3-2018

Addressing Diversity in CALL Evaluation through Arguments and Theory-of-Action

Jim R. Ranalli
Iowa State University, jranalli@iastate.edu

Addressing Diversity in CALL Evaluation through Arguments and Theory-of-Action

Abstract
THE DIVERSE AND EVER-CHANGING list of technologies encompassed by computer-assisted language learning (CALL) presents evaluators with a challenging moving target. At a time when CALL can include everything from school-based telecollaborative projects to Massive Online Open Courses (MOOCs), to smartphone- and tablet-based apps, previous approaches to evaluation reveal their inadequacies. The checklists that once predominated (see Susser 2001) assumed the focus of evaluation to be tutorial software known as "courseware," which now constitutes a much-diminished part of the CALL landscape. Methodological frameworks like the one proposed by Hubbard (2006) assume the role of an instructor and a course in which the technology is situated, which is countered by increasingly autonomous and self-directed applications of CALL (Reinders and White 2016). And sets of criteria based in principles of interactionist second language acquisition (SLA; Chapelle 2001) seem less than ideal for evaluating technologies designed for individual use whose function is primarily facilitative of second language (L2) usage, such as online dictionaries or translation tools.

Disciplines
Educational Technology | English Language and Literature | Modern Languages | Online and Distance Education | Secondary Education

Comments
This book chapter is published as Ranalli, J. Addressing Diversity in CALL Evaluation through Arguments and Theory-of-Action. (2018) in *Useful Assessment and Evaluation in Language Education*, John McE. Davis, John M. Norris, Margaret E. Malone, Todd H. McKay, and Young-A Son, Editors. Georgetown University Press. Posted with permission.

Chapter 13

Addressing Diversity in CALL Evaluation through Arguments and Theory-of-Action

JIM RANALLI
Iowa State University

THE DIVERSE AND EVER-CHANGING list of technologies encompassed by computer-assisted language learning (CALL) presents evaluators with a challenging moving target. At a time when CALL can include everything from school-based telecollaborative projects to Massive Online Open Courses (MOOCs), to smartphone- and tablet-based apps, previous approaches to evaluation reveal their inadequacies. The checklists that once predominated (see Susser 2001) assumed the focus of evaluation to be tutorial software known as "courseware," which now constitutes a much-diminished part of the CALL landscape. Methodological frameworks like the one proposed by Hubbard (2006) assume the role of an instructor and a course in which the technology is situated, which is countered by increasingly autonomous and self-directed applications of CALL (Reinders and White 2016). And sets of criteria based in principles of interactionist second language acquisition (SLA; Chapelle 2001) seem less than ideal for evaluating technologies designed for individual use whose function is primarily facilitative of second language (L2) usage, such as online dictionaries or translation tools.

But the continually shifting and expanding nature of technological forms in L2 learning is only the most obvious type of diversity with which CALL evaluators must contend. We can see diversity in other aspects as well: the contexts in which CALL evaluations take place, the purposes for which they are conducted, the audiences for whom they are intended, the theoretical and methodological perspectives that inform their work, and the standards by which CALL interventions are deemed successful or not.

Consideration of the context of learning is essential in CALL evaluation because the value of technology for L2 learning can only be judged with reference to the needs and characteristics of particular participants in particular situations at

particular times. Meanwhile, the contexts of contemporary CALL usage evidence diversity in the extreme. They include classrooms across formal and informal educational settings, distance and online learning spaces, institutional self-access centers, and countless opportunities for individual, self-directed learning through technology. There is also diversity in the scales at which CALL contexts can be conceptualized. Gruba and Hinkleman (2012) discuss different concerns of evaluation at the micro (e.g., classroom), meso (e.g., institutional), and macro (e.g., policy) levels.

We also see diversity in the purposes for which CALL evaluations are conducted. While decision-making is an important function of evaluation—one for which checklists and methodological frameworks were designed—it is not the only purpose. Evaluation in our field has also contributed to the development of a base of knowledge about technology for language learning. While we address questions about whether a CALL intervention works, we are also interested in why and how it works and under what conditions its benefits may transfer to other users and contexts. Some CALL experts, such as Levy and Stockwell (2006), differentiate between *evaluation* and *research* to refer to these decision-making and knowledge-building functions, respectively, but, as argued by Chapelle (2001), information about why something works can also serve the purpose of decision-making. Importantly, evaluation can likewise inform the iterative development of CALL interventions and artifacts, which are rarely deemed beyond further improvement.

In addition, the audiences for CALL evaluations are diverse. Chapelle (2007) puts them on a continuum from *insiders* (e.g., software developers and other CALL researchers) on one end to *outsiders* (e.g., other applied linguists and program- or policy-level decision-makers) on the other, with *informed critics* (e.g., learners and instructors) in the middle. Other audience types exist as well; for example, in my own situation, a graduate program that trains future CALL professionals, members of thesis or dissertation committees are tasked with appraising student development and evaluation projects. The information needs of these different audiences, the types of evidence that will be most meaningful to them, and the aspects of the evaluation that best align with their concerns and values, will likewise vary.

There is also diversity in terms of the theoretical and research perspectives that CALL evaluators can draw on in conducting their investigations. An analysis of use of the term *theory* across twenty-five years of issues in the *CALICO Journal*, one of the field's major publications, identified 113 unique descriptors, although many of these represented theories within the fields of linguistics and education (Hubbard 2008). Given the necessary relationship between how phenomena are conceptualized and how they will be measured, it is thus not surprising to find diversity in the research approaches and methods employed by CALL researchers, not simply in terms of the qualitative-quantitative spectrum but also the starting points for investigations (e.g., exploring the use of a generic technology for its applicability to L2 learning versus identifying a problem in L2 instructional practice and developing a technology-based solution; Stockwell 2012). Emerging technologies also bring with them the potential for new ways to document and analyze variables of interest.

Finally, there is diversity in the standards by which the success of CALL interventions is evaluated. Looking at the multitude of published CALL studies addressing

the question "Does it work?" may give the impression that effectiveness is the only relevant criterion. While a new tool or task can produce gains in learning, it may yet go unadopted by instructors and learners, in which case it has in an important sense failed. Thus, sustainability of CALL interventions is arguably as important as effectiveness. The fact that CALL is rife with one-off projects (Kennedy and Levy 2009) must of course be attributed in part to the ever-changing nature of technology, but it is a well-known if not well-understood fact that CALL interventions shown to be effective may nevertheless remain stubbornly resistant to widespread adoption in L2 classrooms, such as the use of corpora (Ebrahimi and Faghih 2016).

Given all these forms of diversity, how should evaluation be defined in CALL, and how should it be approached? With respect to the first part of the question, Norris's definition of language program evaluation seems to encompass many if not all of the facets mentioned above: "[Language program evaluation is] a pragmatic mode of inquiry that illuminates the complex nature of language-related interventions of various kinds, the factors that foster or constrain them, and the consequences that ensue. [Evaluation] enables a variety of evidence-based decisions and actions, from designing programs and implementing practices to judging effectiveness and improving outcomes. [It] may provide a heuristic for generating new knowledge; raising awareness; and transforming the … circumstances of individuals and communities" (2016, 469). If we insert the term *technology-mediated* between *language-related* and *interventions*, substitute *CALL intervention* for *programs*, and exchange *stakeholder groups* for *communities*, we have what may be considered a working definition of evaluation in the field.

To complement this inclusive definition, could a single approach to evaluation likewise be found that is capable of embracing the diverse range of contexts, purposes, audiences, theoretical bases, and research approaches, in addition to the manifold forms of technology themselves? This question aligns with the concerns of the organizers of the 2016 Georgetown University Round Table on Languages and Linguistics (GURT) with making evaluation useful, as well as with the specific focus of the CALL evaluation colloquium at that meeting. In responding to these concerns, I wish to propose here that, to the extent a single approach is possible and desirable, it might be found in the idea of arguments.

What Do Arguments Have to Offer?

Conceptualizing evaluation as argument is not new. In the field of program evaluation, House asserts that evaluations are acts of persuasion aimed at "winning a particular audience to a point of view or course of action by an appeal to the audience's reason and understanding" (1977, 6). In introducing their approach to the evaluation of e-learning and distance education, Ruhe and Zumbo describe a "rigorous, evidence-based argument in support of evaluative claims" (2008, 11). In the field of CALL specifically, Chapelle, inspired by the work of Bachman (2005) on validity arguments, proposes that evaluation be conceptualized as "a situation-specific argument" (Chapelle 2001, 52) in which one marshals empirical evidence to show the extent to which a particular CALL task meets six key criteria derived from SLA

theory and research. Since the publication of this groundbreaking CALL evaluation framework, Chapelle has conducted pioneering work with validity arguments in the field of language testing (see Chapelle, Enright, and Jamieson 2008, 2010; Chapelle, Cotos, and Lee 2015) and has recently returned to the notion of CALL evaluation argument to elaborate its potential (Chapelle 2014, 2017). In the following sections, I review Chapelle's current proposal, as well as some argument-based validation research that intersects with CALL evaluation, to illustrate how arguments can be used to this end.

Implicit Arguments in Existing CALL Studies

In a plenary address at the EuroCALL conference at the University of Groningen, Chapelle (2014) proposed using arguments as a conceptual starting point for planning, conducting, and appraising CALL evaluations. In contrast to checklists, methodological frameworks, or SLA-derived sets of criteria, evaluation arguments begin with the *claims* one wants to make about a particular CALL intervention; that is, statements about the value of specific aspects of technology for language learning that are framed with reference to the needs and concerns of the particular audience(s) at whom the evaluation is directed. The evaluator's role is to make these claims explicit and "to plan an investigation that will determine the extent to which the claims can be supported. The results of the investigation are then used in support of an argument about the credibility of the claims" (Chapelle, 2017, 380).

For Chapelle, the lens of evaluation-as-argument can provide significant advantages to the way evaluations are planned, conducted, interpreted and appraised. In particular, it can help novice evaluators, such as graduate students evaluating CALL interventions as part of thesis or dissertation projects, avoid the trap of assuming that evaluation can only be accomplished by comparing between technology and nontechnology conditions. Rather than characterizing this as a new approach, Chapelle asserts that arguments are already evident in much published CALL research. Adopting a perspective of evaluation as argument allows them to be recognized and understood as such.

Chapelle's review of the professional literature in peer-reviewed CALL journals has identified five main types of evaluation argument. These are arguments based on: (1) comparisons, (2) corpus linguistics, (3) authenticity, (4) SLA theory, and (5) general pedagogical principles (Chapelle 2014, 2017). Briefly, comparison arguments are based on studies in which technology and nontechnology conditions are contrasted. Quantitative analysis of scores on outcome measures is typically used to determine if differences in L2 learning gains can be identified across conditions. The second type of argument, based on authenticity, focuses on common, authentic uses of technology for communication outside the classroom, using these as models and rationales for incorporating technology into formal language learning; for example, classroom tasks involving the use of smartphones or tablet computers. The third type of argument also deals with authenticity, but in this case it is with respect to the linguistic forms to which learners are exposed. In this type of argument, corpus linguistics techniques and data, in combined forms referred to as data-driven learning, are claimed to provide more authentic language samples, which in turn are seen to

benefit learning. Theory-based arguments constitute the fourth type in Chapelle's taxonomy. Such arguments make claims based on SLA theory about the value of having students engage in certain technology-mediated activities or conditions such as group discussion via text chat. The theory allows connections to be made between qualities of the students' engagement and interpretations about its value for language learning. The fifth and final type are arguments based on pedagogical principles that do not directly relate to enhanced language proficiency. Teachers' or CALL developers' interpretations of good pedagogical practice provide the rationales for having students engage in activities such as telecollaboration tasks, which are claimed to develop intercultural competence.

In addition to the advantages of evaluation-as-argument identified by Chapelle, there are others related to the diversity concerns outlined earlier. CALL interventions based on emerging technologies that push beyond the parameters of existing frameworks or evaluative criteria will always be expressible in terms of claims. And whereas Chapelle (2017) rightly sees the broad range of theories and research approaches encompassed in evaluation-as-argument as a challenge, it is also an affordance because it can capture so much of the diversity that CALL practitioners already find themselves working with.

However, there are also important limitations in this approach. To the extent that novice evaluators must draw from the professional journals for models, they may confront biased sampling, since evaluations in the professional literature whose primary purpose is local decision-making or ongoing evaluation may be relatively less common compared to those whose primary purpose is knowledge-building. Similarly, there will be bias in favor of effectiveness to the neglect of sustainability, and novice evaluators may likewise find insufficient guidance for shaping evaluations to the needs of audiences other than researchers communicating among themselves in peer-reviewed venues.

Another problem is the difficulty of appraising arguments embedded in a genre of communication that requires appraisal on its own terms, and which, by virtue of its complexity, may mask inadequacies in the argument's content or structure. It is a cognitively demanding task to produce and to interpret a research article, and writers' and readers' concerns will tend to default to the requirements of the genre. When all is said and done, a research article may boil down to the provision of support for a single claim, and while such an article may be judged favorably, the argument embedded within it may yet lack clarity and completeness, with important claims and assumptions left unelaborated or insufficiently developed.

Finally, even if a coherent and complete argument can be discerned in the text of a research article, the finished product provides no clue as to how the argument was developed. To borrow Bachman's description of the challenges of previous approaches to test validation, embedded CALL evaluation arguments may in essence represent groupings of "more or less independent qualities and questions, with no clear mechanism for integrating these into a set of procedures" (2005, 1). What may be helpful, then, is a way to delineate the components of a CALL evaluation argument in notation form, separately from the requirements of conducting and reporting research, and to facilitate more thorough identification of claims,

underlying assumptions, and the types of evidence that might be gathered to investigate them. We turn to such a procedure in the next section.

Explicit Arguments in CALL-Related Validation Studies

Validation is to assessment what evaluation is to instruction. Both are concerned with appraising the extent to which the qualities of a test or task can support the claims that are made about it. Like evaluation, validation is a complex process that must encompass broad and diverse sets of considerations. Contemporary language testers recognize the need to address traditional validation concerns, such as connecting test performance to reported scores, while also dealing with the consequences of test use, such as how scores are employed in decision-making. Arguments can support such a unitary approach to validation by showing how issues of test interpretation and use can be considered at the same time and by allowing diverse forms of validity evidence to be identified, prioritized, and appraised in relation to a larger whole. Because validation may involve a multitude of claims aimed at a variety of stakeholders, especially in the case of high-stakes assessments, explicitness is important. Explicitness can be accomplished by means of an explicit argument structure.

Language testers have drawn in particular upon the argument-based validation work of Kane (2012), who proposed a linear structure in which logical connections are represented by inferences named according to their place in the chain of reasoning. The *evaluation inference*, for example, links the actual observation of performance to its quantification in a test score, while the *utilization inference* links the score to some form of decision-making. Each inference is associated with claims regarding the particular assessment in question, and bridging the inferences requires the provision of support for the claims and their underlying assumptions. In Kane's (2013) approach, validation involves two stages: (1) an interpretation/use argument, in which the inferences, claims, assumptions, and relevant forms of support are specified; and (2) a validity argument, in which the evidence is gathered and then appraised in relation to the overall argument structure. According to Chapelle, the validity argument should take the form of "a narrative that points to a plausible conclusion" (2008, 319).

One of Kane's innovations was to undergird the validity-argument components with a model of inference developed by Toulmin (2003), which provides useful concepts and structures for teasing out logical relationships and anticipating counterarguments. In Toulmin's model, an observation, or "datum," is connected to a claim by an inference. What allows one to infer the claim from the datum is a warrant, a statement which is subject to challenge and which itself rests on one or more assumptions. Supporting the warrant involves the provision of backing for its underlying assumptions. Backing can take the form of empirical, theoretical, or commonsensical evidence. The model allows for conditions of rebuttal, which can also be backed by evidence and which may render the warrant inapplicable, thus undermining the inference. Importantly for the test-validation process, Kane (2012) says the model can provide guidance in the allocation of research effort. Claims can be prioritized for investigation according to how central they are to the interpretation-use argument or on the basis of assumptions considered especially problematic.

The potential of explicit arguments such as these to inform the development of CALL evaluation arguments is evident in recent validation studies involving technology-based formative assessments; that is, assessments that support learning and teaching. Chapelle, Cotos, and Lee (2015) present a validity argument for the use of an automated writing evaluation (AWE) tool to support classroom-based English-as-a-second-language (ESL) writing instruction by providing grammatical feedback and encouraging multiple drafts. The argument for one of these tools, the Criterion Online Writing Evaluation service, includes five inferences and their associated warrants, which are themselves based on twenty-two assumptions. The authors present empirical evidence related to the evaluation inference, which is based on the warrant that the AWE feedback "provides students with accurate information to target relevant areas for revision/improvement/learning," which in turn is based in part on the assumption that "*Criterion* feedback is accurate" (2015, 3). In reviewing their finding that more than half of Criterion's feedback had gone ignored by students, possibly as a result of inaccuracies in the feedback, the authors identified a potential rebuttal for future investigation, stating that students lack confidence in the AWE system. This study is useful first as an example of an explicit argument in which claims and assumptions regarding the value of technology for L2 learning are specified in detail and supporting evidence is gathered and appraised. It is also notable that, while other recent classroom-based AWE research has focused on issues of effectiveness, such as improvements in grammatical accuracy across drafts (e.g., Li, Link, and Hegelheimer 2015), the argument in Chapelle, Cotos, and Lee (2015) directed the authors' attention to sustainability concerns.

A follow-on investigation to this study conducted by myself and two colleagues (Ranalli, Link, and Chukharev-Hudilainen 2016) shows how explicit arguments can serve the purpose of decision-making and help in communicating among stakeholders. I undertook this study in my dual capacity as both a researcher and a coordinator for the ESL writing program in which the Chapelle, Cotos, and Lee (2015) study was conducted. In my latter capacity, I sought help in deciding whether Criterion should continue to be used in the course I oversaw. Our research team investigated assumptions underlying the evaluation and utilization inferences in the Chapelle, Cotos, and Lee (2015) argument and found that students in a lower-level course were better able to make use of the Criterion feedback in correcting errors than their higher-level counterparts, which led to our recommendation that Criterion should be used in the lower- but not the higher-level course. In addition to facilitating this finding, the explicit argument also helped the research and course coordination teams clarify expectations among themselves regarding the extent to which the major writing assignments were intended to support L2 development versus the development of writing expertise, a key consideration that until then had gone unrecognized.

Another recent study straddling argument-based validation and CALL evaluation is Gleason (2013), which focuses on blended learning in college-level Spanish classes. Practitioners of blended learning seek a middle ground between completely online and completely face-to-face language instruction, guided by the question "Which tasks are best delivered in which format?" Gleason develops an

interpretation-use argument and then reviews evidence related to the evaluation inference to determine whether students' engagement in tasks across face-to-face and online conditions demonstrate comparable learning opportunities, with the learning tasks conceptualized as "micro-formative assessments" (2013, 3). The evidence consists of discourse analysis of language produced by students during the tasks, which provides backing for some of the comparability claims but not others; among the latter is the assumption that the online condition affords equivalent chances for learners to spontaneously focus on meaning. Gleason discusses how her findings could support the work of blended-learning designers, thus illustrating how arguments can be used as the basis of formative evaluations informing ongoing CALL development.

Explicit interpretation-use arguments, then, as this very brief review suggests, facilitate greater detail and completeness in the specification of claims and assumptions, helping evaluators identify priority areas for investigation and allowing appraisal of diverse forms of evidence in relation to the larger argument. They expand the scope of concerns beyond mere effectiveness of the intervention and can feed valuable information back into the development process. While making an argument explicit is no guarantee that all stakeholders will be appeased (Bachman 2005), it does increase transparency and the likelihood of more viewpoints being considered. Given the increasing interest in integrating instruction and assessment, as evidenced in many talks at the GURT 2016 meeting, and the ways that technology is allowing new and more powerful means of simultaneously scaffolding and assessing learning, it seems likely that validity arguments involving CALL used for formative assessment will become more common in the future.

One might suppose, then, that validity arguments and CALL evaluation arguments will largely overlap, except insofar as the object of their focus will be, in the case of the former, an assessment, and in the latter, a CALL intervention. There are, however, limitations to the value of applying the Kanean framework to CALL evaluation. Firstly, the validation argument structure will require that a CALL intervention be construable as an assessment, which will not always be possible nor desirable. And while the linear chain of reasoning connecting observations to scores to uses is well suited to language testing, it may not be a good fit for CALL interventions that entail a greater diversity of claims in more complex causal networks. What is needed, then, is a similar method of making evaluation arguments explicit while affording greater flexibility in delineating relationships among components. In the next section, I discuss the potential of Theory of Action (ToA) models to fulfill this purpose.

Theory of Action As a Way Forward

ToA is a concept from the field of program evaluation whose origins lie in efforts to make social programs more goal-oriented and thus likely to succeed. ToA and related approaches such as Logic Modeling and Theory of Change were developed in part to test the readiness of programs to be evaluated by conceptualizing them as identifiable sets of activities and inputs linked to specific outcomes, with these linkages being both logical and testable (Patton 2008). According to Patton, logic

models connect outcomes but do not necessarily specify the causal factors that underlie them. When causal mechanisms are added to a program's logic model, it becomes a Theory of Change. If the scale of evaluation is narrower than a complete program or policy intervention, such as a particular strategy to address a specific problem within a specific time frame (e.g., the "action"), the term Theory of Action is used (Patton 2008, 339).

Procedurally, ToA starts with the specification of the intended long-term outcomes, from which program designers and developers then work backward toward shorter term outcomes, outputs, inputs, and so on that constitute preconditions. A simplification of the theory in notation or diagram form helps stakeholders to develop a theory of action for themselves. As for who is responsible for elaborating a program's ToA, some theorists emphasize the role of program staff and intended beneficiaries while others see the need for social scientists' knowledge and expertise. Another approach is to involve both, with practitioners able to contribute knowledge of how a ToA model gets translated into reality while researchers can complement the model's intended outcomes with other, unintended outcomes that previous research or theory suggest may be likely to occur.

Part of the value of this approach is that, in delineating a ToA model, assumptions underlying the causal linkages among inputs, outputs, and intended outcomes are more readily identifiable. Evaluation theorists have termed these "validity assumptions" (Patton 2008; Suchman 1967). In social programs, a common validity assumption is that acquiring relevant knowledge will lead to beneficial changes in behavior, despite much research and evaluation casting doubt on the universal applicability of this assumption (Weiss 2000, cited in Patton 2008). This is relevant to the situation in CALL, where one-off projects are rife and instructors often fail to buy into technological innovations despite research evidence.

Thus, we can see potential productive overlap between ToA and an argument-based approach to evaluation for use in CALL as outlined above. There are clear parallels between the focus on intended outcomes in the former and claims in the latter, and both stress the need to tease out underlying assumptions. ToA is also agnostic in terms of data types and research methodologies. It lends itself to linear chains of reasoning resembling those found in the validity argument, and yet it can be used to model more complex adaptive systems in which a medium-term outcome has more than one cause or serves as a precondition for more than one long-term outcome.

Kane (chapter 14, this volume) discusses the importance of using ToA in conjunction with argument-based validation when assessments are formative, since the outputs from such assessments serve as inputs to instruction, and these inputs have intended outcomes of their own. The only project in applied linguistics to have combined the argument-based validation and ToA approaches so far is the CBAL project coordinated by Educational Testing Service. CBAL, which stands for Cognitively Based Assessment of, for, and as Learning, is a research initiative to develop a comprehensive kindergarten through twelfth grade assessment system integrated into classroom instruction that provides students and teachers with worthwhile educational experiences. The project, which focuses on English language arts and makes substantial use of computer-based materials, has been described in a report detailing

how evidence for the ToA is gathered and appraised in conjunction with a validity argument. According to Bennett (2010, 71), the CBAL ToA includes

1. the intended effects of the assessment system;
2. the components of the assessment system and a logical and coherent rationale for each component, including backing for that rationale in research and theory;
3. the interpretive claims that will be made from assessment results;
4. the action mechanisms designed to cause the intended effects; and
5. potential unintended negative effects and what will be done to mitigate them.

The first, second, and fourth elements are recognizable as conventional parts of ToA or logic modeling, although the term *effects* has been substituted for *outcomes* and the term *components* for *inputs* and *outputs*. The third and fifth elements, meanwhile, are more clearly associated with validity arguments (i.e., interpretive claims about consequences and uses of assessments and unintended effects that could constitute rebuttals to the claims). According to Bennett (pers. comm. with the author, May 19, 2016), interpretation-use arguments in general, and the Toulmin model of inference in particular, can naturally complement ToA because the former's focus on warrants, assumptions, and rebuttals can help elaborate the explicit and implicit claims in a ToA model or the evidence needed to elaborate those claims.

While the CBAL example is instructive for language testers working with formative assessments, a more CALL-focused illustration will be helpful for present purposes. In the following section, I set forth an example of a CALL evaluation conceptualized using ToA and the Toulmin inference structure.

An Example

As part of my doctoral work, I developed and evaluated a web-based, technology-mediated course in strategy instruction (SI). The evaluation, which is described in Ranalli (2013), did not include an explicit argument, but I have reconceptualized it in such a form here to illustrate some of its shortcomings and to show how these might have been improved upon if ToA and Toulmin's model of inference had been used as the basis of the evaluation.

Background

The SI course, which was called Virtual Vocabulary Trainer or VVT, was created to teach college-level ESL writing students an integrated form of dictionary skills and language awareness of features of pattern grammar (Hunston and Francis 2000). Pattern grammar encompasses a variety of different ways that lexical words can combine with other lexical words as well as function words; in particular, verb transitivity, complementation, and grammatical collocation. Although providing students of English with information about syntactic patterning is a primary purpose of learner dictionaries, such dictionaries are often misused or underused for a variety of reasons, including lack of understanding about syntactic features of English vocabulary (Ranalli and Nurmukhadev 2014).

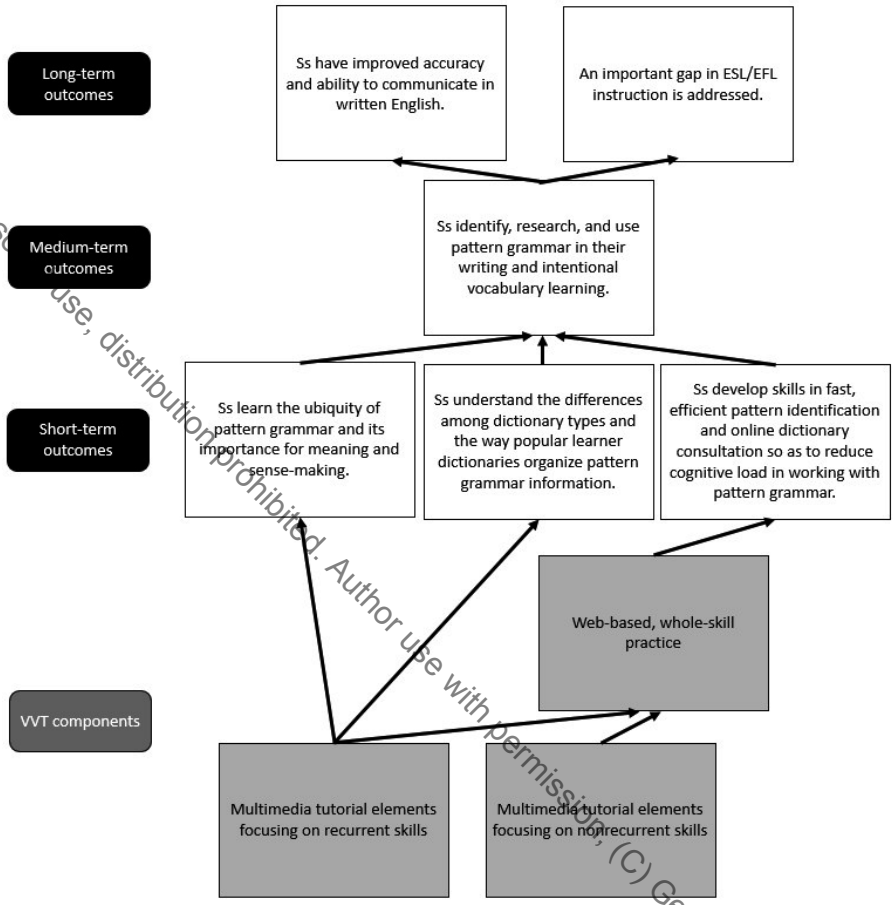
The online course comprised multimedia tutorials consisting of videos and accompanying text-based exercises. The tutorials aimed at developing students' declarative knowledge about learner dictionaries and about pattern grammar—its wide variety of forms, its frequency in English, and the way it helps determine which sense of a word is intended—as well as procedural knowledge in terms of students' abilities at syntactic parsing to identify potential patterns and to perform related searches in learner dictionaries quickly and efficiently so as to minimize cognitive load. An instructional design framework for training complex cognitive skills (Van Merriënboer 1997) was used to differentiate these aims, with declarative aims addressed under the name *non-recurrent* skills, procedural knowledge addressed under the name *recurrent* skills, and the integration of these subcomponents termed *whole-skill practice*.

The original evaluation was based largely on a between-groups experimental design contrasting a VVT condition with a comparison condition that involved learners in repeated dictionary consultations for usage information but no instruction. Participants were assigned randomly to one of the two conditions, which were administered online through a learning management system (LMS). An online task used as a pre- and posttest measure of strategy performance required participants to correct pattern-grammar errors in sentences and to choose among a selection of online dictionaries to assist them in doing so. The results showed large effect sizes for between-groups differences at posttest and within-group differences for the VVT group from pre- to posttest. In addition, user perception data, which was collected via online questionnaires, showed generally positive views of the VVT materials and a majority of participants indicating they would use what they had learned beyond the course.

In the semesters following the original evaluation, the VVT course was used sporadically. It was made available to any interested instructor of the writing course, some of whom chose to use it while others did not. The requirement that students access the course via a separate LMS from that used for the writing course made integration of the materials difficult. Another problem was that instructors who were unfamiliar with pattern grammar expressed uncertainty about the aims of the course, assuming it focused on teaching specific lexical patterns, and were reluctant to better familiarize themselves by completing the course on their own, no doubt because of the investment of time involved. Many opted instead to use a paper-based dictionary consultation assignment that had originally constituted the vocabulary component of the course. The VVT materials, now in need of design and functionality upgrades, are not currently hosted or available online.

Reconceptualizing the Evaluation As a ToA Argument

The original evaluation, then, focused narrowly on effectiveness, with little concern for the needs and views of a key stakeholder group, instructors, and thereby failed to address some of the contextual issues that would influence integration and sustainability. As such, the project may be representative of many CALL interventions that show promise but fail to take root. To illustrate how a ToA argument can work to inform the process of CALL evaluation (as well as iterative development), we can first reimagine the original project as an initial ToA model (figure 13.1).



◆ Figure 13.1. Initial ToA model for the VVT online strategy instruction project

Development of the model starts near the top with specification of the long-term outcomes (cf. use of the term *effects* in Bennet 2010). Outcomes represent "changes in awareness, knowledge, skill, or behavior" (Knowlton and Phillips 2009, 8); in the present case, students will have improved accuracy and communication of lexical meaning in their writing and vocabulary use, and an important gap in ESL and EFL (English as a foreign language) instruction will have been addressed.

Working backward from these ultimate intended outcomes, a medium-term outcome is specified: that students identify, research, and use pattern grammar in their writing and intentional vocabulary learning—in other words, that they transfer application of the strategy to new contexts of use where it is also relevant. This single outcome is then shown to be dependent on three short-term outcomes that constitute preconditions focusing on the relevant declarative and procedural knowledge described above. Below these, the SI components represent the inputs (i.e., the differentiated forms of instruction) and outputs (i.e., descriptive indicators of

what activities the inputs generate; Knowlton and Phillips 2009) that are claimed to result in the short-term outcomes. Together, the elements resemble a logic model, albeit a high-level specification of one for demonstration purposes that could be broken down into a more detailed representation to address specific evaluation needs.

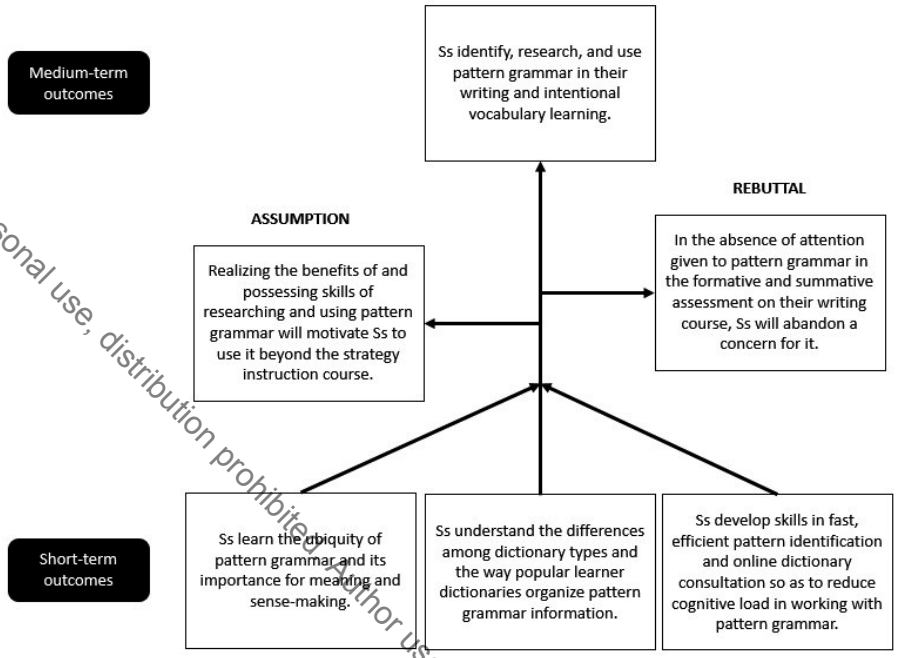
The next steps in developing the ToA model would be (1) to specify any assumptions underlying the causal linkages between project components and outcomes in the model, and (2) to specify forms of empirical or theoretical support that would render these causal inferences warranted. The ToA model would then be appraised for plausibility, coherence, and completeness. The outcomes (i.e., claims), assumptions, and backing could at this point be reviewed by the evaluator to determine which parts of the model are most in need of empirical support. It is interesting to consider where data from the original evaluation study would fit into this model. Scores from the pre- and posttest could be used as support for the inferences linking the VVT components to the short-term outcomes, while the user-perception data could be used as backing for the linkages from one or more of the short-term outcomes to the medium-term outcomes—although in the latter case, this would constitute very limited support, a fact which the model would make more evident.

The process would not stop here, however. The next step in developing the ToA model would be to share it with key stakeholders including, in this case, thesis committee members as well as instructors in the writing course where the study was based. One immediate advantage of using the model for this purpose instead of a research proposal, pilot study report, or other lengthier verbal description, might be to better facilitate understanding that the focus of the intervention is a complex cognitive skill involving use of a certain type of lexical feature rather than direct improvements of L2 proficiency, which was a very common misunderstanding about the project.

More importantly, however, the lack of attention to sustainability issues may have been immediately evident. Instructors, in talking through the model with the researcher, might have raised questions about the causal connection between the short- and medium-term outcomes, noting that it rests on the assumption that possession of new knowledge and skills will necessarily entail students' using these in new contexts of use. Such observations could constitute a potential rebuttal (figure 13.2) stating that the lack of links to pattern grammar in the summative and formative assessment on the writing course would mean that students would be likely to abandon a concern for it beyond the VVT training.

Types of appropriate backing could then be specified for the assumption and rebuttal, yielding evaluation questions for possible investigation. Alternatively, in recognition of the fact that much previous research favors the rebuttal over the assumption, the model could be elaborated to address the lack of attention to integration and sustainability in the project.

Such elaborations are depicted in the revised model (figure 13.3). A new set of short-, medium-, and long-term outcomes is specified that would (hypothetically) address the needs of instructors. In addition, components of the VVT project have been specified at the bottom, consisting of two workshops and integration of the VVT course into the LMS used by the course instructors to facilitate ongoing



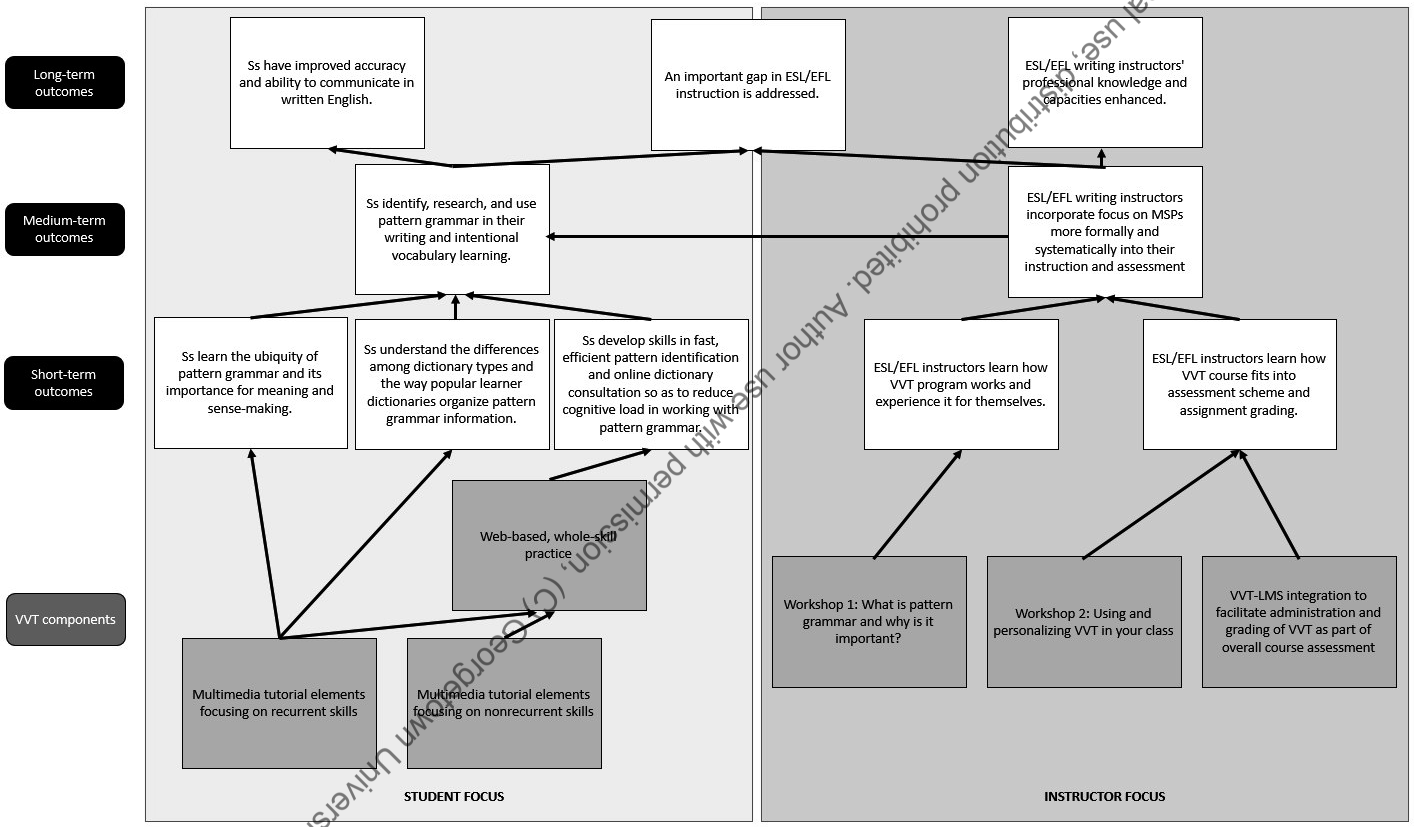
◆ Figure 13.2. Inference in the VVT ToA model linking three short-term outcomes to a medium-term outcome, with an assumption on which that inference depends and a potential rebuttal to the inference

administration and grading. Short-term outcomes focus on important forms of pattern grammar–related knowledge that instructors need to possess; medium-term outcomes address the actions intended to result from that knowledge; and long-term outcomes constitute worthwhile professional development goals in their own right, which are distinct from the original goals of the project.

The next step in the process would be to specify assumptions underlying the causal links in this new section of the model and identify forms of evidence that could be used to support these additional claims. Following this, the evaluation would proceed by gathering and assembling these forms of evidence according to priorities established in consultation with stakeholders. The final step would be to compose the evaluation argument itself; that is, a single narrative in which the evidence is reviewed for all specified claims to determine the extent to which they are supported and in which the coherence and completeness of the ToA argument structure is also assessed. Exemplification of these next steps is beyond the scope of the present chapter, but it is hoped this brief illustration provides a glimpse into the ways a ToA argument can structure and enhance the process of CALL evaluation.

Summary and Conclusion

This chapter has discussed the potential benefits of arguments to support the task of CALL evaluation in all its diversity of focuses, purposes, contexts, and audiences.



◆ Figure 13.3. The revised ToA model for the VVT project, with new intervention components and outcomes targeted at writing-course instructors

Compared to checklists, methodological frameworks, and sets of evaluative criteria, arguments support a broader and more comprehensive approach to evaluation by focusing on the claims one wants to make about a particular CALL intervention and helping clarify assumptions underlying those claims that might otherwise go unrecognized or underappreciated.

Beyond a general recommendation for use of arguments, this chapter has proposed a particular approach based on the notion of Theory of Action. ToA is a participatory approach that can complement social-science expertise with the vital perspectives of stakeholders and thus highlight issues of sustainability as well as effectiveness. It supports decision-making by requiring the contextual characteristics of an intervention to be incorporated into modeling but can also serve purposes of knowledge building and iterative development. It affords attention to the needs and values of particular audiences and allows diverse theoretical and research perspectives to be utilized in principled ways. Used in conjunction with the Toulmin model of inference, ToA arguments may potentially engender a new approach to CALL evaluation similarly to the way interpretation-use arguments enabled a new approach to language-test validation. They provide a flexible yet powerful methodology for understanding how and why CALL can support the needs of learners and instructors as well as mechanisms for accountability and the guidance of research. Importantly, they align with current views of technology mediation in language education, in which the main question posed is no longer "Is it effective?" but rather "Under what conditions, and for whom?" (Chun 2016).

This potential needs to be developed. Among other things, there is a need to adapt ToA to better accommodate aspects of arguments. Differences between the way one frames claims versus intended outcomes may be subtle but important. ToA typically consolidates assumptions in one area of a model, whereas the Toulmin approach entails modeling assumptions for individual claims. This, along with the potential for large and complex arrangements of claims depending on the CALL intervention in question, will present challenges to graphical depiction. To identify and address such challenges, actual evaluations must be conducted, including developmental evaluations undertaken in the early stages of CALL projects so that reciprocities between development and evaluation can be explored.

References

Bachman, Lyle F. 2005. "Building and Supporting a Case for Test Use." *Language Assessment Quarterly* 2 (1): 1–34.

Bennett, Randall E. 2010. "Cognitively Based Assessment of, for, and as Learning (CBAL): A Preliminary Theory of Action for Summative and Formative Assessment." *Measurement* 8:70–91.

Chapelle, Carol A. 2001. *Computer Applications in Second Language Acquisition: Foundations for Teaching, Testing, and Research*. Cambridge: Cambridge University Press.

———. 2007. "Challenges in Evaluation of Innovation: Observations from Technology Research." *Innovation in Language Learning and Teaching* 1 (1): 30–45.

———. 2008. "The TOEFL Validity Argument." In Chapelle, Enright, and Jamieson, *Building a Validity Argument*, 319–52.

———. 2014. "Arguments for Technology and Language Learning." Plenary address at the annual EuroCALL Conference, Groningen, Netherlands, August.
———. 2017. "Evaluation of Technology and Language Learning." In *The Handbook of Technology in Second Language Teaching and Learning*, edited by Carol A. Chapelle and Shannon Sauro. Malden, MA: Wiley-Blackwell, 378–392.
Chapelle, Carol A., Elena Cotos, and Jooyoung Lee. 2015. "Validity Arguments for Diagnostic Assessment Using Automated Writing Evaluation." *Language Testing* 32 (3): 385–405.
Chapelle, Carol A., Mary K. Enright, and Joan M. Jamieson, eds. 2008. *Building a Validity Argument for the Test of English as a Foreign Language*. New York: Routledge.
———. 2010. "Does an Argument-Based Approach to Validity Make a Difference?" *Educational Measurement: Issues and Practice* 29 (1): 3–13.
Chun, Dorothy M. 2016. "The Role of Technology in SLA Research." *Language Learning and Technology* 20 (2): 98–115.
Ebrahimi, Alice, and Esmail Faghih. 2016. "Integrating Corpus Linguistics into Online Language Teacher Education Programs." *ReCALL* 29 (1): 120–35.
Gleason, Jesse B. 2013. "An Interpretive Argument for Blended Course Design." *Foreign Language Annals* 46 (4): 588–609.
Gruba, Paul, and Don Hinkelman. 2012. *Blending Technologies in Second Language Classrooms*. New York: Palgrave Macmillan.
House, Ernest R. 1977. *The Logic of Evaluative Argument*. CSE Monograph Series in Evaluation 7. Los Angeles: Center for the Study of Evaluation.
Hubbard, Philip. 2006. "Evaluating Call Software." In *Calling on CALL: From Theory and Research to New Directions in Foreign Language Teaching*, edited by Lara Ducate and Nike Arnold, 313–38. San Marcos, TX: CALICO.
———. 2008. "Twenty-Five Years of Theory in the CALICO Journal." *CALICO Journal* 25 (3): 387–99.
Hunston, Susan, and Gill Francis. 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.
Kane, Michael. T. 2012. "Validating Score Interpretations and Uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010." *Language Testing* 29 (1): 3–17.
———. 2013. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50 (1): 1–73.
Kennedy, Claire, and Mike Levy. 2009. "Sustainability and Computer-Assisted Language Learning: Factors for Success in a Context of Change." *Computer Assisted Language Learning* 22 (5): 445–63.
Knowlton, Lisa W., and Cynthia C. Phillips. 2009. *The Logic Model Guidebook: Better Strategies for Great Results*. Los Angeles: Sage Publications.
Levy, Mike, and Glenn Stockwell. 2006. *CALL Dimensions: Options and Issues in Computer-Assisted Language Learning*. Mahwah, NJ: Lawrence Erlbaum.
Li, Jinrong, Stephanie Link, and Volker Hegelheimer. 2015. "Rethinking the Role of Automated Writing Evaluation (AWE) Feedback in ESL Writing Instruction." *Journal of Second Language Writing* 27:1–18.
Norris, John M. 2016. "Language Program Evaluation." *Modern Language Journal* 100 (S1): 169–89.
Patton, Michael Q. 2008. *Utilization-Focused Evaluation*. 4th ed. Thousand Oaks, CA: Sage Publications.
Ranalli, Jim. 2013. "Online Strategy Instruction for Integrating Dictionary Skills and Language Awareness." *Language Learning & Technology* 17 (2): 75–99.
Ranalli, Jim, Stephanie Link, and Evgeny Chukharev-Hudilainen. 2016. "Automated Writing Evaluation for Formative Assessment of Second Language Writing: Investigating the Accuracy and Usefulness of Feedback as Part of Argument-Based Validation." *Educational Psychology* 37 (1): 8–25.
Ranalli, Jim, and Ulugbek Nurmukhadev. 2014. "Learner Dictionaries." In *The Encyclopedia of Applied Linguistics*, edited by Carol A. Chapelle, 1–6. Malden, MA: Wiley-Blackwell.
Reinders, Hayo and Cynthia White. 2016. "20 Years of Autonomy and Technology: How Far Have We Come and Where to Next?" *Language Learning and Technology* 20 (2): 143–54.
Ruhe, Valerie, and Bruno D. Zumbo. 2008. *Evaluation in Distance Education and E-Learning: The Unfolding Model*. New York: Guilford Press.

Stockwell, Glenn. 2012. "Diversity in Research and Practice." In *Computer-Assisted Language Learning: Diversity in Research and Practice*, edited by Glenn Stockwell, 147–63. Cambridge: Cambridge University Press.

Suchman, Edward A. 1967. *Evaluative Research: Principles and Practice in Public Service and Social Action Programs*. New York: Russell Sage Foundation.

Susser, Bernard. 2001. "A Defense of Checklists for Courseware Evaluation." *ReCALL* 13 (2): 261–76.

Toulmin, Stephen. 2003. *The Uses of Argument*. Updated ed. Cambridge: Cambridge University Press.

Van Merriënboer, Jeroen J.G. 1997. *Training Complex Cognitive Skills: A Four-Component Instructional Design Model for Technical Training*. Englewood Cliffs, NJ: Educational Technology Publications.

Weiss, Carol Hirschon. 2000. "Which Links in Which Theories Shall We Evaluate?" *New Directions for Evaluation* 2000 (87): 35–45.