TESL-EJ
Teaching English as a Second or Foreign Language

The Electronic Journal for English as a Second Language

August 2016 – Volume 20, Number 2

An Evaluation Framework for CALL

Benjamin L. McMurry
Brigham Young University
<ben.mcmurry@byu.edu>

Richard Edward West
Brigham Young University
<rickwest@gmail.com>

Peter J. Rich
Brigham Young University
<peter_rich@byu.edu>

David Dwayne Williams
Brigham Young University
<david_williams@byu.edu>

Neil J. Anderson
Brigham Young University
<neil_anderson@byuh.edu>

K. James Hartshorn
Brigham Young University
<james_hartshorn@byu.edu>

Abstract

Searching prestigious Computer-assisted Language Learning (CALL) journals for references to key publications and authors in the field of evaluation yields a short list. The *American Journal of Evaluation*—the flagship journal of the American Evaluation Association—is only cited once in both the *CALICO Journal* and *Language Learning and Technology* (Blyth & Davis, 2007). Only two articles in these journals have cited Robert Stake, Robert Yin, Daniel Stufflebeam, Michael Scriven, or Michael Patton, five of the most influential evaluators of our generation. Prestigious CALL journals lacked references to formal evaluation literature, which provides a wealth of information regarding effective evaluation processes.

We reviewed prominent CALL evaluation frameworks and literature in formal evaluation. A comparison of CALL evaluation with formal evaluation showed some gaps in CALL evaluation. Practices and insights from the field of evaluation would benefit CALL researchers and practitioners with regard to conducting systematic evaluations that report evaluation findings that other researchers and practitioners find useful. The proposed evaluation framework includes common evaluation tasks conducted by evaluators in the field of formal evaluation to produce a workflow model for designing and conducting evaluations in CALL. Implications for CALL evaluators and stakeholders indicate several areas for improvement in CALL evaluation.

Introduction

Searching prestigious CALL journals for reference to key publications and authors in the field of evaluation yields a short list. The *American Journal of Evaluation* — the flagship journal of the American Evaluation Association — is only cited once in both *CALICO Journal* and *Language Learning and Technology* (Blyth & Davis, 2007). Only two articles in these journals have cited Robert Stake, Robert Yin, Daniel Stufflebeam, Michael Scriven, or Michael Patton, five of the most influential evaluators of our generation. Whereas now it seems largely ignored, the field of evaluation could provide valuable insight to CALL researchers and practitioners with regard to conducting systematic evaluations that report evaluation findings that other researchers and practitioners find useful.

Chapelle (2010) stated that "the amount of published work on materials evaluation [in second language learning (SLL)] is surprisingly small in view of the impact that materials have in the instructional process" (p. 67). However, few authors have provided guidelines for evaluation of instructional materials in SLL (Cunningsworth, 1984; Sheldon, 1988; Skierso, 1991; Tomlinson, 2003). Additionally, prominent authors in CALL have proposed evaluation frameworks for evaluating CALL (Beatty, 2010; Burston, 2003; Chapelle, 2001, 2007, 2010; Garrett, 2009; Hubbard, 1988, 1996, 2006, 2011; Leakey, 2011; Levy & Stockwell, 2006; Reeder, Heift, Roche, Tabyanian, Schlickau, & Golz, 2004; Susser, 2001; Villada, 2009). While these frameworks have their strengths, the systematic approach and practices of formal evaluators such as those of the American Evaluation Association (AEA), would help CALL evaluators design and conduct evaluations that are methodologically similar to formal evaluations, while maintaining the diversity offered by these various frameworks.

While the proposed frameworks for CALL evaluation are far from unsystematic, none explicitly describe workflows for conducting evaluations. They all implicitly mention evaluation tasks such as identifying the object of evaluation, determining the purpose of the evaluation, collecting and analyzing data, and reporting findings and implications. However, looking through the lens provided by formal evaluators, some aspects of each task are overlooked. Furthermore, the task of metaevaluation (i.e., evaluating the evaluation) does not appear to be addressed in CALL evaluation frameworks. A framework for conducting evaluation of CALL must provide systematic steps that are based on proven evaluation practices. A systematic approach to designing and conducting quality evaluations would produce a body of transferable and usable data to help inform developers, researchers, practitioners, and students.

In this article, we propose a conceptual framework for designing and conducting evaluations in CALL that builds on the works of Hubbard (1987, 1988, 1996, 2006), Chapelle (1999, 2001, 2007, 2010) and others (Beatty, 2010; Burston, 2003; Leakey, 2011; Levy & Stockwell, 2006; Reeder et al., 2004; Susser, 2001; Villada, 2009) and incorporates literature from the field of evaluation. The purpose of the conceptual framework is to answer the following question: What is an appropriate systematic approach for conducting evaluations in CALL?

Methodology

This review of the literature followed Machi and McEvoy's (2010) six steps for producing a literature review: selecting a topic, searching the literature, developing an argument, surveying the literature, critiquing the literature, and writing the review. When looking at literature, we decided to search for journal articles and key authors in the Linguistics/Language Behavior Abstracts (LLBA) database because it provides the richest search options for CALL when compared with Educational Resources Information Center (ERIC). Through LLBA we looked for publications containing the thesaurus entry *Computer-Assisted Language Learning*, which narrowed the results to 585 articles. We further limited the search by including the term *evaluation*. From these 82 results we looked for articles that discussed evaluation in terms of proposed and actual processes. Apart from LBBA queries, it was even more helpful to look at landmark articles that were frequently cited and review the works of prominent CALL scholars, such as Chapelle and Hubbard. Searching the contents of both the *CALICO* and *Language Learning and Technology* journals did yield a plethora of results relating to CALL and evaluation, but as previously mentioned, only two articles in these journals mention prominent evaluation articles or authors.

Next, we searched for literature on evaluation processes and standards. ERIC provided these three thesaurus terms: *evaluation*, *evaluation methods*, and *evaluation research*. A search for program evaluation standards provided several articles that address the implementation of the standards. Looking for literature regarding evaluation processes, tasks, and standards in ERIC was much less successful. Many articles resulting from the search were key articles in evaluation with which we were familiar. However, these publications by key evaluation authors such as Patton, Scriven, Stake, and Stufflebeam proved to be the greatest source of information.

Evaluation Defined and Described

Evaluation is a systematic process that seeks to determine the value, worth, importance, significance, or merit of an object, person, or activity (Stake & Schwandt, 2006; Yarbrough, Shulha, Hopson, & Caruthers, 2010). At the core of evaluation, evaluators seek to gather and interpret data that helps us make judgments about the evaluated (Nunan, 1992). These judgments are arrived at by the interpretation of data, whether quantitative, qualitative, or both.

Systematic evaluation methodologies are similar to research methodologies. Dunkel (1991) and Leakey (2011) use *evaluation* and *effectiveness research* synonymously. In fact, Nunan (1992) declared, “I believe that evaluations, incorporating as they do questions, data and interpretation, are a form of research” (p. 184). Levy and Stockwell (2006) referred to this as the research-evaluation nexus. However, there are notable differences in purpose between evaluation and research. The aim of most research is to produce “credible, generalizable knowledge about the nature of the world around us,” whereas “evaluations help stakeholders answer specific questions or make decisions . . . [and] investigate such things as a program’s development, processes, theory, viability, outcomes and impact” (Yarbrough, Shulha, Hopson, & Caruthers, 2010, p. xxv).

We use a broad definition of evaluation in this article. There are several types of evaluation that are regularly used in language learning contexts. For example, assessment is a form of evaluation where testing tools are used to evaluate learner proficiency or aptitude. Teacher evaluations serve to inform stakeholders and the teachers themselves regarding their performance as educators. Accrediting bodies evaluate programs against set standards. Evaluation is a large umbrella that covers many facets of education. In this article, we talk about this umbrella form of evaluation.

One example worth mentioning is the way evaluation is often mentioned in articles concerning technology delivered learning experiences such as MOOCs and Learning Management systems. While the authors are unaware of any frameworks that are similar in purpose to those by the aforementioned CALL scholars, there is an abundance of articles that address evaluation as it applies to student and teacher assessment. Frameworks presented in these articles propose models of the overarching learning experience with evaluation of student learning being an integral part. They also report the results of evaluations and research, but do not explicitly provide a framework or model for evaluation in the terms we previously outlined.

The term *formal evaluation* is used throughout the article to represent the field of professional evaluation; it refers to the ideals, principles, standards and practices of *formal evaluators*. The evaluation field does not belong to one specific group or domain; rather, it refers to the professional practice of formal evaluation. One example of the breadth of professional evaluation is the number of organizations represented on the Joint Committee on Standards for Educational Evaluation (JCSEE). Sponsored by more than 15 different evaluation organizations, JCSEE has published standards in program evaluation, personnel evaluation, and student evaluation. These standards are widely accepted among professional evaluators and are a driving force in the field of evaluation.

Analysis of Popular CALL Frameworks

The most popular evaluation frameworks in CALL have been posited by Hubbard (1987, 1988, 1996, 2006, 2011) and Chapelle (2001), each of which has strengths and weaknesses. In the following paragraphs, we provide an overview of Hubbard's and Chapelle's frameworks. To analyze the frameworks, we considered each principle identified in the evaluation literature as key for successful evaluations and then considered how and in what ways each principle applied to the Chapelle and Hubbard frameworks. We listed similarities and overlapping ideas. We then asked two professional evaluators and two CALL scholars to review the framework for deficiencies, overlapping ideas, and clarity.

Hubbard's Framework

We will first look at Hubbard's (2011) latest publication on his framework. From the onset, he mentions four distinct purposes of CALL evaluation: selection for use in a course, selection for use in self-access environments or for other instructors, published reviews, and feedback during the development process. Hubbard presented these purposes not as a comprehensive list but as a specific subset to which his framework can be applied.

Hubbard narrowed the list of evaluation approaches or methodologies to three specific types: checklists, methodological frameworks, and SLA research. Checklists are essentially a combination of criteria that evaluators review and to which they assign some type of score using either a Likert scale or other rating systems. While this is a common methodology used to evaluate CALL, in many cases it assumes that the evaluation criteria are one size fits all. Evaluators can change and alter a checklist to match the criteria specified by stakeholders, but the use of checklists as an approach or *methodology* for CALL evaluation may be confounded by its overlap with evaluation *criteria*. These checklists tend to be a list of evaluation criteria and may not provide adequate methodological concerns to CALL evaluation, omitting key procedures in the evaluation process.

The other two approaches mentioned are not exempt from similar phenomena. Methodological frameworks, as described by Hubbard, allow evaluators to form their own questions. In this regard, the evaluator may be the sole stakeholder. This may be a limitation for evaluators with little experience who rely on CALL evaluation frameworks such as this one to guide their evaluation. Thus, it could become increasingly easy for evaluators to neglect potential stakeholders.

Hubbard (1988) based his original framework on the works of Phillips (1985) and Richards and Rogers (1982). While the confusion between criteria and methodologies in Hubbard's framework is not as prominent as it is with checklists, evaluators may interpret his suggestions as prescribed criteria. Lastly, SLA approaches seem to focus on methodologies and criteria concerning language acquisition issues. Once again, it seems as though criteria and methodologies are being grouped together. It is clear that there is a relationship between evaluation criteria and evaluation methodologies; however, a versatile evaluation framework for CALL should tease these apart to allow more options to evaluators when evaluating CALL. In short, evaluation criteria consist of the attributes by which the evaluand is judged, and methodology refers to the approach used to learn about the evaluand with regard to those criteria.

Hubbard mentioned that his description of the framework reflected the purpose he felt most common, which is selection by a teacher for use in the classroom. However, he also argued that the framework could be applied to the other three purposes: selection for use in self-access environments or for other instructors, published reviews, and feedback during the development process. While possible, the framework might not be as accommodating to these other evaluands.

Figure 1 is a diagram of Hubbard's (2011) framework. The processes Hubbard outlined in various iterations of his proposed framework include steps such as giving a technical preview, creating an operational description, considering learner and teacher fit, making appropriateness judgments, and implementing schemes.

Following his outlined workflow may limit the quality of the evaluation because it fails to explicitly mention important steps in the evaluation process, which include considering the values of stakeholders and outlining clear evaluative criteria. Additionally, his framework is often associated with and geared toward the evaluation of CALL courseware and websites.

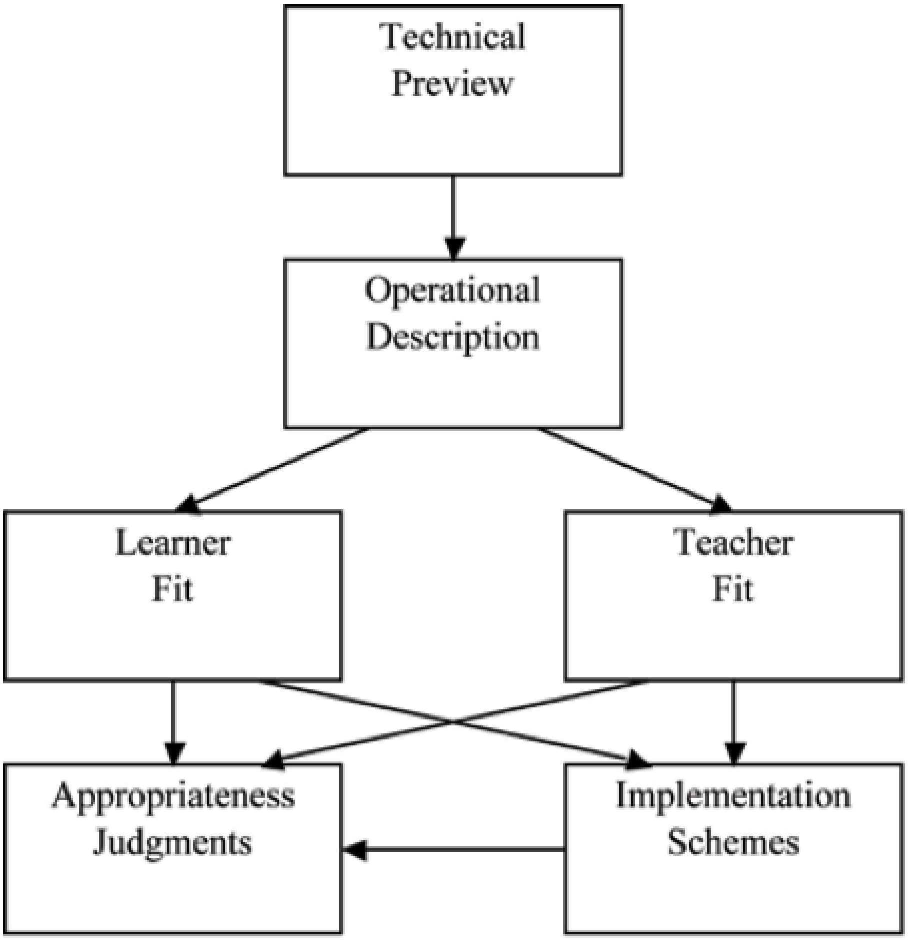

***Figure 1*. Hubbard's (2011) Evaluation Framework**

Hubbard did not specifically address the various stakeholders that may need to be considered when evaluating CALL. However, other stakeholders were mentioned in Hubbard's CALL methodology, of which evaluation is only one module. The three modules—development, implementation, and evaluation — do interact with one another, which may provide for more interactions with stakeholders. Nonetheless, with regard to the evaluation module, Hubbard focused primarily on teachers and learners. One purpose mentioned, the provision of feedback during development, entails the considerations of CALL developers in an evaluation but fails to include other possible stakeholders such as parents. Additionally, school administration and staff are not mentioned as possible stakeholders, both of whom are important because of their responsibility with financing and implementing CALL. In short, Hubbard's framework may be a good place to start when evaluating courseware or websites pertaining to language learning, but its specificity may limit its effectiveness for evaluators with different evaluation purposes, evaluands, stakeholders, and criteria. A CALL evaluation framework should be broad enough to guide potential CALL evaluators in various situations and purposes.

Chappelle's Framework

Chapelle's (2001) framework for evaluation varies from Hubbard's. From the outset, it is clear that she did not limit the types of evaluands as strictly as Hubbard. Her framework was broad enough to consider CALL software, teacher-planned CALL activities, and learners' performance during CALL activities. She also lists standards for selecting the evaluation criteria and even suggests some specific criteria. Chapelle (2001) discussed the importance of criteria based in SLA research and stated that, "learning language potential should be the central criterion in the evaluation of CALL" (p. 52). She also lists learner fit, meaning focus, authenticity, positive impact, and practicality as criteria to be considered in CALL evaluations.

Chapelle suggested that CALL evaluations should be looked at from two perspectives: (a) a judgmental analysis of CALL software and activities, and (b) an empirical analysis of the learner's performance. In many ways, this could be a recommendation for various research methodologies. She implied that the evaluand dictates, at least to some degree, the type of analysis that should be done in an evaluation. Table 1 shows three types of analyses with her suggested evaluand (object of evaluation), question, and evaluation type (method of evaluation) for each analysis.

While some CALL evaluands would appear to be best evaluated qualitatively and others quantitatively, this too may be limiting. Rather than basing the type of approach on the nature of the evaluand, it should be based on a series of factors including the nature of the evaluand, the evaluation questions, and the evaluation criteria.

Hubbard (2011) and Chapelle's (2001) frameworks differ in their focus, with Hubbard emphasizing process, including its parts and the details and specific suggestions for each step. For example, he spelled out various purposes of a CALL evaluation. Chapelle on the other hand focused less on creating a procedural map for conducting evaluations and more on purpose, criteria, and methodologies.

Table 1. ***Chapelle's (2001, 2007) Evaluands, Questions, and Evaluation types***

Level of analysis	Object of evaluation	Example question	Method of evaluation
1	CALL software	Does the software provide learners the opportunity for interactional modifications to negotiate meaning?	Judgmental
2	Teacher-planned CALL activities	Does the CALL activity designed by the teacher provide learners the opportunity to modify interaction for negotiation of meaning?	Judgmental
3	Learner's performance during CALL activities	Do learners actually interact and negotiate meaning while they are working in a chat room?	Empirical

To summarize, Chappelle's framework may be effective in certain circumstances, particularly those evaluating CALL for language issues, but it may not be helpful when considering other non-SLA issues. Evaluators who are considering non-SLA issues, such as the finances, infrastructure, or other administrative aspects, would benefit from a framework that allows for other considerations such as financial issues or hardware requirements. Following Chappelle's framework may also generate evaluations that do not consider the values of underrepresented stakeholders such as software developers and program administrators.

Leakey (2011) took Chapelle's framework a few more steps forward. He argued that CALL evaluation needs a common agenda and a set of benchmarks by which such evaluation can be measured. Starting with the criteria suggested by Chappelle, Leakey outlines 12 *CALL enhancement criteria*. He also outlines possible data collection methodologies that mirror research methodology. His three case studies focus on the evaluation of platforms, programs, and pedagogy. Leakey also provided several tables that CALL evaluators can use to guide the evaluation with regard to the 12 criteria and three evaluand types. While the expanded list of criteria and detailed tables may be helpful, the evaluation may simply result in completed checklist. Nonetheless, Leakey, in many ways, did what he intended by bringing the criteria together into one model and suggesting a common agenda based on said criteria.

While the frameworks of Hubbard and Chappelle (as well as similar frameworks based on these) have strengths and may be viable in certain situations, we propose a framework that is adaptable to various contexts and dictated by the values of stakeholders, which may match values and evaluative purposes of the frameworks of Hubbard and Chapelle, but are not limited to them.

Framework for Designing and Conducting CALL Evaluations

After reviewing the popular CALL evaluation frameworks, we propose a framework that is essentially borrowed from frameworks and practices in formal evaluation. It aims to provide guidance in conducting more effective CALL evaluations. Its purpose is to provide direction to evaluators in conducting systematic evaluations using procedures from seasoned evaluators resulting, we believe, in comprehensive and reliable evaluations that would be more informative, efficient, useful, replicable, and to some degree transferable.

Figure 2 shows each task in relation to the others. It focuses on the careful crafting of a purpose-driven evaluation that helps evaluators identify the evaluand and stakeholders, set evaluative criteria, and determine the purpose and type of the evaluation. Performing the aforementioned tasks leads to the constructing of evaluation questions. Based on the results of previous tasks, evaluators can design the data collection and evaluation procedures, collect and analyze the data, and report the findings and implications. The rounded rectangle at the background of the figure represents *metaevaluation* and emphasizes the constant need to evaluate each task throughout the process of the evaluation. As each task is evaluated, evaluators may need to return to previous tasks or look at and plan for future tasks as indicated.

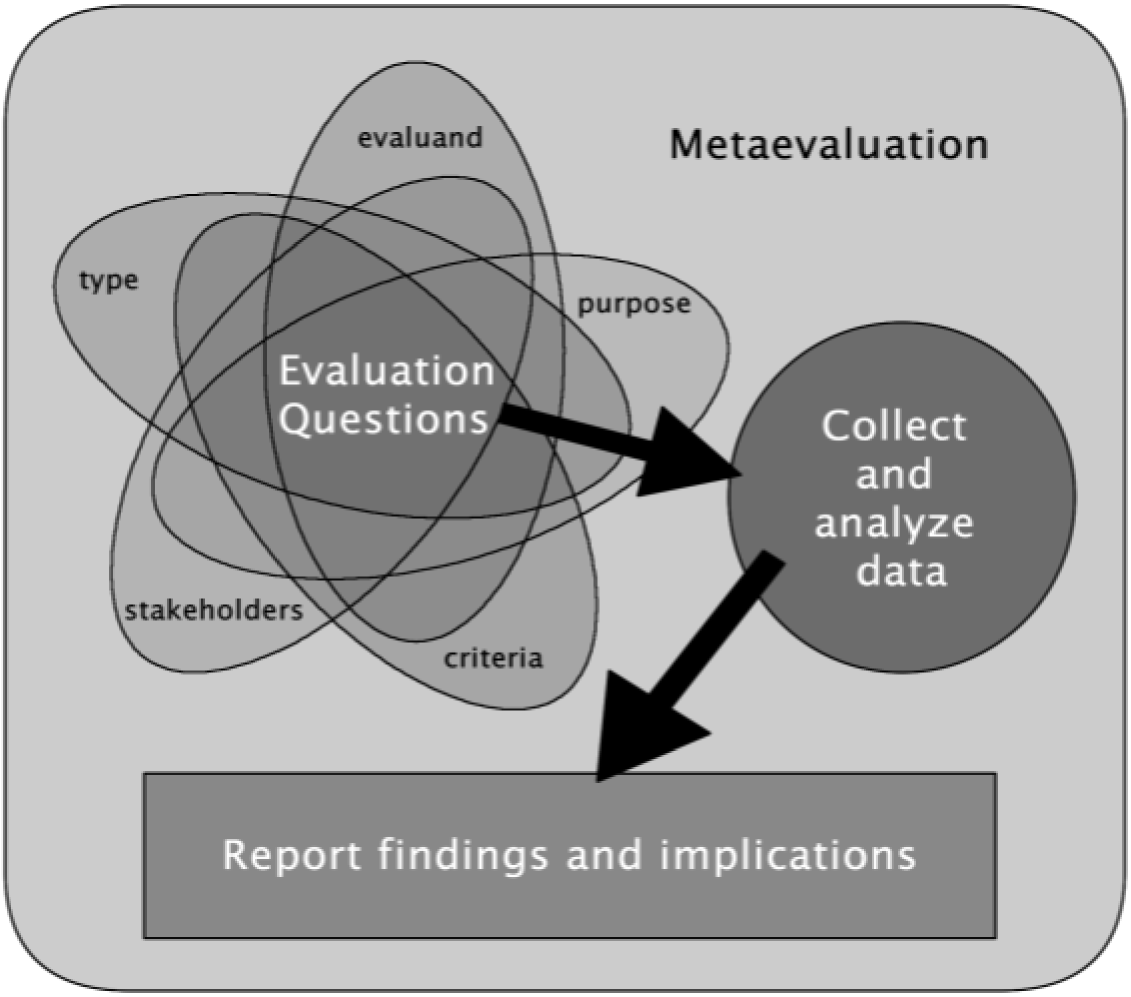

Figure 2. **Framework for systematic CALL evaluation.**

Figure 2 also illustrates tasks for CALL evaluators. Though this is not a new framework to those in evaluation, this paradigm may be new to some CALL evaluators. Rather than limit or specify the details of an evaluation, we suggest a framework that is specific enough to guide evaluators through a tried and tested process and broad enough to accommodate the evaluation of any activity or material related to CALL. This framework provides CALL educators an additional tool to use to gather evaluation data from multiple sources in order to make the best decisions to improve language learning.

Comparison of CALL Evaluation Frameworks to Formal Evaluation Tasks

Popular CALL evaluation frameworks have many similarities and differences when compared to formal evaluation tasks. Table 2 maps formal evaluation tasks to activities mentioned in the frameworks of Hubbard (2011) and Chapelle (2007, 2011). Essentially there are activities that match, but there are a few differences. For example, many areas focused on by Hubbard were narrow and highly specific. The framework we propose is less constrained and can be applied to the evaluation of any evaluand — any CALL activity or any CALL material.

Table 2. ***Formal Evaluation Tasks, Hubbard's (2011) Framework, and Chapelle's (2001, 2007) Framework***

Evaluation task	**Hubbard's framework**	**Chapelle's framework**
Identify the evaluand	Courseware and websites	Complete course, technology component of a course, and technology pedagogy
Identify stakeholders	Teachers and learners	Insiders (software developers, other CALL researchers), informed critics (other teachers, learners, other applied linguists), and Outsiders (other applied linguists, program decision makers, policy decision makers)
Set evaluative criteria	Technical considerations, operational description, teacher fit, learner fit	Language learning potential, meaning focus, learner fit, authenticity, positive impact, and practicality
Define a purpose	Selection for a course, selection for self-access or other instructors' use, reviews, and providing feedback for development	*Connected to identifying the evaluand*
Select an evaluation type	Grouped with data collection	Teachers' judgment, performance data, and synthesis of judgment and performance data
Develop evaluation questions	*Based on evaluative criteria*	*Implies the use of research questions*
Collect and analyze data	Checklists, methodological frameworks, SLA-based approaches	Qualitative: ethnographic and case study, interaction analysis and discourse analysis, and experimental and quasi-experimental
Report findings and implications	*Connected to evaluation purpose*	*Mentioned when talking about audiences*
Evaluate the evaluation	—	—

Evaluands

Hubbard claimed that his framework could be adapted toward aspects of CALL outside of courseware and websites. However, it is only focused on these two evaluands. Chapelle's suggested evaluands are placed on a continuum ranging from the evaluation of a complete course to technology used in a traditional face-to-face learning. While the spectrum allows for a wide range of evaluands, it may exclude other CALL tools or activities. The continuum focuses on technology as used in the context of a course, whether in face-to-face or online environments. Leakey (2011) provides three evaluand types: platforms, programs, and pedagogy. Other possible evaluands may include technologies and activities that learners engage in independent of formal class settings.

Stakeholders

Hubbard's (1996) CALL methodology may mention several potential stakeholders as part of the development and implementation modules, but these are not always explicitly or specifically tied to the evaluation module of his methodology. Although Hubbard briefly mentions developers as stakeholders when evaluating developing products, he limits stakeholders to teachers and learners. While in many cases these are the more prominent, there are always other stakeholders to consider. For example, school administrators, parents, and developers may be vital and perhaps underrepresented groups that can help increase the completeness and validity of the evaluation. Chapelle included a lengthy list of types of evaluations and possible audiences that may also be counted as stakeholders. This can serve as a list of audiences to whom evaluations might be reported and a list of possible stakeholders.

Criteria

One strength of Hubbard's framework is the extensive list of criteria to be considered when evaluating CALL courseware and websites. However, this is not an exhaustive list and should only be a starting point. The suggestions are specific to courseware and website evaluations, but there may be other criteria important to stakeholders. Additionally, there are no suggestions for criteria for evaluating CALL activities or even hardware used in language learning. However, Leakey's (2011) emphasis on platforms, programs, and pedagogy may cover a more inclusive list of evaluands.

Purposes

While Chapelle made no explicit mention of evaluation purposes, her continuum of evaluands (ranging from the evaluation of a complete course to technology used in a traditional face-to-face learning) may indicate implied purposes similar to those Hubbard mentioned: selecting technology for a course, selection for self-access or other instructor use, and published reviews.

Evaluation Types

Unlike many evaluation types used by formal evaluators, Hubbard's evaluation types seemed to be tied to the data collection methods rather than the purpose of the evaluation. Hubbard (2011) cited Levy and Stockwell's (2006) three types of courseware evaluation in CALL: checklists and forms, methodological frameworks, and SLA-based approaches. Formal evaluators may adopt an evaluation type based on the purpose of the evaluation, and while some types may favor a particular type of data collection, they are not limited by it.

While Chapelle (2001) did not specify any specific type of evaluation, she did group evaluations into three distinct categories. She mentioned evaluations that are based on teachers' judgment, those that are focused on the analysis of performance data, and then evaluations that may be a synthesis of the two former types. These seem to be largely based in the participants of the evaluations (i.e., teachers and students).

Evaluation Questions

Both Hubbard and Chappelle grouped the creation of evaluation questions with other formal evaluation tasks. Hubbard suggested that the questions be connected to data collection methods, while Chapelle implied the use of research questions. Both have merit and are essential areas to consider when drafting evaluation questions, but evaluation questions come from information from stakeholders and the identification of the evaluand, criteria, purpose, and type of evaluation.

Collection and Analysis of Data

As mentioned previously, Hubbard's mixed types of evaluation with types of data collection. In contrast, Chapelle (2001) suggested using research methodologies to collect data to be used in evaluations. She mentioned qualitative and quantitative approaches. She discussed the use of ethnographies, case studies, interaction analyses, discourse analyses, and both experimental and quasi-experimental methodologies. Depending on the questions asked about the evaluand, evaluators can choose a methodology that would best lend itself to the collection of valid and reliable data. Leakey's (2011) model included a workflow that addressed both qualitative and quantitative data collection methodologies.

Report of Findings

In Hubbard's framework, the method for reporting findings and implications is linked to the evaluation's purpose. For example, the findings of a courseware evaluation may only be reported to the teacher and perhaps the students. When writing a review of software or a website, the findings may be published in a journal or website. Chapelle largely discussed her list of possible stakeholders as possible audiences, but she provided little elaboration.

In addition to the suggested audiences for evaluation reports, there are several venues for reporting findings, and evaluators should be encouraged to find ways to share results. Leakey (2011) emphasized the importance of reporting on evaluations for the benefit of the field as a whole. Because many CALL research articles focus on the efficacy of a tool or activity, they are not entirely different from an evaluation. If an evaluation uses research methods to examine the evaluand, turning a CALL evaluation into a publishable article may be possible.

The frameworks of Hubbard and Chapelle are useful frameworks that incorporate several principles as described in formal evaluation. They also provide examples or prescriptions for conducting CALL evaluation. Both may be too narrow to apply to a broad range of evaluands and evaluations. Neither emphasized the importance of metaevaluation. Our proposed framework is broad enough to capture the majority of evaluation needs of the CALL community. It also encourages delineation between various tasks that lead to a more focused, methodological, and systematic evaluation.

Evaluation Tasks

We have briefly mentioned the nine evaluation tasks used by formal evaluators and compared them to popular CALL frameworks. In this section, we will look at each of the nine evaluation tasks. With regard to CALL evaluation, Nunan (1992) provided some guiding questions for designing evaluations that mirror the processes of formal evaluators. He suggested identifying the purpose, audience, procedures, instruments, evaluators, timeframe, costs, and reporting procedures. Formal evaluators pose similar questions that can be divided into nine primary tasks: (a) identifying the evaluand, (b) identifying stakeholders, (c) determining the purpose of the evaluation, (d) selecting an evaluation type, (e) setting evaluation criteria, (f) asking evaluation questions, (g) collecting and analyzing the data, (h) reporting findings and implications, and (i) evaluating the evaluation.

Before conducting an evaluation, evaluators and stakeholders work together to design the evaluation. Stufflebeam (2003) suggested that identifying the evaluand, identifying stakeholders, determining an evaluation purpose, selecting the type of evaluation, and identifying values and criteria were essential activities evaluators engage in when designing the evaluation. These are the first five tasks that should be considered when designing and conducting an evaluation.

Identifying the Evaluand

The evaluand is the object, person, or activity being evaluated. Program evaluation is often used to discuss evaluands that include educational programs, policies, or projects. In order to conduct an evaluation, evaluators need to identify both what the evaluand is and what the evaluand should be. Identifying the evaluand is not always an easy process, and often times there are several evaluands that can be explored. Hubbard's (1988, 1996, 2006, 2011) framework focused on tools such as courseware and websites as evaluands. Levy and Stockwell (2006) suggested software, online courses, websites, computer-mediated communication, and combinations of elements (i.e., learning management systems) as evaluands. Reeder et al. (2004) suggested three types of software that could be the object of evaluations: microcosm situations, microethnographies, and online programs.

Chapelle's (2001, 2010) guidelines for evaluating CALL tended to focus on activities as evaluands. She later identified what is taught in a complete course, what is taught through technology in a complete course, and what is taught through technology as three evaluation targets. She defined a complete course as one that is web-delivered or technology based. By looking at what is taught through technology in a complete course, she suggested that evaluators look at a subset of the total course objectives as the evaluand. What is taught through technology refers to technology that is used to enhance a traditional face-to-face classroom (Chapelle, 2007).

Because the most common frameworks for CALL evaluation focus on materials and activities, many published evaluations focus on these two evaluands. However, the following evaluands may also be considered when conducting an evaluation: a student using CALL, a class using CALL, a teacher using CALL, or a school using CALL. When looking at materials, evaluations do not need to be limited to software but can also include hardware. When looking at CALL activities as evaluands, these need not be constrained by teacher-led or classroom activities but may include autonomous language learning activities. In short, evaluands need not be limited by those suggested in evaluation frameworks. Furthermore, any framework for evaluation should be adaptable to a myriad of evaluands that may include, but are not limited to, MOOCs and other online courses, social networks, multimedia use, mobile tools, and other emergent mediums.

Identifying Stakeholders

Stakeholders are all those who have some interest in the evaluand and by extension the evaluation. In educational settings, students, teachers, and administrators are the most frequently identified stakeholders. Often evaluators overlook curriculum and materials developers, parents, funding agencies, and other members of the community. Evaluators are sometimes overlooked as stakeholders, but indeed become an interested party in the evaluand when conducting an evaluation (Sechrest, Babcock, & Smith, 1993). They may bring expertise and unique views to the evaluation design that can help other stakeholders in the development of quality evaluations. Carefully identified stakeholders can provide essential information about the evaluand and help shape the evaluation.

Key authors in CALL have identified possible stakeholders for CALL evaluations. Levy and Stockwell (2006) discussed how designers and developers of CALL materials are evaluators. They pointed out that many published CALL research articles focus on tools and their evaluation. They stated, “The designer-evaluator perspective is a very important one in contemporary CALL evaluation” (p. 52). Chapelle (2001) argued that due to the complexity of CALL evaluation, all those who use CALL should be involved in the evaluation process. In a later article, Chapelle (2007) discussed audiences for evaluations. Regardless of the differences between stakeholders and audience, there is some overlap. She identified insiders, informed critics, and outsiders as three separate audiences. Insiders include software developers and other CALL researchers. Other teachers, learners, other applied linguists, and program decision makers are among the informed critics. Outsiders may overlap with critics and include program and policy decision makers. Hubbard (1988, 1996, 2006, 2011) gave high priority to teachers and learners as stakeholders.

Villada’s (2009) proposed interpretive evaluation framework focused on multivocality as one of its main tenets. He defined multivocality as multiple voices or perspectives. Of the 24 articles on evaluations that he reviewed, 14 only addressed the perspective of the teacher. Guba and Lincoln (1981) stated that, “the evaluator has a duty to identify all audiences, to do his best to determine what their concerns and issues are, and to honor and respond to those concerns and issues” (p. 306).

CALL evaluators can ensure that stakeholders are represented by asking the right questions. Guba and Lincoln (1981) suggested asking three types of questions focusing on developers, users, and nonusers of the evaluand: Who was involved in producing the evaluand, and who is using the evaluand? Who benefits from the evaluand? Who does not benefit or is disadvantaged by the evaluand? Asking these questions and including perspectives of the stakeholders will produce more useful and effective evaluations. These questions lead the evaluator to explore previously underrepresented groups including (but not limited to) students, nonusers, and developers.

Each evaluation has its own unique purpose and design, which are mentioned later. Due to this, in many occasions the values and opinions of one group of stakeholders may carry more weight than others. The purpose of the evaluation and the venue in which the results are intended to be published affect the target audience and included stakeholders. Evaluations that are intended to be generalized to other similar situations and that resemble evaluative-research have a different readership when published in journals, monographs, or other books than when used for internal, decision-laden purposes. As the audience of the evaluation changes, some individuals or entities may no longer be stakeholders. Others may still be stakeholders whose input regarding evaluative criteria and procedures are not needed, and in some situations, intentionally avoided.

One example of this is the role of developers as stakeholders. During the development phase, developers are stakeholders whose input into the evaluation is of upmost importance. In order to improve upon a product or service, they may have pointed out questions that they would like to have answered as they continue development. Post-production evaluations of CALL products do not invalidate developers as stakeholders. They are still concerned with the success of their product, but at this point their input into the evaluation design might prove to be biased or skewed. Their role as stakeholders has changed from one that informs evaluation design to one that is more concerned that the evaluation is not biased to their detriment.

Evaluators as stakeholders can also prove to be problematic. Granted, in many "in house" evaluations, the evaluators often fill other roles such as that of teacher or administrator. Evaluators with multiple roles or vested interests in the outcome of the evaluation must recognize potential issues and conflicts of interests. In some cases, this may mean that the evaluators clearly outline their roles and biases from the beginning. In other cases, evaluators might need to step back and assume a role more in line with that of a researcher to maintain the integrity of the evaluation.

Regardless of the make-up of the stakeholders, it is important that evaluators consider all stakeholder concerns and values. Often, the ideas and values of the individual or group soliciting the evaluation may be given priority. Of course, the prioritization of the values of stakeholders depends on the other aspects of evaluation such as the purpose.

Determining the Purpose of the Evaluation

With the evaluand in mind and in collaboration with stakeholders, evaluators should define a clear purpose for conducting the evaluation and ask detailed questions to guide the rest of the process. Hubbard (1988, 1996, 2006, 2011) suggested the following possible purposes: (a) selection for a course; (b) selection for self-access or other instructor use; (c) reviews; and (d) feedback for development. Levy and Stockwell (2006) suggested that investigating the effectiveness of new materials as a purpose for evaluation is "one of the unique, defining features of CALL" (p. 43). They also discussed purposes such as seeing if CALL materials are working as they should, assessing value and effectiveness of CALL materials, and learning about viability and effectiveness of specific methodologies and strategies. Their review of the literature also led them to other purposes such as assessing student attitudes and perceptions, obtaining feedback from students about CALL courses and courseware, and investigating learners' views on features of the tools they are using.

Tomlinson (2003) argued that materials evaluation "involves making judgments about the effect of the materials on the people using them" (p. 15). He continued by including an exhaustive list of possible purposes that materials evaluation seeks to measure. Below are a few that relate to CALL:

- Appeal of the materials to learners
- Credibility, validity, and reliability of materials
- Ability of the materials to interest and motivate students and teachers
- Value of the materials in terms of short and long-term learning
- Learners' and teachers' perceived value of the materials
- Flexibility of materials

Authors of CALL evaluations have generally articulated clear purposes of their evaluations. However, their purposes have resembled those mentioned by Tomlinson (2003). Reeves and Hedberg (2003) emphasized that the purpose of evaluation is to drive decision-making. Looking at decisions that will result from the evaluation may help CALL evaluators identify purposes not previously explored that will direct the evaluation process and produce more useful evaluations.

Selecting the Type of Evaluation

Having defined purposes that were selected by stakeholders while keeping in mind the evaluation's resulting decisions strongly influence what kind of evaluation should take place. Many CALL evaluators and formal evaluators refer to evaluations as either formative or summative. CALL evaluators have delineated other types of evaluations that are tightly connected to their evaluative purposes. Formal evaluators suggest various models for conducting evaluations. These models provide CALL with a wealth of methods and frameworks for conducting evaluations for unique purposes. As CALL evaluators shift their independent evaluation paradigms and adopt or borrow from models that formal evaluators use; the efficacy, efficiency, and quality of CALL evaluations may improve.

Evaluation types in SLL and CALL. It is important to understand the types of evaluations proposed in SLL and CALL materials evaluation. Tomlinson (2003) discussed three types of evaluation. As mentioned earlier, these are connected to the purposes of the evaluation. Pre-use evaluation looks at how the materials might benefit the users. Whilst-use evaluations investigate the value of the materials while they are being used. Post-use evaluation, he argued, is the "most valuable (but least administered) type of evaluation as it can measure the actual effects of materials on users" (p. 25). Chapelle (2001) described two types of CALL evaluation: judgmental and empirical. She argued that judgmental analyses examine the "characteristics of the software and the tasks," while empirical analyses are based on "data gathered to reveal the details of CALL use and learning outcomes" (p. 54). Reeder et al. (2004) promoted a similar dichotomy in evaluation calling one type introspective and the other empirical. They argued that introspective evaluations often result in completed checklists or reviews. Introspective evaluations often use similar criteria and provide information about the material and are based on the reviewers' perspectives. Like Chapelle (2001, 2007), they argued that empirical evaluations involve looking at students in authentic situations. CALL authors tended to mix the type of evaluation with the methods for *collecting data*.

Evaluation types in formal evaluation. Formal evaluation literature is abundant in proposed types of evaluation. Table 3 provides brief descriptions of a selection of formal evaluation types. The purpose of describing these is to introduce specific ways to guide CALL evaluators. However, we do not provide extensive information for conducting each type of evaluation.

Table 3. ***Selection of Formal Evaluation Types, Descriptions, and Possible CALL Applications***

Model and author	Description	Possible CALL application
Responsive Evaluation Stake (1975, 2003, 2004)	Focuses on adapting the evaluation to changing, diminishing, or emerging concerns and issues of stakeholders.	Evaluation of a CALL resource or activity in response to concerns from students, teachers, or administrators.
Illumination Evaluation Parlett and Hamilton (1974)	Discovers and describes the underlying principles and issues of the evaluand.	Evaluation of a CALL resource or activity prior to use in a class, lab, or program.
Goal-Free Evaluation Scriven (1972)	Evaluators work independently of evaluand users to determine what the evaluand actually is or does instead of determining whether it meets goals and objectives.	Evaluation of CALL tools independent of learning outcomes. (i.e., software reviews in journals)
Effectiveness Evaluation Reeves and Hedberg (2003)	Determines if the evaluand is reaching short-term goals or objectives.	Evaluation of learning outcomes from use of CALL during the course of a semester.
Impact Evaluation Reeves and Hedberg (2003)	Used to determine if what is learned through the evaluand is actually transferred to its intended context.	Evaluation of language skills acquired via CALL in comparison to actual language proficiency gain.
CIPP Model Stufflebeam (2003)	Uses the core values of stakeholders to evaluate goals (context), plans (input), actions (process), and outcomes (product) of the evaluand.	Comprehensive evaluation that considers the context, use, and outcomes of a CALL tool.

Model and author	Description	Possible CALL application
Utilization-Focused Evaluation Patton (2003, 2008)	Focuses on intended use of the evaluation by the intended users of the evaluation.	Evaluation of CALL designed and conducted for use by program administrators or other decision makers.
Developmental Evaluation Patton (2010)	Used to evaluate innovative evaluands and adapts to issues in complex environments.	Formative evaluation of CALL software or hardware during the development process.

Lynch (1996) is possibly the only author in SLL that has situated evaluation ideals in the context of language learning. He discussed the responsive model, the illumination model, and goal-free evaluation, among others. Other prominent models include Stufflebeam's (2003) CIPP model that includes context, input, process, and product evaluations. Reeves and Hedberg (2003) discussed effectiveness evaluation and impact evaluation. Patton has also introduced utilization (participant-oriented) evaluation (2003, 2008) and developmental evaluation (2010).

Examples of formal evaluation types used in CALL evaluation. Several of the proposed purposes for evaluation of CALL correspond well to the types of evaluations that formal evaluators use. Although we provide only a small description of each evaluation type, we also suggest some scenarios in which each evaluation may be appropriate in the context of CALL. We elaborate on a few types and situations in the following paragraphs.

Responsive evaluation. For example, Stake's (1975, 2003, 2004) responsive evaluation may be suited to the evaluation of prototypes used in teaching. The nature of changing values and concerns of stakeholders can be difficult to account for, but Stake's model may be helpful in addressing such issues.

Illuminative evaluation. What Chapelle (2001) and Reeder et al. (2004) called judgmental or introspective evaluation may be a closer resemblance to illuminative evaluation. However, this type of evaluation is not limited to materials as evaluands and can help evaluators conduct judgmental or introspective evaluations of activities because it also provides for the exploration of student tasks and experiences (Parlett & Hamilton, 1976).

Utilization-focused evaluation. Utilization-focused evaluations may have a unique fit in CALL. As academic journals increasingly publish evaluations of CALL materials, editors and publishers should consider how these evaluations will actually be used by the stakeholders. Patton (2003) stated that, "utilization-focused evaluation is concerned with how real people in the real world apply evaluation findings and experience the evaluation process. Therefore, the focus in utilization-focused evaluation is on the intended use by intended users" (p. 223).

Developmental evaluation. The developmental evaluation model, with its affordances for complexity, is also well-suited for CALL, which in itself incorporates the complexity inherent in technology use and language learning. Perhaps one application of this model could be throughout the development process of new products and even changing and evolving language learning curricula that employ or rely heavily upon CALL. These are only a few models that formal evaluation literature has to offer CALL. When CALL evaluators consider the various types of evaluation at their disposal, they can take advantage of tried practices by experienced evaluators, which will strengthen their evaluative skills and the evaluations that they produce. The type of evaluation is connected to its purposes and affects the questions, criteria, methods, and reporting of findings of the evaluation.

Setting Evaluation Criteria

Levy and Stockwell (2006) stated, "The nature of the object of the evaluation is important in choosing suitable criteria" (p. 71). With a clear understanding of the evaluand and the purpose of the evaluation, evaluators select criteria or standards by which to judge the evaluand. Criteria should reflect the values of stakeholders and be linked to the purpose of the evaluation and the questions you are asking about the evaluand. For example, administrators may consider low operating costs and ease of teacher adoption to be important. These criteria would be considered when determining the merit or worth of the evaluand.

The two prominent frameworks by Hubbard (1987, 1988, 1996, 2006, 2011) and Chapelle (2001, 2007, 2011) provided limited suggestions for determining criteria by which to judge their proposed evaluand. These frameworks and other evaluations do not specifically articulate exploring the nature of the evaluand, its intended purposes, and their relationship with the values of stakeholders in an organic approach generated by evaluators and stakeholders. In fact, the authors of these frameworks tend to be more prescriptive regarding criteria used in evaluation. Hubbard suggested looking at technical considerations, operational descriptions, teacher fit, learner fit, and implementation. Burston's (2003) suggestions mimicked Hubbard's, but he also suggested that software be pedagogically valid and adaptable to the curriculum as well as efficient, effective, and pedagogically innovative. Hubbard and Burston inferred that these criteria are sufficient for evaluation of software. While they may serve as starting points, evaluations that follow these guidelines lack the consideration of stakeholder values and may fail to address the intended outcomes of the evaluation.

Chapelle (2001) emphasized that "evaluation criteria should incorporate findings and theory-based speculation about ideal conditions for SLA" (p. 52). Language learning potential, learner fit, meaning focus, authenticity, positive impact, and practicality are the six criteria she recommended. While these overlap with some of Hubbard's (1987, 1988, 1996, 2006, 2011) criteria, the focus revolves more around language learning. Leakey's (2011) list contains a fairly comprehensive list of criteria ranging from language learning potential to tuition delivery modes. With regard to this, Reeder et al. (2004) argued that there is a lack of identified criteria that addresses both learning outcomes and learning processes. They also stated that evaluative criteria often fail to connect to design and instructional methodologies. While Chapelle's (2001) criteria may help evaluators and stakeholders determine their own evaluative criteria, selecting criteria based on the values of evaluators is curiously absent. Considering the desired outcomes of the evaluand

and incorporating them with stakeholders' values may address the concerns of Reeder et al. (2004).

It is important to recognize that all evaluations will not have the same criteria. The purpose shapes the way criteria are used. Language learning potential may be one of the more prominent criterion in CALL evaluations. However, other criteria may have a higher priority based on the overall purpose of the evaluation. For example, administrators may consider practicality to be a more important criterion when evaluating two pieces of software that are considered similar with regard to language learning potential.

Evaluation seeks to determine what should be. By clearly articulating the criteria or standards by which the evaluand will be measured, evaluators can have a directed study that lends itself to clear and defensible results, leading to clear and defensible decisions. Only by considering the intended evaluand, outcomes, and stakeholders' values can effective criteria be selected and used in the evaluation process.

Developing Evaluation Questions

With a purpose and type of evaluation in mind, a clearly identified evaluand, and set evaluation criteria, asking questions about the evaluand should be a well-informed task. Figure 3 demonstrates the relationship this task has with the previous ones.

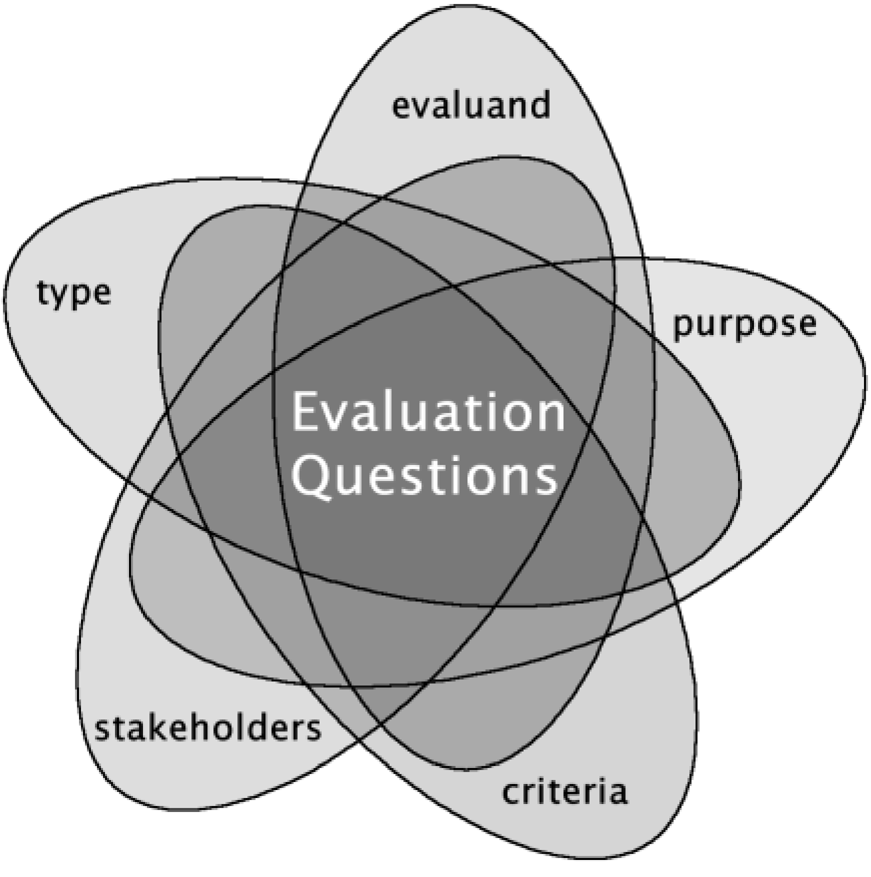

***Figure 3*. Evaluation tasks with the development of evaluation questions.**

Stake (2010) emphasized the importance of asking questions before selecting methods to collect data. However, he also pointed out that in some situations one may define evaluation or research questions, select methods to collect data, and then return to the question to tweak it to work better with the chosen methods. Figure 3 shows that each of the previous tasks are intertwined and culminate in the questions that drive the evaluation.

In order for questions that match the evaluation, evaluators must work with stakeholders to help arrive at overall evaluation questions. What do we (the evaluators and stakeholders) want to know about the evaluand? What questions do we need to ask to achieve our purpose? In many ways, the process of determining appropriate and evaluable questions resembles the same process that researchers undertake when asking research questions. It is essential that evaluators and stakeholders realize the connectivity among the first five evaluation tasks and developing evaluation questions. In concert, those five tasks help generate effective evaluation questions.

For example, personnel in an intensive English program (stakeholders) use a mobile application for L2 literacy (evaluand), and want to see if it leads to an increase in reading fluency over the course of a semester (criteria) to determine if they should continue using the application (purpose). The evaluators and stakeholders decide that an effectiveness evaluation (type) is best suited for this situation. With this information they develop the main evaluation question: What effect does the mobile application have on the reading fluency of students?

Collecting and Analyzing Data

With clear evaluation questions, evaluators can design the data collection procedures and begin collecting data. Once again, the nature of the evaluand, the purpose of the evaluation, the criteria, and the evaluation questions are paramount in determining the data collection methods. The epistemological preference of the evaluator and stakeholders may inform the methodologies to be used and the types of data that are collected.

As mentioned earlier, CALL researchers and evaluators often divide evaluation into two groups: judgmental (also referred to as introspective) and empirical (Chapelle 2001, 2007; Reeder et al., 2004).

Judgmental evaluation. Evaluators commonly use checklists in judgmental evaluation of CALL materials. However, Susser (2001) suggested that the use of checklists for CALL and SLA evaluation have been criticized. He identified several areas in which various scholars described the shortcomings of checklists. Decoo (1994) questioned the accuracy, compatibility, and transferability of checklists. Squires and McDougall (1994) suggested that checklists may place too much focus on the technical features of courseware and neglect to consider pedagogical concerns. Susser (2001) referred to several experimental studies that questioned the objectivity of checklists; he argued that some checklists may favor a particular theory of language acquisition or computer-assisted instruction. He also pointed out that the background knowledge of those completing the checklist may affect the criticality of the reviewer and the overall accuracy of the completed checklist.

Despite these arguments against checklists, Susser (2001) defended their use and explained that if they are used in the right contexts for the right purposes, checklists can be an excellent tool. Chapelle (2001, 2007), Hubbard (1988, 1996, 2011), and Reeder et al. (2004) all acknowledged the value of checklists for judgmental purposes, or in other words, for describing CALL materials and activities.

Empirical evaluations. Chapelle (2001, 2007) advocated that quantitative, qualitative and mixed methods approaches can be used in empirical evaluations. She discussed that much of the research in CALL is based on theoretical comparison research that employs experimental, quasi-experimental designs, as well as interaction and discourse analysis. A theoretically motivated approach to CALL evaluation may include qualitative methodologies such as ethnographies and case studies in addition to the more quantitative experimental and quasi-experimental methodologies.

Leakey (2011) gave a fairly detailed description of the role of data collection in evaluation. He expounded on the *Research Onion* (Saunders, Thornhill, & Lewis, 2006) to classify various types of data collection. He also proposed an evaluation diamond that illustrated qualitative and quantitative methods over time that focused on the three evaluand types (platform, programs, and pedagogy).

Reeves et al. (2005) argued that design research might be more effective than traditional experimental research. For example, experimental designs often do not control for all the variables and treatment groups are often unnatural and lack resemblance to a typical classroom. They also noted that when questionnaires are used to collect data from control and treatment groups, they are often narrow in their scope and any learning results not predicted and indicated in the questionnaire may be missed. Quantitative data collection for evaluations is not necessarily inappropriate, and in many cases, it may be the best data collection method depending on the evaluand, the purpose of the evaluation, the evaluation criteria, and the evaluation questions.

Reporting Evaluation Results

Just as in research, evaluation results need to be reported. Unlike researchers who publish their work in books and academic journals with the intention of providing generalizable or transferable findings that add to the body of research, evaluation reports might only be made available to the stakeholders. While in certain contexts a complete report with all the details may be appropriate, for many stakeholders, a brief report that includes a summary of the evaluation is often preferred. Because evaluations are extremely context dependent, they are not generalizable; however, their findings may be transferable. Any insights gained from a specific evaluation could be used to direct evaluations of other programs and some implications might be beneficial.

Sharing evaluation reports with other teachers, students, administrators, and additional interested parties can be helpful. In language programs, evaluations could be made available to others working in the program. Software evaluations are often published in print journals and online. Additionally, evaluations of CALL tools and activities could be posted to a website making them available to a wider audience. Because evaluations are not intended to be generalizable to all contexts, published evaluations should include sufficient information, such as a detailed description of the evaluand, the purpose of the evaluation, and the criteria by which the evaluand was judged so others can understand the context for interpreting evaluation results and apply the findings to their own circumstances as they feel appropriate.

Evaluations are only effective if they are used. Scriven (1990) argued that software developers and users need, "to be willing and able to do and use evaluation" (p. 3), and Patton (2003, 2008) emphasized a utilization-focused evaluation approach for making evaluations useful to stakeholders. Beatty (2010) argued that there is value in making evaluations available so they can be used by others to make conscientious decisions about the technology used in the classroom.

Many published evaluations in CALL journals are either software reviews or research articles exploring software developed by the researcher. The target for these publications tends to be those who read the journals, but research reports and evaluation reports differ in purpose. Thus, CALL frameworks of evaluation and instruction given to a prospective evaluator should stress the importance of working with stakeholders and producing effective reports to help them understand the evaluation and guide resulting decisions. Such decisions may be contextually specific and inappropriate to publish in research or other widely available reports.

Evaluating the Evaluation

Stufflebeam (1974), Scriven (1969), and others have suggested that a metaevaluation (i.e., an evaluation of the evaluation) be included in the evaluation design. Whether done internally or externally, metaevaluations not only provide a type of validation of the evaluation in question, but they also help evaluators conduct the evaluation with clear goals and systematic procedures that these scholars believe often lead to more reliable and unbiased evaluations. The metaevaluation should be conducted throughout the process and not just following the completion of all the evaluation tasks. To this end, the Joint Committee on Standards for Educational Evaluation has outlined program evaluation standards (Program Evaluation Standards, 2011) that can be used to conduct metaevaluations and help evaluators focus on the most essential aspects of evaluation. Five areas are emphasized in these standards: utility, feasibility, accuracy, propriety, and evaluation accountability.

In some situations, the metaevaluation need not be formally articulated. In other situations, performing a formal comprehensive metaevaluation both during and after the evaluation can help evaluators recognize limitations or gaps not addressed in the evaluation. When compared with traditional research and design research, a somewhat analogous term might be *meta-analysis*. However, a key difference between meta-evaluation and meta-analysis is that the former involves looking at a single evaluation, while the latter concerns research examining effect size across a wide body of research studies. While a meta-analysis is retrospective, a meta-evaluation is introspective.

Evaluations that are used internally and not made available to the public benefit greatly from metaevaluation. Whereas journal articles and other academic research are examined through peer review, internally used evaluations use metaevaluation as a way to verify the methodology and analysis of an evaluation. Metaevaluations may be conducted by the evaluators, by other external evaluators, or both. In fact, if those conducting the evaluation are carefully metaevaluating through the process and external evaluators provide a comprehensive metaevaluation upon its completion, the overall validity and utility of the evaluation can be better ascertained.

Discussion

Implications

While some prospective reviewers might see these recommendations as being more complex or less practical than traditional frameworks for evaluation in CALL, we believe they empower the evaluator with a methodical approach to ensure that the underlying purposes for the evaluation are achieved. With fewer constraints and systematic opportunities to consider aspects of the evaluation that might otherwise be overlooked, we believe that these recommendations will lead to more effective evaluation in CALL.

Administrators, publishers, designers, and others should strongly consider the following implications. Doing so will result in evaluations that are more systematic, thorough, and useful. Applying the evaluation tasks outlined in this paper might lead to the following suggestions for improving CALL evaluations. The list includes several variations from common trends in CALL evaluation to incorporate more practices from formal evaluators.

1. Evaluators should not limit themselves to CALL materials and activities as evaluands.
2. Evaluators have the responsibility to identify and determine the concerns or perspectives of all possible stakeholders. Looking at only one group (i.e., teachers, students, or developers) may not address all the issues considered in the evaluation.
3. Evaluators need to include stakeholders in articulating the purposes of evaluations. They need to consider their values and concerns.
4. Evaluators should rely on the values of stakeholders and research to establish criteria used to evaluate the evaluand.
5. Evaluators should use research methods to conduct dependable and systematic evaluations.
6. Evaluation reports (i.e., publications, software reviews) should be useful to their audiences.
7. Evaluators should continually evaluate their own evaluations throughout the entire process.
8. Academic journals that publish CALL evaluations should either adhere to evaluation standards such as the Program Evaluation Standards (2011) or develop their own to guide evaluators throughout the evaluation process and help determine the merit of worth of other evaluations.

In addition to these implications, CALL evaluators should try to incorporate formal evaluator practices in their evaluation projects. While the framework is simple and the description limited, we believe that evaluators who use this in their evaluations will produce evaluations that consider a broader range of evaluands, criteria, and stakeholders and benefit CALL.

Using the Framework

Consider the following abbreviated example. Matthew is the coordinator for ESL reading classes in an Intensive English Program (IEP) at a large university. Students in a low proficiency class regularly use a software program called *Phonics4Fun* to help with reading proficiency. Administrators in the IEP are wondering if the software is worth the price they pay each semester to keep the software licensed. Matthew realizes that a formal evaluation may provide administrators with answers to their questions. As he goes through the evaluation framework, he addresses each task.

Purpose, evaluand, and type of evaluation. The purpose of this evaluation is to determine if the institution should continue to use the software program. The evaluand is the actual software program that the reading classes have been using. Matthew decides that an effectiveness evaluation (Reeves & Hedberg, 2003) would be appropriate because it may best help him fulfill the purpose of the evaluation, which is to determine the value of the software program.

Stakeholders and criteria. Matthew identified three main groups of stakeholders: administrators, teachers, and learners. Although there are other stakeholders, such as the software developers and IT personnel, he limited his evaluation to those three groups. In collaboration with them, he was able to articulate some common criteria by which to evaluate the software program. Both the teachers and learners considered the program's effectiveness to be of the utmost importance. Administrators and students were concerned with the financial cost of continuing to use the program. Teachers and administrators were also concerned with the practicality of continued use of the program.

Evaluation questions. At this point, Matthew works with stakeholders to design a small number of evaluation questions, such as: *In what ways does the software impact the students' reading proficiency? How does the overall cost of the software compare to similar phonics curricula? How do teachers and administrators view the practicality of using the software as part of class instruction?*

Data collection. For the first question, Matthew decides to conduct a pre-test and post-test on the students using the software. He decides to use a quasi-experimental design and has one class of students use the software while the other class does not. He decides to collect cost information for comparable programs. In terms of practicality, Matthew decides to take a qualitative approach and interview and observe teachers using precise questions to gauge teacher experience with regard to practicality.

Reporting the findings. Throughout the course of the evaluation, Matthew has been taking notes and keeping detailed records. He prepares a written report that is made available to all stakeholders. His findings show that student reading proficiency is benefited from using the phonics software; they also show that teachers generally feel like its use is practical for implementation in the classroom. Finally, his review of similar software revealed that the software license is quite a bit more expensive that competing products. As part of his report, he provides the suggestion that phonics software not be abandoned, but that perhaps the evaluation of competing products may lead to an equally effective yet less expensive solution.

Meta-evaluation. Throughout the course of the evaluation, Matthew was constantly reviewing his evaluation against the standards outlined in the Joint Committee on Standards for Educational Evaluation (Program Evaluation Standards, 2011). He also worked with a colleague familiar with evaluation and asked for his assistance in evaluating the work he had done.

In this example, Matthew completed the tasks as outlined in this article. The evaluation framework provided the breadth to adapt to his situation and clearly and adequately addressed the needs of stakeholders.

Suggestions for Future Research

From here, there are several questions regarding CALL evaluation that may need to be addressed. First, the efficacy and utility of evaluations that follow this framework should be researched and evaluated. Does this framework provide a more guided approach that is adaptable to various evaluands, stakeholders, and criteria?

Additionally, how might this proposed framework benefit CALL publications including peer-reviewed research and software reviews? Many publications in top-tier CALL journals publish peer-reviewed research regarding the efficacy of author-generated CALL materials or CALL activities. While some may argue that evaluation and research are similar, future research and initiatives regarding the appropriateness of such publications would be helpful. Or in other words, should research that is essentially evaluation be portrayed as research? How does the evaluation of author-generated products benefit the body of CALL research it aims to contribute to?

Similarly, the current conventions for software reviews in these same top-tier journals need to be evaluated. How effective are these published reviews? As outlined in the article, our proposed framework for CALL evaluation includes essential tasks that are missing from popular CALL evaluation frameworks. Software reviews that follow our proposed framework will provide more information to readers and make them more readily usable by those same readers.

Separate from the previous questions regarding the role of evaluation in published literature, CALL evaluators should consider the use of evaluation standards. Is there a need for standards or guidelines similar to those proposed in formal evaluation, or would currently adopted formal evaluation standards such as the Joint Committee on Standards for Educational Evaluation be sufficient for CALL evaluations? Regardless of the answer to this question, it needs to be asked and studied, and the field of CALL should look to the field of evaluation for guidance and understanding.

Lastly, and perhaps more global in nature to the previous suggestions, research may be needed to verify that following procedural models such as the one recommended here positively impact the quality of evaluation. What strengths do frameworks such as the one described here as well as the other frameworks by Hubbard and Chappelle have in regard to conducting evaluations? How do frameworks or procedural approaches strengthen or weaken evaluation reports and their ability to communicate results with stakeholders? How do they constrain the scope of evaluation? These questions may prove to be helpful as we continue to evaluation language learning materials.

Conclusion

In this paper, we have reviewed the popular CALL frameworks and formal evaluation tasks and illustrated the gap between formal evaluation and CALL evaluation. Our proposal to implement formal evaluation practices into CALL evaluation may help provide evaluations that address several issues that have been overlooked in CALL evaluation. The field of CALL needs to be more aware of the practices in mainstream evaluation and apply them when evaluating CALL materials and activities. Formal evaluators and publications have much to offer CALL. There are a plethora of principles, ideals, and practices from which CALL evaluation may benefit. CALL evaluation publications should reflect the expertise that experienced evaluators bring to evaluation and be based on principles similar to those to which formal evaluators espouse.

While this framework is somewhat prescriptive in nature, evaluators must be flexible and adaptable as one framework for evaluation may not be successful in every situation. When evaluators consider the values, goals, questions, and concerns of stakeholders, quality evaluations are generated that provide insightful data that can create positive changes. Finally, as SLA stakeholders commit to use these tried and true methods of evaluation, which have been established across a variety of contexts in myriad domains, we may begin to see greater improvement in understanding both the tools and processes that support second language acquisition.

About the Authors

Benjamin McMurry is the Curriculum Coordinator at Brigham Young University's English Language Center. He actively develops language-learning materials and conducts qualitative research. His current research interests include instructional design, evaluation, and the experiences of language teachers and learners.

Richard E. West is an assistant professor in the Instructional Psychology and Technology department at Brigham Young University, where he teaches courses on instructional technology foundations, theories of creativity and innovation, technology integration for preservice teachers, and academic writing and argumentation. He researches the design, support, and evaluation of environments that foster collaborative innovation, as well as online learning communities and K-16 technology integration. His research is available through his files on Mendeley, Google Scholar, and his personal website (richardewest.com).

Peter Rich is an associate professor in Instructional Psychology and Technology at Brigham Young University. His research focuses on the cognitive aspects of learning and on the use of video self-analysis improve teaching and learning. Dr. Rich has organized video-analysis researchers from over a dozen different universities and across several countries to synthesize knowledge of how video is best used to aid in teacher self-evaluations. Through these associations, he seeks to advance understanding and use of video annotation software in educational situations.

David D. Williams is an Instructional Psychology and Technology professor at Brigham Young University. He conducts and studies evaluations of teaching and learning in various settings. He studies interactions among stakeholders as they use their values to shape criteria and standards for evaluating learning environments. He conducts research on informal and formal evaluation and how people use evaluation to enhance learning.

Neil J Anderson is a Professor at Brigham Young University-Hawaii. His research interests include second language reading, motivation, language learning strategies, and teacher leadership. He has had two Fulbright research/teaching fellowships, one in Costa Rica (2002-2003) and in Guatemala (2009-2010).

K. James Hartshorn has been involved in second language education in the United States and Asia for more than two decades. He currently serves as the Associate Coordinator of Brigham Young University's English Language Center. Professional interests include curriculum development, teacher training, pronunciation, and second language writing.

References

Beatty, K. (2010). *Teaching and researching: Computer-assisted language learning* (2nd ed.). London: Pearson Education Limited.

Blyth, C., & Davis, J. (2007). Using formative evaluation in the development of learner-centered materials. *CALICO Journal, 25*(1), 1-21.

Burston, J. (2003). Software selection: A primer on sources and evaluation. *CALICO Journal, 21*(1), 29-40.

Chapelle, C. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing and research*. Cambridge: Cambridge University Press.

Chapelle, C. (2007). Challenges in evaluation of innovation: Observations from technology research. Innovation in *Language Learning and Teaching, 1*(1), 30-45.

Chapelle, C. (2010). The spread of computer-assisted language learning. *Language Teaching, 43*(1), 66-74.

Cunningsworth, A. (1984). *Evaluating and selecting EFL teaching materials*. London: Heinemann.

Decoo, W. (1994). In defense of drill and practice in CALL: A reevaluation of fundamental strategies. *Computers & Education, 23*(1-2), 151-158.

Dunkel, P. (1991). Research on effectiveness of computer-assisted instruction and computer-assisted language-learning. In P. Dinkel (Ed.), *Computer-assisted language-learning and testing: Research issues and practice* (pp. 1-36). New York: Newbury House.

Garrett, N. (2009). Computer-assisted language learning trends and issues revisited: Integrating innovation. *The Modern Language Journal, 93* (focus issue), 719-740.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches*. San Francisco, CA: Jossey-Bass Inc.

Hubbard, P. (1987). Language teaching approaches, the evaluation of CALL software and design implications. In W.F. Smith (Ed.), *Modern media in foreign language education: Theory and implementation* (pp. 227-254). Lincolnwood, IL: National Textbook Company.

Hubbard, P. (1988). An integrated framework for CALL courseware evaluation. *CALICO Journal, 6*(2), 51-72.

Hubbard, P. (1996). Elements of CALL methodology: development, evaluation, and implementation. In M. Pennington (Ed.), *The power of CALL* (pp. 15-32). Houston, TX: Athelstan.

Hubbard, P. (2006). Evaluating CALL software. In L. Ducate & N. Arnold (Eds.), *Calling on CALL: From theory and research to new directions in foreign language teaching* (pp. 313-318). San Marcos, TX: CALICO.

Hubbard, P. (2011). Evaluation of courseware and websites. In L. Ducate & N. Arnold (Eds.), *Present and future promises of CALL: From theory and research to new directions in foreign language teaching* (pp. 407-440). San Marcos, TX: CALICO.

Leakey, J. (2011). *Evaluating computer-assisted language learning: An integrated approach to effectiveness research in CALL*. Bern, Switzerland: Peter Lang Publishing.

Levy, M., & Stockwell, G. (2006). *CALL dimensions: Options and issues in computer-assisted language learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Machi, D. L. A., & McEvoy, B. T. (2008). *The literature review: Six steps to success*. Thousand Oaks, CA: Corwin Press.

Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.

Patton, M. Q. (2003). Utilization-focused evaluation. *International Handbook of Educational Evaluation, 9*(1), 223-244.

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

Patton, M. Q. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York: The Guilford Press.

Reeder, K., Heift, T., Roche, J., Tabyanian, S., Schlickau, S., & Golz, P. (2004). Toward a theory of e/valuation for second language learning media. In S. Fotos & C. Browne (Eds.), *New perspectives on CALL for second language classrooms* (pp. 255-278). Mahwah, NJ: Lawrence Erlbaum Associates.

Reeves, T. C., & Hedberg, J. G. (2003). *Interactive learning systems evaluation*. Englewood Cliffs, NJ: Educational Technology Publications, Inc.

Reeves, T. C., Herrington, J., & Oliver, R. (2005). Design research: A socially responsible approach to instructional technology research in higher education. *Journal of Computing in Higher Education, 16*(2), 96-115.

Richards, J. C., & Rodgers, T. (2012). Method: Approach, design, and procedure. *TESOL Quarterly, 16*(2), 153-168.

Sechrest, L., Babcock, J., & Smith, B. (1993). An invitation to methodological pluralism. *American Journal of Evaluation, 14*(3), 227-235.

Scriven, M. (1969). An introduction to meta-evaluation. *Educational Products Report, 2*(5), 36-38.

Scriven, M. (1972). Pros and cons about goal-free evaluation. *Evaluation Comment, 3*(4), 1-7.

Scriven, M. (1990). The evaluation of hardware and software. *Studies in Educational Evaluation, 16*(1), 3-40.

Sheldon, L. E. (1988). Evaluating ELT textbooks and materials. *ELT Journal, 42*(4), 237.

Skierso, A. (1991). Textbook selection and evaluation. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (pp. 432-453). Boston: Heinle & Heinle.

Squires, D., & McDougall, A. (1994). *Choosing and using educational software: a teachers' guide*. Philadelphia, Pennsylvania: Routledge.

Stake, R. E. (1975). *Program evaluation, particularly responsive evaluation*. Kalamazoo, MI: Western Michigan University Evaluation Center.

Stake, R. E. (2003). Responsive evaluation. In E. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 63-68). Boston, MA: Kluwer Academic Publishers.

Stake, R. E. (2004). *Standards-based and responsive evaluation*. Thousand Oaks, CA: Sage Publications, Inc.

Stake, R. E., & Schwandt, T. A. (2006). On discerning quality in evaluation. In I.F. Shaw, J.C. Greene, & M.M. Mark (Eds.), *The SAGE handbook of evaluation* (pp. 404-418). London: Sage.

Stufflebeam, D. L. (1974). Alternative approaches to educational evaluation: A self-study guide for educators. In J. Popham (Ed.), *Evaluation in education: Current applications* (pp. 97-143). Berkeley, CA: McCutchan Publishing.

Stufflebeam, D. L. (2003). The CIPP model for evaluation. *International Handbook of Educational Evaluation, 9*(1), 31-62.

Susser, B. (2001). A defense of checklists for courseware evaluation. *ReCALL, 13*(2), 261-276.

Tomlinson, B. (2003). Materials evaluation. In B. Tomlinson (Ed.), *Developing materials for language teaching* (pp. 15-36). Trowbridge, Wiltshire: Cromwell Press.

Villada, E. G. (2009). CALL Evaluation for early foreign language learning: A review of the literature and a framework for evaluation. *CALICO Journal, 26*(2), 363-389.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage Publications, Inc.