TESOL QUARTERLY Vol. 28, No. 4, Winter 1994

Criterion-Referenced Language Test Development: Linking Curricula, Teachers, and Tests

BRIAN K. LYNCH
University of Melbourne

FRED DAVIDSON
University of Illinois, Urbana-Champaign

In discussing the use of both criterion-referenced measurement and norm-referenced measurement techniques for item analysis, Brown (1989) has called for strengthening the relationship between testing and the curriculum. Alderson and Wall (1993) have pointed out the need for actual studies on the existence of *washback*, or the influence of tests on teaching. This article answers those calls by presenting criterion-referenced language test development (CRLTD) as a means for linking ESL curricula, teacher experience, and language tests.

CRLTD focuses on the generation of *test specifications*, as adapted from Popham (1978), and their refinement following the production of items or tasks from those specifications. Sample specifications are presented from university ESL/EFL programs at the University of Illinois, Urbana-Champaign, and the University of California, Los Angeles. CRLTD is elaborated further in the form of a workshop designed to translate curricular goals into test instruments with the active participation of teachers. The article concludes by examining data from teachers who have used CRLTD and with a discussion of its benefits as a proactive process for teaching and assessment.

Several decades ago, a dichotomy was introduced to the educational measurement literature: *norm-referenced measurement* (NRM) and *criterion-referenced measurement* (CRM). Glaser (1963) is generally credited with making the distinction and coining the term CRM, and Popham (1978, 1981) has discussed both approaches. The language-testing literature has evidenced a growing attention to the NRM/CRM dichotomy (Bachman, 1990; Brown, 1989; Cartier, 1968; Cook, 1992; Cziko, 1982; Hudson, 1989; Hudson & Lynch, 1985; Hughes, 1989). Discussions have focused primarily on the differences in the measure-

ment philosophy of NRM and CRM and on their associated statistical techniques for item analysis, reliability, and validity.

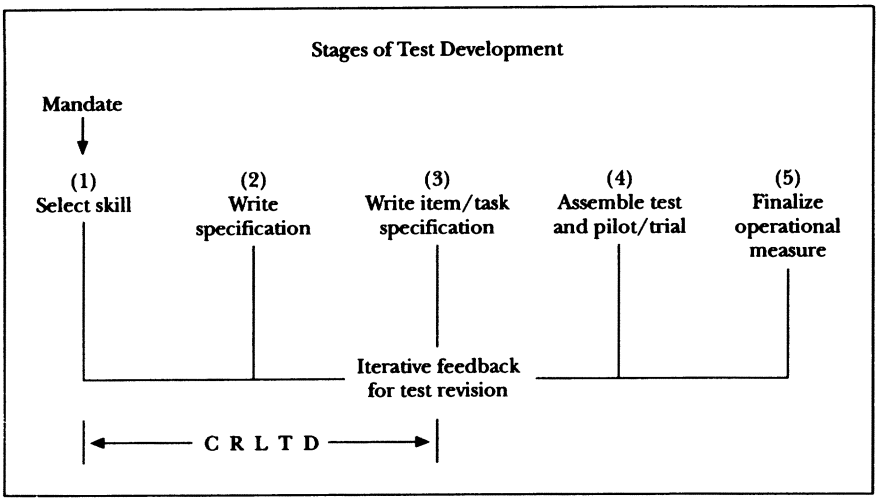
In this article we focus on the potential benefits of using CRM in the initial stages of a larger process of test development. As such, we comment on the use of this measurement approach in the linking of ESL curricula, teacher experience, and language tests. We also demonstrate the potential for the positive influence of testing on language teaching, traditionally referred to as *backwash*, or *washback*. The existence of washback has recently been called into question by Alderson and Wall (1993). These same authors have followed with a study that finds some evidence of washback on the content of teaching but not on the methods used (Wall & Alderson, 1993).

This article presents an approach called *criterion-referenced language test development* (CRLTD). We define CRLTD as the development of test specifications (i.e., test blueprints) and test items or tasks through a process that works back and forth between the specification and the item to refine the notion of the *criterion*, or what is being tested. Whereas the notion of criterion referencing can affect all stages of test development, CRLTD focuses on the earlier stages, as depicted in Figure 1. (The *mandate* refers to contextual requirements impinging on test design; we discuss this further below.) This focus on the link between a test plan and the test instrument should also yield the kind of teaching/testing evidence necessary to prove the existence of washback. In particular, we present evidence of what we term *reverse washback*—the influence of language teaching on language tests—with that evidence becoming formalized in the process of CRLTD.

First, we define the terms NRM and CRM as used in this article. Rather than a dichotomy, it is perhaps best to think of these terms as representing a continuum: Most tests can have characteristics of both measurement approaches and can best be described as tending to one end of the continuum or the other. NRM refers to the measurement approach that is concerned with determining the relative standing, or rank order, of examinees. From this perspective, we are concerned with relative decisions—for example, achieving a clear spread of students' scores on a test so that we know that student A is 2 points better than student B, who is 2 points better than student C. Often teachers are called on to make just this sort of relative decision. For example, a school district may award a limited number of scholarships each year to the top 10 students. From year to year, the overall ability of the students may vary somewhat; however, the district wants to award the top 10 with scholarships each year whether they are better or worse than the previous year's top 10 students. The decision is relative to the particular group of students each year not to some absolute criterion of academic success or potential.

FIGURE 1

The Relationship of CRLTD to the Stages of Test Development



CRM refers to the measurement approach that is concerned with determining the absolute standing of examinees in reference to a specific ability or behavior. The criterion is this ability or behavior. CRM, then, calls for an absolute decision—absolute not in the sense of being perfect or without doubt but in the sense of the examinee's score being tied to a specifically defined criterion rather than to the performance of others taking the test. ESL teachers encounter such a decision context at the end of every instructional term. Teachers want to be able to determine whether each student has mastered the course material sufficiently, often in order to determine whether or not they should pass that student on to the next level of instruction. In this case, teachers are not interested in who the top 5, 10, or 20 students are. The fact that students vary in their ability from term to term is of concern to the teachers, and they cannot assume that a certain percentage of their students will always have mastered the course material by the end of the term. Their decision is in reference to the criterion—the course objectives—not to the rank ordering of students on the test.

It is possible to integrate NRM and CRM approaches, as Brown (1989) and Cook (1992) have demonstrated. However, the usefulness of CRM for teaching rests in the degree to which the behavior or ability being tested is clearly defined. This is not to say that norm-referenced tests (NRTs) will make no effort to clearly specify what is

being tested, or that criterion-referenced tests (CRTs) will always succeed in doing so. However, CRTs can be distinguished from NRTs in part by the tendency toward greater detail in their test specifications. Specifications at the NRM end of the continuum often do nothing more than label their items as measuring something like reading comprehension and specify the number of such items on the test—in part due to the NRM assumption that item statistics will be used later in the test development process to ensure test quality. We discuss this assumption further in the section on the CRLTD workshop.

In contrast, the CRT specification will define in detail what skills and abilities are intended as the object of measurement and will operationalize those skills in the test format. Although clearly defined constructs are necessary for any valid test, be it a NRT or CRT, we argue that test specifications at the level of detail we propose are more likely to be associated with the latter. For this reason we refer to our approach as *criterion referenced*—we consciously focus on the test specification and item-or task-writing stages of test development as a means of clarifying the criterion being tested. This clarification is the result of the iterative nature of CRLTD: the experience from the item-writing stage feeding back to the elaboration of the test specification. In this way, the test specification also provides a detailed record of evidence for judging how well the test items or tasks match what the test claims to be measuring.

THE CRLTD PROCESS

The process of CRLTD involves, first and foremost, the creation of a detailed test specification. It arrives at this specification through a series of steps presented at the end of this section. The format we have developed for test specifications (Davidson & Lynch, 1993) essentially follows that of Popham (1978, 1981) and is represented in generic form in Figure 2.

The Specification Number, Title, and Related Specifications are meant to aid the test developer in keeping the test development process organized. This addition to the general format came from our experience with CRT specifications in various workshops and in the development of the revised University of California, Los Angeles (UCLA), English as a Second Language Placement Exam (ESLPE); the revised University of Illinois, Urbana-Champaign (UIUC), ESL Placement Test; and the General Tests of English Language Proficiency (GTELP) exam (see Hudson, 1989). Note that we have also adopted the term *prompt attributes*, as used in Brown, Detmar, and Hudson (1992) rather than Popham's (1978) *stimulus attributes* to avoid confusion with stimu-

FIGURE 2
Components of a CRM Specification

Specification Number: an index number

Title of Specification: a short title that generally characterizes each specification.
The title is a good way to outline skills across several specifications.

Related Specification(s): the numbers and/or titles of specifications related to this one, if any.
For example, in a reading test separate detailed specifications would be given for the passage and for each item.

General Description (GD): a brief general statement of the behavior to be tested.
The GD is very similar to the core of a learning objective. The purpose of testing this skill may also be stated in the GD. The wording of this does not need to follow strict instructional objective guidelines.

Prompt Attributes (PA): a complete and detailed description of what the student will encounter.

Response Attributes (RA): a complete and detailed description of the way the student will provide the answer, that is, a complete and detailed description of what the student will do in response to the prompt and what will constitute a failure or success.
There are two types of RAs:
a. selected response: a clear and detailed description of each choice in a multiple-choice format
b. constructed response: a clear and detailed description of the type of response the student will generate, including the criteria for evaluating or rating the response.

Sample Item (SI): an illustrative item or task that reflects the specification, that is, the sort of item or task the specification should generate.

Specification Supplement (SS): a detailed explanation of any additional information needed to construct items for a given specification.
In grammar tests, for example, it is often necessary to specify the precise grammar forms tested. In a vocabulary specification, a list of testable words might be given. A reading specification might list in its supplement the textbooks from which reading test passages may be drawn.

lus-response as expressed in behaviorist theories of learning. Finally, under Response Attributes, selected response (a), the description of the alternatives in a multiple-choice format are called required. Because these alternatives are a part of what the student will encounter, they could logically be classified under the Prompt Attributes instead. However, because they represent what the student must sort through and select from in order to answer, we have included them as a part of the Response Attributes. Furthermore, as the following sample specifications will demonstrate, this generic specification format is not meant to be rigidly fixed. Depending on what is being tested and the context in which the specification is developed, the order and characterization of the components can vary.

A well-written specification should result in a document that will enable similarly trained teachers working in a similarly constituted

teaching context to produce a set of representative and homogeneous test items or tasks. That is, any item writer should be able to produce an item or task that the specification writer would acknowledge as being consistent with, or fitting, the test specification. As a further illustration of what a specification is and what it can do, consider Figure 3. The figure is not intended as an example of a flawless specification but rather as a working draft. Variations in the format are possible. Also, no specification is ever in its final form and, much like the process of establishing validity, we continue to test our specifications over time.

In the example in Figure 3, the Sample Item is presented after the General Description. In some situations the specification writer may feel it helpful to illustrate with an example before detailing the Prompt Attributes and Response Attributes. The point is that the specification format is a flexible tool that test developers can reshape to respond to specific testing requirements.

CRLTD, then, is the process by which these test specifications are developed. Basically, the process consists of a series of steps carried out by individual test developers or by a test development team. First, the testing context needs to be identified. Here, the concept of the *mandate* expresses the motivation and articulation of testing needs. In an instructional setting, it may come from the curriculum, the textbooks, the administration, the teachers, or other such sources. Next, a preliminary draft of the test specification, following the format presented above, is developed. The specification is then used to produce an item or task. At this point, the experience of item writing is used to provide feedback for refining the specification. This, in turn, can suggest changes in the originally produced item or task. Depending on time constraints and other limitations of the test development context, this process can continue through several iterations until the test developer is prepared to trial the test items/tasks.

When teachers develop test specifications together, their collaboration can serve to illuminate larger testing and evaluation issues for them and give them a voice in the discussion of those issues and the policies and tests that result from such efforts. In particular, the CRLTD process can help them to better articulate their understanding of their curriculum objectives and help them to link those objectives to the testing mechanisms used to evaluate student achievement. A workshop approach to CRLTD can develop the ability of teachers to engage in this collaboration. The steps in such a workshop are given in Davidson and Lynch (1993) and are reproduced here as Figure 4. The workshop is, essentially, an elaboration of the CRLTD process outlined above.

Like the test specification format, the CRLTD workshop steps are not meant to be fixed and immutable. We have used this process with

FIGURE 3

Writing Letters of Complaint Specification

TITLE: Writing: Letters of Complaint: Business Products

GD: It is important for learners in an ESL environment to know how to write culturally appropriate letters of complaint. Students will demonstrate their knowledge of cultural appropriateness by using proper letter format, relevant information, and proper register.

SI (The student will receive a printed card.): You are a student. You have just purchased a radio from Radio Shack. When you take it home, you find that you cannot tune in your favorite station. Write a letter of complaint to the manager of the Customer Service Department and ask for a refund or exchange. Make sure that your letter describes your situation (include who, what, when, where, why/how), is written in the format of a standard business letter, and is of the proper register.

PA: Each student will be given a card that includes his/her role, the role of the addressee, and a minimum of one more piece of relevant information (see SS) concerning a complaint about a business product.

RA: The student will write a letter of complaint to describe the problem. This implies that the letter will contain relevant information and be written in proper letter format and proper register.

SS: Relevant information for a letter of complaint about a business product should include the following factors:

- *who* (who the sender of the letter is), such as housewife, secretary of a company, or student (optional)
- *what* (what the problem is/what the product is), for example, item damaged at time of purchase; broken very shortly after purchase but not complainee's fault; not satisfied with quality of the item
- *where* (where the product was purchased)
- *when* (when the product was purchased and/or when the problem occurred)
- *why/how* (if known, how or why the problem occurred) (optional)

Proper letter format should include elements of a standard business letter:

address of sender
date

address of company

salutation
body of letter

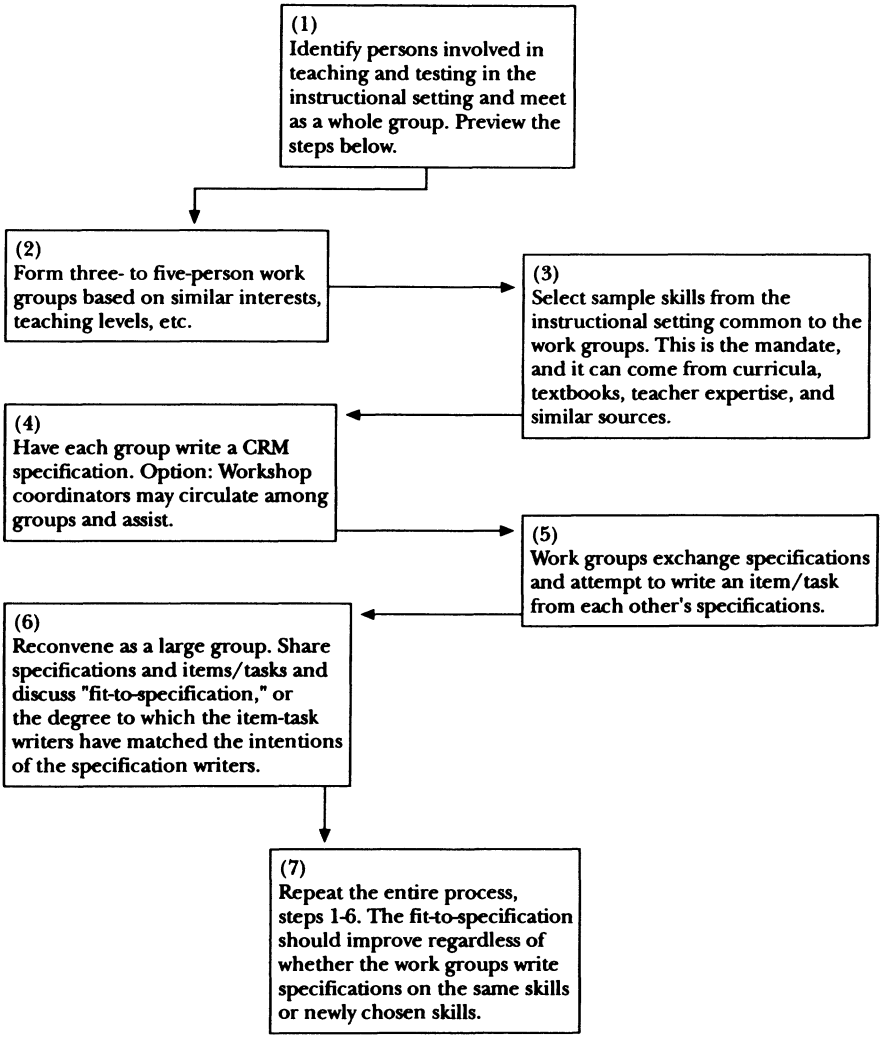
closing,
signature

Note. From the Testing Seminar taught by Fred Davidson at the University of Illinois, Urbana-Champaign, Spring 1992.

teachers in Costa Rica, Egypt, Finland, Guatemala, Panama, the UK, and the U.S. Each setting provides a somewhat different dynamic, which can suggest different orderings or characterizations of the steps. For example, in Step 1 the group discussion may generate Step 3's sample skills from the instructional setting, which will in turn suggest the composition of Step 2's three- to five-person work groups.

Another minor revision to the process would be to have the work

FIGURE 4
The CRLTD Process



groups exchange specifications in Step 5, without giving the exchange group a Sample Item. That is, each group would receive a specification with the General Description, Prompt Attributes, Response Attributes, and optional Specification Supplement and would write their item/task without seeing a Sample Item from the specification writers. Experience in a recent set of workshops suggested that the participants

tend to focus too much on the Sample Item when attempting to use the specification for the first time.

GUIDELINES FOR USING THE CRLTD PROCESS

Although the process of test development is obviously larger than the stages focused on here (refer to Figure 1), we believe that CRLTD is a crucial component. We have found that CRLTD stimulates very provocative discussions, and our experience in using the CRLTD workshop has also suggested that common problems tend to arise. The following are guidelines for avoiding those common problems, whether CRLTD is used in a workshop format or by individual test developers.

1. Strike a balance between generality and specificity. Be careful not to write an item instead of a specification for generating items. Conversely, be careful not to write a specification so general that it fails to provide the item writers with sufficient information to carry out their task.
2. Keep a distinction between the Prompt Attributes and the Response Attributes. That is, make certain that the Prompt Attributes are what will be presented to the examinee and the Response Attributes are what the examinee will be expected to do as a result of the Prompt Attributes.
3. Consider the potential item writer. Throughout the specification-writing process, ask questions such as the following: "Does the item writer have access to the necessary materials to produce an item from the specification?" "Does the item writer need any special training in order to be able to interpret and use the specification?" "Does the specification clearly state, in language accessible to the item writer, what it intends the items/tasks to look like and do?" (Note: Individual test developers can ask these same questions of themselves.) As a specific instance, and perhaps the most common problem of this type, consider whether the item writer is to write an item, find a passage, or both. If the task is to find a passage, be certain to specify the passage fully.
4. When beginning the cycle of CRLTD again (Step 7), be careful not to lose sight of the original criterion to be tested. It is easy to let problems with the test item/task format (the Prompt Attributes and Response Attributes) lead to solutions that result in a more efficient item/task that does not truly capture the intended criterion skill or ability. For example, an original criterion of "ability to successfully

communicate orally in an academic discussion section" could wind up as a test of pronunciation.

5. Consider the mandate—the motivation for developing the test, which comes from a combination of curriculum philosophy and political reality. The mandate may flow freely from the curriculum and teachers or it may be more or less imposed by forces external to the instructional setting. Teachers developing tests may or may not have much flexibility in how they select the skill to be assessed and the method of its assessment. CRLTD can provide feedback to the mandate, but even that may be systematically restricted because of the particular administrative structure.

Figure 5 presents a CRLTD exercise used in a recent testing seminar that further illustrates what we intend by the mandate. Students produced specifications and commented on each others' product in a class discussion. Interestingly, the class agreed that the kind of semirigorous context, or mandate, given in the handout was actually beneficial. It seemed to reduce the time necessary to write the specification.

Realistically, testing is never done in isolation. External pressures—time, politics, money, people, and so on—impinge on many choices. These factors will obviously shape the mandate and affect the outcome of any CRLTD activity. A strength of the process outlined here is that teachers are afforded the opportunity to systematically work toward an accommodation of the mandate to their collective vision for ESL teaching and learning.

THE INTEGRATION OF TEACHING AND TESTING: INSIGHTS FROM CRLTD WORKSHOPS

We have accumulated a good deal of qualitative data from teachers who have used the workshop approach to CRLTD in a variety of settings. Some of these data illustrate the potential for CRLTD to illuminate the connections or mismatches between curriculum, texts, methods, and tests. The process can also lead to the clarification of instructional objectives. Of course, it will usually provide some insight into refining the nature of CRLTD itself.

Building Consensus on Test Objectives

At one institution, workshop participants were able to articulate the gap between their teaching methodology (which was described as the communicative approach) and their exit test (which, ultimately, was measuring fairly discrete-point grammar). However, the discussion

FIGURE 5
CRLTD Exercise: The Mandate

Imagine that you and your fellow teachers are working at a top-down, fairly conservative language-teaching institute, the Fairly Conservative Language Institute (FCLI). You and the other teachers have proposed for several years that the FCLI administration try more communicative language-teaching methods. Yet generally a grammatical-core syllabus still informs the FCLI curriculum.

You are now given an opportunity to implement a communicative test activity as part of the placement exam process. The goal of the test is to determine if students can already do certain tasks and should therefore be placed into the top level, where performance of these tasks is assumed. The FCLI Curriculum Coordinator has given you three course objectives that s/he feels are more communicative and that represent assumed proficiency in the top level of the FCLI course sequence:

Objective 1: Students will be able to listen to fast, unsimplified L2 speech as spoken on radio or TV. (Note: S/he is particularly concerned with, as she puts it, the "rapid-fire speech you hear on nightly newscasts or between songs on morning radio shows.")

Objective 2: Students will be able to role play an information-gathering encounter in the L2. (Note: S/he conceives of the following as typical information-gathering encounters: talking to a reference librarian, consulting with a computer specialist, asking for advice from a psychological counselor, and so on.)

Objective 3: Students will be able to write an L2 letter of thanks for a service or kindness received for free. (Note: S/he imagines things like thank-you letters when somebody has done someone a favor or has been particularly nice to that person.)

Each group in the room will be assigned one objective above. Recalling that curricular-mandated objectives, like these, can become a CRLTD Specification GD, you and your group should write a specification where the assigned objective triggers the whole specification. In your discussions, try to make the test task palatable to the conservative administration of the FCLI.

Note. From the Testing Seminar taught by Fred Davidson at the University of Illinois, Urbana-Champaign, Spring 1993.

did not merely end with this relatively superficial recognition of the form/accuracy versus meaning/fluency tension in L2 teaching. In attempting to write specifications that would work with their methodology and curricular goals, the participants were able to isolate particular problems. For example, in their attempt to revise the existing rating scale for assessing students' written communicative ability, the workshop participants realized they were potentially penalizing individual students again and again for the same mistake. Such an error correction scheme might make sense in another curriculum, but it did not match the goals of this particular institution. This kind of mismatch can often go unnoticed unless teachers are included in the test development process. This mismatch also represents the potential for what we have termed *reverse washback*—knowledge from teaching having an influence on testing. Qualitative data such as these may help answer Alderson and Wall's (1993) call for more research on washback, in particular by clarifying the directionality of the influence, which we found from teaching to the test rather than vice versa.

The CRLTD workshop experience can also uncover mismatches between textbook materials and existing tests. In one setting, the teachers came to the realization that the metalanguage of textbooks and teaching plays a crucial role in the testing process. They were able to pinpoint ways in which their textbooks used terms that were different from those used for the same features in the existing test instructions. For example, the textbook used a term such as *present progressive*, whereas the test used a term such as *present continuous*. Furthermore, even though the students were accurately producing this form orally in their classroom activities, many seemed unable to do so on the test. Teachers in the workshop discussed the possibility that students were being tested on their knowledge of linguistic terminology rather than their ability to use these forms communicatively, which was the instructional objective. A focal point for their specification-writing process became to detail the terminology to be employed in the test items/tasks, attempting to match it to that used in the instructional materials and activities. This kind of information is typically presented in the Specification Supplement.

In another workshop discussion, the accuracy versus fluency issue arose again, but in this case the CRLTD process led the participants to clarify their instructional objectives. As they attempted to write items/tasks from the first round of specifications (Step 5 of the workshop—see Figure 4), the teachers began to discover differing conceptions of what their communicative curriculum entailed. Some felt that the curriculum had nothing to do with the formal accuracy of their students' speech; others, that there was a certain amount of concern for accuracy; and others, that the curriculum had an important component of attention to linguistic form. This discussion and the need to clarify the Response Attributes for the test specification in order to write an appropriate item/task led the teachers to a consensus that their objectives included accuracy of form in addition to communicative fluency. The specification revision process, then, became one of clarifying when and how grammatical form would be a part of the assessment of communicative ability.

Feedback on the CRLTD Process

In terms of refining the CRLTD process itself, every workshop makes its own unique contribution. In a recent graduate seminar in language testing, students experienced a CRLTD workshop as a part of their course work. Their reactions and critical observations made us aware of aspects of the specification format and CRLTD process that needed attention. Several of the students were concerned about

the feasibility of doing CRLTD in real-world teaching contexts. They pointed out the large amount of time and effort required to properly carry out CRLTD. As a related issue, they questioned whether the teacher-driven approach to testing that CRLTD represents is likely to be adopted by the power structures that are responsible for testing and evaluation.

This type of questioning has made us aware of the need to provide a clear rationale for the approach to test development that CRLTD represents. In part, such a rationale is what distinguishes the CRT end of the continuum from the NRT end. The development of NRTs is governed by careful attention to item statistics. This attention comes from staff members who have training in psychometrics and who judge the quality of test items through various statistical procedures. Teachers are not typically expected to involve themselves with such procedures and thus tend to have a minimal role, if any, in the test development process. We contend that any CRT, to be fully realized, needs to evolve from something like the workshop process presented here. CRLTD involves teachers in the generation and refinement of test items and gives priority to teachers' knowledge and experience over item statistics when deciding on the value of test items. Statistical characteristics of the item are consulted, but only after the CRLTD process has established preliminary evidence of its validity.

In a very real sense, this process challenges the existing structure and authority of testing and evaluation practice. As such it underscores both the positive potential of empowerment for teachers and the importance of presenting a rationale for the benefits and feasibility of CRLTD within existing practice. Testing experts need to understand the enhanced evidence of validity provided by CRLTD (because of the clear link between the specification and the instructional goals, or reverse washback), and administrators will need to be convinced it is worth the time and effort. We have found the time commitment necessary for specification writing to decrease with experience. The first few times a group produces a specification, it can take quite a while. If the specification process—either in a workshop or in actual test development operations—continues, and if the participants try additional specifications, they become more adept at the process and are able to produce specifications more quickly. The specifications, in turn, result in test items and tasks that will require less time and effort to refine via trialing and item analysis than those produced without specification guidance.

In part this increased proficiency in specification-writing may be thought of as the acquisition of what we have come to call *Speclish* (the language of test specifications). Language like "The texts and words

will be of both a scientific/technical and a general/nontechnical nature, to tap into a student's background in a variety of areas" is, at first, very difficult for CRLTD participants to use. Usually, it does not occur to them to be that specific about the range of text sources. Or, if they do attempt to specify the range, they are not always able to come up with the guiding language that is necessary for Speclish. This skill is very difficult to develop, but it is often acquired with CRLTD practice.

Comments from workshop participants have also resulted in an emphasis on the communication between specification writers and item/task writers. This, in part, echoes point 3 mentioned above in the guidelines for using the CRLTD process. One refinement in the process that was suggested in a recent workshop experience is to have the participants keep a systematic record of comments and questions that occur during Step 5 (writing an item/task from another group's specification). Often, the CRLTD workshop presents the results of Step 5 by having the work groups write their items/tasks out on newsprint paper (the specification itself, from Step 4, is also written out) and display the newsprint on the walls of the workshop room. Comments and questions from the item writers could be incorporated in the margins of this display, as could the comments and questions that result from the large-group discussion in Step 6. The newsprint displays thus become a permanent record of the CRLTD session, which would be of great value for future test development.

Another element common to all of these workshop experiences is that they make evident the link between teaching and testing. When CRLTD participants struggle over the components of the test specification, they are experiencing the essence of operationalization (we have argued this point elsewhere: Davidson, Hudson, & Lynch, 1984). That is, they are attempting to translate their teaching constructs, which can also be thought of as research hypotheses, into operations that can be measured and interpreted. Sometimes these constructs are specific teaching objectives, for example, the ability to write an argumentative essay for academic purposes. At other times the constructs are more general notions of second language acquisition, for example, fluency in oral communication. In either case, the CRLTD process helps guide the teacher to find the clear and interpretable link between the object of inquiry—the teaching objective or research hypothesis—and the means of inquiry—the language test. Scholars like Bachman and Clark (1987) and Bachman (1989) have called for developing CRTs of L2 proficiency as the cornerstone of their research and development program. We see CRLTD as an effective means of involving large numbers of teachers in that program.

CONCLUSION

CRLTD obviously needs to be integrated into a larger, multistage process of test design and development. CRLTD's focus is primarily on linking writing specifications and writing test items or tasks. Additionally, the selection of the skill to be tested is often a prime target during specification revision, and we have noted that CRLTD can yield honest questioning of an external mandate. We believe that clear communication between specification writers and item writers (even when that communication occurs within one individual doing both tasks), and the consequent refinement of the specification and items, both contribute to a better measure. However, we acknowledge the continuing need (as in all good testing) to pilot and monitor all operational measures, which involves another important channel of feedback in the test development process.

As we have attempted to demonstrate, CRLTD, with its focus on test specifications, represents a means of linking ESL curricula, teacher experience, and language tests. Although we have illustrated CRLTD in the form of a workshop, the process can easily be applied to ongoing test development in a variety of institutional settings. The workshop flow chart presented in Figure 4 might be realized via interoffice memo, in staff meetings, or through electronic mail exchanges.

Implicit in the examples from the workshops we have conducted using CRLTD is the notion of washback. We also point to the UCLA ESLPE and the UIUC ESL Placement Test as evidence of reverse washback—the influence of teaching on language testing. These language tests, both of which are used to place international students into ESL curricula, have been significantly revised using CRLTD, resulting in more detailed test specifications. The hallmark of those revisions has been to make the tests more representative of the teaching that characterized the curricula into which they were placing students. In the case of the ESLPE, the test has changed from a TOEFL-like exam to one that incorporates genuine samples of academic discourse on the listening and reading subtests as well as an academic essay. At UIUC, CRLTD resulted in a similar evolution, and the test now includes a video/reading-based academic writing task. In both cases, the evolving CRLTD specifications represent a level of detail not found in previous versions of the tests and are considered to be useful evidence of test validity.

Finally, at the heart of CRLTD and its potential for washback is the notion of the mandate, discussed earlier. We believe that this notion is critical for understanding CRLTD within the wider language-testing context and for reaping the maximum benefit from this process.

THE AUTHORS

Brian K. Lynch is a Lecturer in the Department of Applied Linguistics and Language Studies at the University of Melbourne. His primary research interests are program evaluation and testing.

Fred G. Davidson is Assistant Professor in the Division of English as an International Language at the University of Illinois, Urbana-Champaign. His primary research interests are language testing and data structures in applied linguistics.

REFERENCES

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 115–129.

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

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Clark, J. L. D. (1987). The measurement of foreign/second language proficiency. *Annals of the American Academy of Political and Social Science, 490*, 20–33.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 22*, 65–84.

Brown, J. D., Detmar, E., & Hudson, T. D. (1992, February). *Developing and validating tests of cross-cultural pragmatics*. Paper presented at the 14th Annual Language Testing Research Colloquium, Vancouver, BC.

Cartier, F. (1968). Criterion-referenced testing of language skills. *TESOL Quarterly, 2*, 27–32.

Cook, G. (1992, March). *The place of placement tests*. Paper presented at the 26th Annual TESOL Convention, Vancouver, BC.

Cziko, G. A. (1982). Improving the psychometric, criterion-referenced, and practical qualities of integrative language tests. *TESOL Quarterly, 16*, 367–379.

Davidson, F. G., Hudson, T. D., & Lynch, B. K. (1984). Language testing: Operationalization in classroom measurement and L2 research. In M. Celce-Murcia (Ed.), *Beyond basics: Issues and research in TESOL* (pp. 137–152). Rowley, MA: Newbury House.

Davidson, F. G., & Lynch, B. K. (1993). Criterion-referenced language test development: A prolegomenon. In A. Huhta, K. Sajavaara, & S. Takala (Eds.), *Language testing: New openings* (pp. 73–89). Jyväskylä, Finland: University of Jyväskylä, Institute for Educational Research.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.

Hudson, T. D. (1989). *Measurement approaches in the development of functional ability level language tests: Norm-referenced, criterion-referenced, and item response theory decisions*. Unpublished doctoral dissertation, University of California, Los Angeles.

Hudson, T. D., & Lynch, B. K. (1984). A criterion-referenced measurement approach to ESL achievement testing. *Language Testing, 1*, 171–201.

Hughes, A. (1989). *Testing for language teachers.* Cambridge, England: Cambridge University Press.
Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice Hall.
Popham, W. J. (1981). *Modern educational measurement.* Englewood Cliffs, NJ: Prentice Hall.
Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10,* 41–69.

References

The Measurement of Foreign/Second Language Proficiency
Lyle F. Bachman; John L. D. Clark
Annals of the American Academy of Political and Social Science, Vol. 490, Foreign Language Instruction: A National Agenda. (Mar., 1987), pp. 20-33.
Stable URL:
http://links.jstor.org/sici?sici=0002-7162%28198703%29490%3C20%3ATMOFLP%3E2.0.CO%3B2-U

Improving ESL Placement Tests Using Two Perspectives
James Dean Brown
TESOL Quarterly, Vol. 23, No. 1. (Mar., 1989), pp. 65-83.
Stable URL:
http://links.jstor.org/sici?sici=0039-8322%28198903%2923%3A1%3C65%3AIEPTUT%3E2.0.CO%3B2-H

Criterion-Referenced Testing of Language Skills
Francis A. Cartier
TESOL Quarterly, Vol. 2, No. 1. (Mar., 1968), pp. 27-32.
Stable URL:
http://links.jstor.org/sici?sici=0039-8322%28196803%292%3A1%3C27%3ACTOLS%3E2.0.CO%3B2-A