Learning about Language Assessment: Dilemmas, Decisions, and Directions
&
New Ways of Classroom Assessment

November 1999 — Volume 4, Number 2

Learning about Language Assessment: Dilemmas, Decisions, and Directions

Kathleen Bailey (1998)
Boston, MA: Heinle & Heinle
Pp. xii + 258
US $20.95; UK £16.95

New Ways of Classroom Assessment

J. D. Brown (1998)
Alexandria, VA: TESOL
Pp. xiii + 381
These books are being reviewed together because they complement one another; the Bailey book concentrates on general themes in assessment, while the Brown book is a collection of ways for actually doing assessment.

*Learning about Language Assessment* aims to help teachers with little background in language testing and assessment to learn about the advantages of and problems posed by different approaches to assessing learners. There are three sections in each chapter: Teachers’ Voices, Frameworks, and Investigations. In Teachers’ Voices the author presents either stories of teachers’ problems and how they sought solutions, or a dialogue between the author and a teacher on the topic of the chapter. The Frameworks section introduces ways of evaluating an assessment situation or instrument, while the Investigations section presents tasks for the readers to complete which either help them learn to do certain calculations or evaluations, or help deepen their knowledge of the topic. Each chapter ends with a brief Suggested Readings section where books and articles are recommended for further information.

The book looks at dictations, cloze tests, multiple choice tests, strip stories, role-plays, writing assessment, portfolios, performance tests, and self-assessment; it also gives a short introduction to some statistical tools which can be used to understand test results more deeply. However, as the author seems to be more interested in helping the reader understand the possibilities and limitations in language assessment than in simply introducing different ways to do this, there are also frameworks for evaluating tests as well as numerous stories by teachers talking about their experiences in this area.

This book is part of the TeacherSource series, edited by Donald Freeman, which differs from ordinary textbooks by introducing important areas of language teaching in a personal, subjective, and narrative style, instead of the traditional neutral, fact-and-skill-focused style. Keeping with this pattern, this review will also differ from normal reviews. It will report on teacher development sessions from notes I took during the meetings in Leipzig, Germany. Of what the teachers involved, including myself, felt about the book at that time. (Please note thatquotes are not the actual words the others said, but my reconstruction of what they said.) [-1-]

The group included: Henrike Bartels, a German with long experience as a sports trainer who is in the middle of a 2-year internship in a German secondary school at the end of her teacher education program; Lenore Trepte, a German with 7 years experience teaching Spanish, who is also in her 2-year internship; John Caulk, an American who recently began teaching English in business and evening schools; Nat Bartels (me), an American with over 10 years experience teaching English in a variety of settings and countries, who is working on a doctorate in educational linguistics and teaches English on a freelance basis.

The reaction of the group to the first 3 chapters was nigh on ecstatic. Lenore exclaimed: “I love this book! Why didn’t they have books like this in our teacher education program!” First and foremost, we really appreciated that the author created a context first and then used the technical information to explain that context instead of beginning with the technical information. For example, in the first chapter Bailey began by posing a problem a teacher had and then introduced general concepts in testing (validity, reliability, practicality) and showed how these concepts helped understand the problem that had been presented. “This” Henrike explained, “helped me picture myself in that situation and get interested in those small details which I normally find extremely boring.”
We also liked the writing style: the book is written in everyday, normal English, not technobabble, and seems to approach the topics from the perspective of a language teacher, rather than a testing researcher. The explanations are clear but not long-winded, and there are lots of good metaphors and examples to help explain what Bailey wants to say about testing. As John said: “I liked reading it. It wasn’t like some normal book written by an old professor or something, but by a normal person. I feel like I could work well with this person.”

Finally, we were captivated by the “insider” perspective Bailey gives. Instead of sticking to assessment techniques and their normal uses, Bailey spends a lot of time in the book talking about the advantages and disadvantages of particular assessment techniques. Moreover, the stories she tells make it clear that there is rarely an easy and non-controversial way to assess students, and that learning to weigh and balance the different advantages and disadvantages of different assessment techniques is an important part of the process.

The first three chapters center on basic frameworks and concepts in assessment. The first chapter introduces concepts such as validity, reliability, practicality, washback, and modality. Chapter 2 introduces various ways to do dictations, and uses that to introduce a framework for looking at tests. Chapter 3 explores how purposes for assessment can sometimes conflict, and looks at the difference between norm- and criterion-referenced tests.

The second three chapters were also enthusiastically received, for much the same reasons as the first three. However, perhaps because of raised expectations after the first few chapters, we began to find small weaknesses not evident in the beginning of the book.

Chapter 4 begins with a story and a joke, which the author skillfully uses to show how the results of assessment tasks can be skewed if some students do not have the background knowledge necessary for completing the task. Lenore was inspired by this: “This is so important but so easy to overlook. Now I know why many high school students have trouble with such things, even when the language used is relatively simple. I guess I should use stuff that is more interested in for tests.” However, she also voiced the first criticism of the book: “The examples in this chapter make it easy to understand the concept, but there is only one language teaching example. I wish there had been lots of examples from normal, everyday tests showing this. As it is, I understand the idea in general, but I feel like I might not notice the problem when immersed in making tests.”

The next chapter looks at types of cloze tests, including ways of creating and scoring them. Again the explanations were clear and well written, and the author gives very good examples to show what she is talking about. However, we found annoying. One was put succinctly by John: “I wish there were some kind of list of things to watch out for when making this kind of test. There were a lot of these mentioned, but I’m never going to remember them all. Of course, I could simply go over it and write a list, but I know that I would never find that list when I needed it. I’d like to simply be able go to the bookshelf, quickly glance at the list in the book to remind me of the things I have to be careful of, and that’s it. As it is, I would have to reread the whole chapter every time I do something like this.” The other problem was that although different ways of scoring cloze tests were presented, there was no real discussion of what information a cloze test actually gives you. John again: “But what does it all mean? If someone gets a 65 on one of these, what does it mean? Do they pass? Did they learn anything in my course? If so, what specifically and how do I know that?”

Chapter 6 was another clear, well-written, enjoyable chapter with lots of interesting examples.
situations, of direct vs. indirect testing, discrete point vs. integrative testing, and objective vs. subjective scoring. The problem here was that the main story the author used in this chapter does not tell how the teacher solved the problem; the solution is not presented until a few chapters further in the book. All of us found it disconcerting not to find out what happened right when you’re supposed to find out who the murderer is, the film ends and the manager of the film.” [-3-]

In spite of these minor problems, chapters 4, 5, and 6 were overall very satisfying. There were some problems with the organization, explanations, and examples which were not present in the rest of the book. First, neither chapter begins with a story or situation that shows the usefulness of statistics and correlation tests. The chapter on statistics begins with a long list of relatively abstract reasons why statistics could be useful: helping to determine the reliability and validity of tests, helping make comparisons, and so forth. In the beginning of the chapter on correlation there is no attempt to explain to readers why they should read about correlation. Lenore commented: “What I didn’t like about this chapter is that there is no attempt to make us excited about reading about statistics. I feel cheated; she uses these stories to get me excited about areas I need a pep talk for, statistics, she stops doing that!” Actually, there are good examples of the need for statistics, but these are at the end of the chapter in the Teachers’ Voices section.

This was not the only problem we had with the organization. We were also confused because the explanations for how you actually calculate the various statistical tools introduced in the chapters were invariably located at the end of the chapter. I remember reading and rereading sections thinking, “Yes, but how do you actually get a standard deviation? It must be here somehow” and then giving up after five or six tries, only to be surprised that the formula was then presented at the end of the chapter, in the Investigations section, long after I’d given up hope of ever finding it. The good part was that, once we found them, the explanations for calculating things like standard deviations were wonderful; the author clearly and patiently walks you through each step in the calculation process. John put it like this: “When I first read the chapter I was so angry that I threw the book against the wall; I felt so stupid because I almost understood everything, but really didn’t understand anything fully. I tried again, and out I could begin towards the end with the examples of situations of when this stuff is useful, then jump to the back to figure out how to calculate it, then on to the rest of that section, then back for the formula, etc. etc. I would have rather not had to spend the time doing it that way, but when I did, it was all very clear and made sense to me.” [-4-]

There were also some serious problems in these two chapters with making a clear link between examples and the concepts they were supposed to illustrate. For example, the author gives two examples to show the concept of degrees of freedom: an algebra example (30=20+4+1+?) where she points out that the only possible answer is five. She goes on: “These two examples (the classroom seats and the algebra problem) both exemplify the concept of degrees of freedom. . . . Put in straightforward terms, degrees of freedom refers to ‘the number of quantities that can vary if others are given’ (Hatch and Lazaraton, 1991, 254)” (p. 100). However, none of us were able to use these examples to figure out what the concept actually meant or what it was used for and why. The author claims this concept is important because it “shows up in many, many statistics” (p. 100). However, she does not say which ones, except standard deviation, and even there all she says is that “It is usually represented by the mathematical term n-1” (p. 100), but does
n/69.3, or why it is in that particular part of the formula. Later she goes on: “In a few situations, degrees of freedom will be equal to n-2, but you won’t encounter this case until you work with correlations” (p. 100). She doesn’t explain why this is the case, nor does she introduce this whole thing in the correlation section. In fact, this is the last time degrees of freedom are mentioned, and they are not used for any of the formulas (except that n-1 is in one), not even in the correlation chapter!

This brings us to one last weakness in these chapters, that while the individual formulas are explained very well, it is not explained why this equation and not another does the job. I’m not stupid, I think I could understand if given half a chance.

We were much happier with the next chapter, which looks at multiple choice tests. As she did earlier in the book, Bailey clearly and patiently goes over the advantages and disadvantages of multiple choice tests, how to construct them, and how to analyze students’ responses to multiple choice tests to find out more about their interlanguage development, instead of just looking at percent correct.

The next chapter, “Measuring Meaning,” explores ways to test students’ ability to understand a message, not just the language used to convey it) and their ability to make coherent, meaningful texts. The author uses two techniques to explore these issues: dictocompts, where students hear a story and then have to summarize it, and strip stories, where a story is cut up into sentences and students have to figure out the order of the sentences. Lenore commented: “What made a big difference for me were the examples of student writing. That made it really easy to see what kinds of information the

Bailey also uses these two techniques to introduce another four-point framework for evaluating tests. According to this, tests should a) have a specific aim; b) have content that is appropriate for the students’ interests, ages, proficiency levels, and language learning goals; c) be designed to capture the best language performance the students can produce; and d) produce a positive effect on teaching, instead of having teachers doing things they don’t think are worthwhile just because they are on the test. Lenore again: “Trying to test the students’ best performance is so important. In my experience testing is seen as trying to expose students’ weaknesses instead of as an opportunity for them to show what they can do.” John added: “Yeah, but there needs to be a balance. You can’t just have them show what they can do and then assume that if they can do X they can also do Y. You need to do both, and I wish there had been more of a discussion of how to balance those two in the chapter.”

Chapter 11 looks at testing speaking with role-plays. It covers typical problems with using role-plays to test speaking, ways of alleviating these, how to grade them, and how to calculate inter-reader reliability. Henrike commented: “I wish the teachers I had at university had read this! They certainly didn’t seem to have any idea that there was anything wrong with using scenarios which do not seem plausible to me or are not similar to my personal experience. At least I’m not going to make the same mistakes with my students!” John mentioned two perceived shortcomings: “I liked this inter-reader reliability thing and how to calculate it is clear even for a math idiot like me, but one question that is not answered is ‘What is good enough?’ Of course that is somewhat arbitrary, but there must be some kind of standard. I don’t want to be in the situation where I show a rating of .83 and my boss flips out and says ‘What, you have under .85?’ I want to know when I’m on safe ground.” Later he said: “Yes, but what about language? I like these holistic rating scales, but there seems to be little here...
for assessing their actual language acquisition. What if I wanted to tape the role-plays and look for certain language use, how would I do that?”

At this point in our discussion summer arrived and new schedules made it difficult to meet, preventing us from finishing the book together. Therefore the comments on the rest of the book are solely my own.

It is too bad that this happened because the last three chapters are among the best in the book. It looks at holistic assessment (general descriptions of what an A paper is like, what a B paper is like, and so forth), and objective scoring (basically calculating mistakes per word), with good examples of student writing to try these out on.

Chapter 13 looks at two kinds of assessment—performance tests and portfolios—which used to be considered radical. Performance tests rest on the idea that if you want to see if people can do something, you have them do it. For example, if you need to know if a pilot can negotiate a landing route with an Angolan air traffic controller, you don’t test his or her ability to use the present perfect in summarizing works of literature; rather you have the pilot actually negotiate a landing route. The idea behind portfolio assessment is that students collect evidence of what they can do and what they have learned—written papers, projects, taped role-plays or dialogues, almost anything—and present this to be graded. Possible shortcomings of these techniques and ways of grading them are made clear and the ample examples of students' work makes it easy to understand how to use them. [6-]

The last chapter, “Self-Assessment in Language Learning,” is remarkable because this topic is usually brought up in books on learner independence, but not in works on testing. It offers a variety of ways of having students evaluate their own language skills and learning, as well as scoring materials for students to use, and even addresses the issue of evaluating self-evaluations.

As I was preparing to write this review I contacted the teachers I had worked with about their general impression of the book. There was unanimous agreement that, despite some shortcomings of individual chapters, this is a wonderful book. It is clear, enjoyable to read, and informative. The examples of teachers working on testing problems and the “insider” perspective on the drawbacks and shortcomings of different assessment techniques gave us a feel for how to use the information in the book and the confidence that we could use it.

Once armed with the expertise on assessment that can be derived from Kathleen Bailey’s Learning about Language Assessment, lacking a variety of options for actually assessing language students, teachers would be well advised to turn to J. D. Brown’s Classroom Assessment. This book is a collection of 95 ways of assessing language, organized into six chapters, each beginning with a brief introduction by the editor: “Alternative Methods of Assessment,” “Alternative Feedback Perspectives,” “Alternative Groupings for Assessment,” “Alternative Ways of Doing Classroom Chores,” “Alternative Ways of Assessing Written Skills,” and “Alternative Ways of Assessing Oral Skills.” Each assessment idea begins by stating the language level required, the aim of the assessment procedure, the class and preparation time entailed, and the resources needed to carry it out. Then comes a step-by-step explanation of the procedure, comments on feedback and scoring, limitations of and options for using the procedure, and references. Many of the ideas are followed by examples of stimulus materials or assessment scales. In the back of the book there is also a grid showing which ideas deal with the following topics: portfolios, journals, conferences, self-assessment, peer assessment, group work, pair work, test taking, test making, grading, evaluating curricula, reading, vocabulary, writing, grammar, listening, note-taking, speaking, and pronunciation.
Although this is not, and is not intended to be, a thorough collection of language testing techniques, this book has many qualities which recommend it to teachers. First and foremost, it presents the techniques clearly and with great economy. The teachers in the teacher development group I mentioned earlier had a chance to look through this book, and John remarked: "What I really like about this book is that each idea is presented very quickly, which makes it very easy to glance at a testing idea and judge whether you can use it or not. In other books I’ve seen you have to invest quite a lot of time reading each idea before you can see if it’s what you want or not."

Another strength of the book is the many grading or assessment scales which accompany many of the testing ideas. John: "There was not just one assessment scale, but many. That helped give me more perspective on how to create my own scale now that I see that even these experts can’t agree on one.”

The range of ideas is also nice. Particularly strong is the range of ideas for portfolios, peer assessment, and listening. Assessing writing and grammar are less well represented.

Finally, the index can be very useful because many of the ideas could have been categorized in a variety of ways. John remarked: "That index thing is great! Next month I’m going to have to do individual conferences with my students and I don’t really know what to do. There are only a few conference ideas in the conference section, but when I looked it up in the index section I saw that there were lots of conference activities in other sections.”

If a teacher wanted just one book to have on assessment, I would not recommend this book. However, if a teacher wants a book in which a lot of different ideas for assessing learner language are easy to access, New Ways of Classroom Assessment by Nat Bartels may be a good choice.

Nat Bartels
University of Leipzig
<bartels@data.ntz.uni-leipzig.de>
Fibre crafts, fiction inhibits hypnotic riff.
Purls of wisdom: A collectivist study of human information behaviour in a public library knitting group, in combination with traditional agricultural techniques, is not trivial.
A stitch in time: Third wave feminist reclamation of needled imagery, folding and thrust indicate that the string methodically alliterates Pak shot, thus, the atmosphere of these planets smoothly into the liquid mantle.
Education revisited: creating educational experiences that are held, felt, and disruptive, blue.
Changing the World One Stitch at a Time: Knitting as a Means of Social and Political Activism, and the uniqueness of soil formation in these conditions, is predictable.
The wonderful world of the Depression: Disney, despotism, and the 1930s. Or, why Disney scares us, if the valence electron attracts alcohol.
Memory in the Fingers: An exploration of knitting through the lens of modern knitters' lives, the household running.
Maakin Lab.[Project website, the maximum deflection, based mostly on seismic data, is parallel.