clinical communication of students with English as a second language using online technology: A small scale evaluation study. *Nurse Education in Practice*, 13(5), 400–406. <https://doi.org/10.1016/j.nepr.2012.12.003>



- Ross, S. (1992). Program-defining evaluation in a decade of eclecticism. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 167–195). Cambridge University Press.
- Ross, S. (2003). A diachronic coherence model for language programme evaluation. *Language Learning*, 53(1), 1–33. <https://doi.org/10.1111/1467-9922.00209>
- Ross, S. J. (2009). Program evaluation. In M. Long & C. Doughty (Eds.), *Handbook of second and foreign language teaching* (pp. 756–778). Blackwell.
- Sanders, R. F. (2005). Redesigning introductory Spanish: Increased enrollment, online management, cost reduction, and effects on student learning. *Foreign Language Annals*, 38(4), 523–532. <https://doi.org/10.1111/j.1944-9720.2005.tb02519.x>
- Schmitt, E. (2014). Seat time vs. proficiency: Assessment of language development in undergraduate students. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 110–130). Cengage Learning.
- Schneider, A. I. (2000). Title VI funding for undergraduate international study programs: Long-term impact on language offerings. *ADFL Bulletin*, 32(1), 42–47.
- Schulz, R. A. (2007). The challenge of assessing cultural understanding in the context of foreign language instruction. *Foreign Language Annals*, 40(1), 9–26. <https://doi.org/10.1111/j.1944-9720.2007.tb02851.x>
- Shaver, A. (2012). Fostering integrative motivation among introductory-level German students through a language partners program. *Die Unterrichtspraxis/Teaching German*, 45, 67–73. <https://doi.org/10.1111/j.1756-1221.2012.00119.x>

- Seçer, Ş. Y. E., & Erişen, Y. (2020). Evaluation of English language teaching undergraduate programs in terms of document analysis. *International Journal of Curriculum and Instruction*, 12(2), 671–703.
- Shintani, N. (2011). A comparative study of the effects of input-based and production-based instruction on vocabulary acquisition by young EFL learners. *Language Teaching Research*, 15(2), 137–158. <https://doi.org/10.1177/1362168810388692>
- Shintani, N. (2013). The effect of focus on form and focus on forms instruction on the acquisition of productive knowledge of L2 vocabulary by young beginning-level learners. *TESOL Quarterly*, 47(1), 36–62. <https://doi.org/10.1002/tesq.54>
- Shneyderman, A., & Abella, R. (2009). The effects of the extended foreign language programs on Spanish-language proficiency and academic achievement in English. *Bilingual Research Journal: The Journal of the National Association for Bilingual Education*, 32, 241–259. <https://doi.org/10.1080/15235880903372829>
- Slimani, A. (1992). Evaluation of classroom interaction. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 197–221). Cambridge University Press.
- Snow, M. A., & Brinton, D. M. (1988). Content-based language instruction: Investigating the effectiveness of the adjunct model. *TESOL Quarterly*, 22(4), 553–574. <https://doi.org/10.2307/3587256>
- Spaventa, L., J., & Williamson, J., S. (1991). Participatory placement: A case study. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 75–97). NAFSA.

- Su, Y. (2011). The effects of the cultural portfolio project on cultural and EFL learning in Taiwan's EFL college classes. *Language Teaching Research*, 15(2), 230–252.  
<https://doi.org/10.1177/1362168810388721>
- Sullivan, J. H. (2006). The importance of program evaluation in collegiate foreign language programs. *Modern Language Journal*, 90(4), 590–593. [https://doi.org/10.1111/j.1540-4781.2006.00466\\_6.x](https://doi.org/10.1111/j.1540-4781.2006.00466_6.x)
- 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). Heinle Cengage.
- Towell, R., & Tomlinson, P. (1999). Language curriculum development research at university level. *Language Teaching Research*, 3(1), 1–32.  
<https://doi.org/10.1177/136216889900300102>
- Tribble, C. (2000). Designing evaluation into educational change processes. *ELT Journal*, 54 (4), 319–327. <https://doi.org/10.1093/elt/54.4.319>
- Tsou, W., & Chen, F. (2014). ESP program evaluation framework: Description and application to a Taiwanese university ESP program. *English for Specific Purposes*, 33, 39–53.  
<https://doi.org/10.1016/j.esp.2013.07.008>
- Tucker, G. R., & Cziko, G. A. (1978). The role of evaluation in bilingual education. In J. E. Alatis (Ed.), *International dimensions of bilingual education* (pp. 111–124). Georgetown University Press.

- Uzun, L. (2016). Evaluation of the latest English language teacher training programme in Turkey: Teacher trainees' perspective. *Cogent Education*, 3(1).  
<https://doi.org/10.1080/2331186X.2016.1147115>
- Van den Branden, K. (2006b). Training teachers: Task-based as well? In K. Van den Branden (Ed.), *Task-based language teaching in practice* (pp. 217–273). Cambridge University Press.
- Van Viegen, S., & Russell, B. (2019). More than language—Evaluating a Canadian university EAP bridging program. *TESL Canada Journal*, 36(1), 97–120.  
<https://doi.org/10.18806/tesl.v36i1.1304>
- Wang, P. (2013). Can automated writing evaluation programs help students improve their English writing? *International Journal Of Applied Linguistics & English Literature*, 2(1), 6–11.  
<https://doi.org/10.7575/ijalel.v.2n.1p.6>
- Warford, M. K. (2006). Assessing target cultural literacy: The Buffalo state experience. *ADFL Bulletin*, 37(2–3), 47–57.
- Weir, C., & Roberts, J. (1994). *Evaluation in ELT*. Blackwell.
- Wenzhong, Z., & Cheng, Z. (2013). An Empirical Study on the Curriculum Construction of Business English for International Trade Based on the Case of GDUFS. *English Language Teaching*, 6(4), 19–28. <https://doi.org/10.5539/elt.v6n4p19>
- Wherritt, I., & Cleary, T. A. (1990). A national survey of Spanish language testing for placement or outcome assessment at B.A.-granting institutions in the United States. *Foreign Language Annals*, 23(2), 157–165.

- Williams, E. (2007). Extensive reading in Malawi: Inadequate implementation or inappropriate innovation? *Journal of Research in Reading*, 30(1), 59–79.  
<https://doi.org/10.1111/j.1467-9817.2006.00328.x>
- Windham, S. (2008). Redesigning lower-level curricula for learning outcomes: A case study. *ADFL Bulletin*, 39(2&3), 31–35. <https://doi.org/10.1632/adfl.39.2.31>
- Winskowski-Jackson, C. (1991). Evaluation of culture components in ESL programs. In M. C. Pennington (Ed.), *Building better English language programs: Perspectives on evaluation in ESL* (pp. 98–134). NAFSA.
- Wright, B. D. (2006). Learning languages and the language of learning. *Modern Language Journal*, 90(4), 593–597. [https://doi.org/10.1111/j.1540-4781.2006.00466\\_7.x](https://doi.org/10.1111/j.1540-4781.2006.00466_7.x)
- 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. <https://doi.org/10.1111/j.1944-9720.2013.12005.x>
- Yamada, M. (2018). Evaluation of an EFL teacher training program in Japan. *Journal of Applied Social Science*, 12(1), 25–45. <https://www.jstor.org/stable/26370954>
- Yang, W. (2009). Evaluation of teacher induction practices in a US university English language program: Towards useful evaluation. *Language Teaching Research*, 13(1), 77–98.  
<https://doi.org/10.1177/1362168808095524>
- Yavuz, A., & Zehir Topkaya, E. (2013). Teacher educators' evaluation of the English language teaching program: A Turkish case. *Novitas ROYAL (Research on Youth and Language)*, 7(1), 64–83.
- 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(1), 160–181.

- Yu-Chih, D. S., & Yun-Pi, Y. (2019). Evaluating an English elite program in Taiwan using the CIPP model. *Journal of Asia TEFL*, 16(1), 1738–3102.  
<http://dx.doi.org/10.18823/asiatefl.2019.16.1.13.200>
- Zannirato, A. (2014). Promoting Change in “Two-tiered” departments: An exploratory evaluation of conflict and empowerment among language and literature faculty. In N. Mills & J. M. Norris (Eds.), *AAUSC 2014 volume – Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 154–182). Cengage Learning.
- Zapata, G. (2011). The effects of community service learning projects on L2 learners’ cultural understanding. *Hispania*, 94, 86–102. <https://doi.org/10.1353/hpn.2011.0024>
- Zareva, A., & Fomina, A. (2013). Strategy use of Russian pre-service TEFL university students: Using a strategy inventory for program effectiveness evaluation. *International journal of english studies (IJES)*, 13(1), 69–88. <https://doi.org/10.6018/ijes/2013/1/148711>