UC Riverside
Issues in Applied Linguistics

Title
An Interview with J. Charles Alderson

Permalink
https://escholarship.org/uc/item/2qc611zc

Journal
Issues in Applied Linguistics, 12(1)

ISSN
1050-4273

Authors
Vongpumivitch, Viphavee
Carr, Nathan

Publication Date
2001-06-30

DOI
10.5070/L4121005041

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

An Interview with J. Charles Alderson

Viphavee Vongpumivitch
University of California, Los Angeles

Nathan Carr
University of California, Los Angeles

PROFILE

For more than 20 years, J. Charles Alderson has been an internationally respected scholar in language testing. He has published research in a wide variety of areas, including reading assessment, test development, test validation, test impact (*washback*), computer-based testing, English for specific purposes testing, the effect of background knowledge on student performance, and the relationship between language testing and second language acquisition theory. His most recent book, *Assessing Reading* (Alderson, 2000), explores the nature of reading ability as well as issues involved in constructing and evaluating reading tests. Dr. Alderson is also a co-author of *Language Test Construction & Evaluation* (Alderson, Clapham & Wall, 1995), and he is also co-editor of the journal *Language Testing* and the Cambridge Language Assessment Series. Dr. Alderson is currently Professor of linguistics and English language education in the Department of Linguistics and Modern English Language at Lancaster University, United Kingdom. He is also the Scientific Coordinator of the DIALANG project, a web-based diagnostic test of 14 European languages, and he is an advisor to the British Council on the Hungarian English Examination Reform Project.

INTRODUCTION

The interview is divided into six sections. In the first section, Dr. Alderson briefly explains how he became involved in various areas of research. In the second section, *Washback Research*, Dr. Alderson discusses his work on the impact of tests on classroom teaching and learning, as well as teaching materials development and education policy. Dr. Alderson and Dr. Dianne Wall were among the first researchers to initiate systematic investigation of positive and negative impacts that tests may have on education. From their field research in Sri Lanka, Alderson and Wall (1993) proposed the groundbreaking *Washback Hypotheses* which describe the nature of washback and identify the types of influence that a test can have on both teachers and students. In this interview, Dr. Alderson reflects on the effort to create a theory of Washback in Alderson and Wall (1993), the need for

Issues in Applied Linguistics
ISSN 1050-4273
© 2001, Regents of the University of California
Vol. 12 No. 1, 91-109

empirical studies, and the role of Washback in his current research in Hungary on English education reform.

Shifting to another area of research, we then ask Dr. Alderson about the challenges, advantages, and disadvantages of computer-based and web-based language testing in the third section of this interview, *Computer-based and Web-based Testing*. At present, there is a growing interest in using computers and the Internet as media for delivering tests (see, for example, *System*, Vol. 28, No. 4, a special issue on technology and language testing). Based on his experience as the Scientific Coordinator of the DIALANG project, Dr. Alderson discusses the issues involved in developing web-based tests. He urges language testers developing computer- and web-based tests to avoid the use of traditional test item types such as multiple-choice and cloze, and suggests they move toward creating more useful kinds of test items that will make the best use of the new testing medium. He also discusses other challenges to web-based testing, such as overcoming the lack of security and finding ways to effectively score open-ended items, and the challenges faced by emerging research on ways to incorporate corpus analysis into language test development.

In the fourth section, *Test Validation and the Quantitative/Qualitative Paradigms,* Dr. Alderson shares his views on the characteristics of effective language test validation research and addresses issues surrounding the commonly discussed dichotomy between quantitative and qualitative research. Dr. Alderson challenges the either/or nature of this purported distinction, emphasizes the need to integrate both paradigms, and calls for more collaborative study by teams of researchers with different specializations.

In the fifth section of the interview, Dr. Alderson discusses second language reading research. He argues that even though much research has been conducted over the past 30 years, remarkably little is known about this area. Dr. Alderson holds that it is still difficult to make definite conclusions about any particular area of reading research except the role of background knowledge in reading comprehension.

Finally, in the last section we ask Dr. Alderson for suggestions for beginning researchers about conducting language testing research and about good introductory level books for those interested in learning more about language testing.

THE INTERVIEW

Carr: *How do you view the relationship between the various areas of research that you've conducted over your career?*

Alderson: I consider myself an applied linguist first and foremost. Obviously, I'm interested in language testing, but I'm interested in all things to do with language use and language learning, particularly in the second language context. To do testing you have to understand the construct you're trying to test. Pretty much

everything that I have done relates to teaching problems, one way or the other, and how to evaluate them.

Vongpumivitch: *You have said that your early work stemmed from problems you faced as an ESL teacher. Now that you are no longer an ESL teacher, what is it that sparks your interest in research projects?*

Alderson: Often what happens is somebody with a problem to solve comes along and says, "Can you help?" So back in '86, the British Council came along and asked if I could help them revise the ELTS (English Language Testing Service) test, which was the old British Council test. They wanted to bring it up to date with applied linguistic theory, improve it, and make it shorter. I said OK and became director of the project, and we created the IELTS test (International English Language Testing System), which you've probably heard of. It was reasonably successful. It was an interesting test, and we wrote some interesting research articles as a result of that. So that was one example. In Hungary, where I've been living for the last two years until last summer, one of my jobs was to advise on exit-exam reform. They wanted to help people develop decent tests and to train them in testing-related matters. So again, people asked me if I could help.

Washback Research

Vongpumivitch: *Your article with Dianne Wall (Alderson & Wall, 1993) argued that tests have an impact on classroom teaching and learning and explored the nature of such impacts, called washback, as well as ways to measure them. How did you become interested in studying washback?*

Alderson: The notion of washback came out of Sri Lanka. I was involved in Sri Lanka because we at Lancaster, in the Institute for English Language Education where I was director at the time, had teachers and teacher trainers coming to Lancaster to improve their understanding of language education. We were involved in advising partly on test reform and teacher-training reform, and through that connection they asked us if we would get involved and advise them on creating a new school-leaving exam. What we were particularly interested in doing was proving that washback works. So we got some money from the British government and designed a study that lasted about three years to prove that improving the tests as well as improving the textbooks and the teacher training would have the impact required. That was really the agenda. Unfortunately we were wrong. We showed that there was washback on content of teaching, but not on methodology of teaching, and that's where the article came from—the surprise of the lack of success of that project.

Carr: *Methodology would probably be the hardest thing to get a language teacher to change.*

Alderson: That's right, because it involves changing the way they think, basically.

Vongpumivitch: *Beliefs about teaching exist at the individual level, which may be the hardest place to bring about change.*

Alderson: Yeah, it probably is. I mean the stuff we've done since looking at TOEFL, for example, does show some impact on methodology, but it's not predictable because what we see is that teachers are very much influenced by their own teaching styles which relate to personality as well as their experience from the past. Very often people teach the way they were taught.

Vongpumivitch: *Most washback studies seem to be reports of case studies. What can we do with this information? At the end of the day, how do you make sense of the whole field of test impact research?*

Alderson: Read the Alderson and Wall paper (1993) where we talked about the Washback Hypotheses. Those hypotheses are the beginning of theorizing about impact, going beyond the primitive assumption that tests have negative impact to a more general statement that tests will have general influence and not necessarily a negative influence. So at that level, those Washback Hypotheses are a theoretical framework, and the case studies are ways of exploring that theoretical framework.

What we suggested in that article was that there are two theories from outside applied linguistics that we need to understand: innovation theory and motivation theory. More work has been done looking at innovation theory than has been done at the individual level of motivation. We were thinking of students' motivation when we wrote that article. But since then, I've come more and more to realize that we also need to understand teachers' thinking and cognition. In other words, what are teachers' belief frameworks? Why do they do what they do? What drives what they do? And so washback studies, I think, can potentially contribute to studies of teachers' thinking and theories of teacher cognition, as well as draw from research on teacher cognition in order to explore reasons why teachers do what they do. There are bits and pieces in a couple of papers I've written but I've not fully developed anything yet. A lot of people who have been working in teacher training and teacher education are doing that kind of research already. But one of the interesting things about a lot of what happened, particularly in Britain, is that it's not very empirical. It's argumentative; it's assertive; it's theoretically oriented. But there is not very much research that actually goes into the classroom, interviews the teachers, and asks why they did what they did. It tends to be the washback research where we do that sort of thing.

Vongpumivitch: *So there's definitely a need for empirical study.*

Alderson: Oh, yeah, absolutely.

Carr: *Do you think that it will be useful for an empirical study to come up with a general definition or sets of criteria for assessing the effects of washback?*

Alderson: What I've done in relatively recent work, before the Hungarian stuff, was to try to predict washback; that might be what you mean by criteria. So for example, for the University of Cambridge Local Examination Syndicate, several of my students and I designed a set of instruments to investigate the impact of IELTS. And the way we did that was by predicting what a positive influence of IELTS on classrooms would look like, in terms of the usual content—text types, task types, skills—but we also tried to predict what sort of methodology teachers would be using in the classroom. We can turn our results into a framework for doing washback research on IELTS and developing instruments for conducting washback studies, such as classroom observation schedules.

Carr: *So it would be more situation dependent. It wouldn't necessarily be one that can be generalized to all situations, just like there cannot be one test that can be used in every situation.*

Alderson: There are probably generalizable elements you could arrive at once you've done the particular. I think we now understand more from the case studies that have been published than we did before. Those case studies were embedded in individual contexts, so we can generalize from the mass of case studies, and I think methodologically we're beginning to understand what sorts of things we might want to predict regardless of the tests. In other words, we want to look at test constructs, test contents, test methods and predict from those.

Vongpumivitch: *Have you seen any example of a successful washback driven test reform?*

Alderson: The work we are doing in Hungary, I hope, will be one example. It hasn't happened yet.

Vongpumivitch: *You have briefly mentioned that you are currently involved in an education reform project in Hungary. Could you tell us more about it? Isn't it a teacher-training project?*

Alderson: There's a lot of teacher training going on in order to engineer positive washback. What we've developed is a test design which now has been piloted in three rounds. The test has not yet been introduced into the system and won't be

introduced for another four years. We have been working on a project in collaboration with the Hungarian authorities to develop tests which are communicative. The tests relate to the sort of teaching teachers say they want to do and are different from what they say they don't want to do. The difficulty is that the tests must also be for credit, which means that some students will fail, and the tests are therefore high-stakes tests and can have negative impact if teachers don't understand what is required by the test. So parallel with the technical things we are doing like standardizing, standard-setting, and piloting, colleagues have been developing an in-service teacher training course to help teachers understand what the test is about, why it's the way it is. They are trying to get teachers to think about how they would teach in class when they are given a test like this. For example, if you're going to teach reading in an exam preparation class, then what's the best way to teach reading in general, or listening in general? So it's relating best test preparation practice to best classroom practice.

Vongpumivitch: *It seems that if the Hungarian project is successful, it can be a model that other countries can use.*

Alderson: That's what we hope, certainly. We now have three publications; the third volume has just come out, giving details of the teacher-training course. Two books have been published already called *English Language Education in Hungary: Baseline Study* and *English Language Education in Hungary, Part II: Examining Learners' Achievement in English* (Alderson, Nagy, & Oveges, 2000). Part III is about the in-service course that we will develop, and is subtitled *Training Teachers for New Examinations.* It is edited by Együd Györgyi, Gàl Ildikó and Philip Glover and is available from Edit Nagy in the British Council, Budapest.

When doing a washback study, you have to be careful. There's always a danger that teachers will fool you. They'll do things that are unprofessional as a shortcut. Teachers vary enormously, as you know. Some think seriously about what's the best way to prepare their students in general as well as for tests, and some don't. Particularly in countries like Hungary or Sri Lanka where teachers are not well paid, they have to go and do other jobs. They take shortcuts, just as they do here in America when teaching TOEFL. So one of the tricks, I think, in engineering washback is to think how teachers can possibly subvert the test and then try to find ways around that subversion. A lot of people don't like such an approach. When people talk about teaching, they tend to talk about teachers as if they were essentially professional people who want the best for their students, but not all teachers are like that. So you have to think of those teachers and how they might teach—you have to consider the worst case as well as the best case.

Computer-based and Web-based Testing

Carr: *What do you see as some of the advantages, disadvantages, limitations, and doors that are opened by computer-based testing and web-based testing?*

Alderson: The biggest danger of computer-based testing in general is that the methods used in the tests tend to be conservative. After all these years, we still have multiple-choice, and we still have gap-filling or cloze tests. So it's a real worry that the test methods will be conservative compared to what we do in paper-and-pencil tests or face-to-face tests. But with the web-based tests and with the increasing power of IT in general, I think people are starting to experiment with method; for example, the TOEFL listening test is enhanced by having visuals. We don't actually know whether it improves the test, but my hunch is that it does. Similarly when on-stream video becomes widely available that could also have an advantage.

The biggest disadvantage of web-based testing for many situations, though, is not the method, but the lack of security. Having secure web-based tests is problematic. So the TOEFL, for example, will never become web-based, I don't think, at least not in my lifetime. Where the most interesting developments in web-based testing or computer-based testing are going to happen is in low-stakes environments rather than high-stakes environments such as placement testing, diagnostic testing, classroom testing. The problem is most money doesn't go into producing classroom tests, or into producing an achievement test or if it does, it tends to become a high-stakes test. You need to have resources to be able to develop interesting test methods.

Carr: *There are placement tests here [at UCLA] that are moderately high stakes, and we are trying to take them web-based.*

Alderson: It will be interesting to see what you're going to do. Internet-based? Well, security will be a problem. Personation will be a real problem. How do you guarantee that the person who's taking the test is the person who is applying to come to your classes?

Carr: *For now we're planning on lab-based administration.*

Alderson: Right. There you go. Which means that you're eliminating the advantages of web-based testing. It no longer individualizes but it's group-based.

Vongpumivitch: *Are you saying that very high-stakes tests should be paper-and-pencil as opposed to web-based?*

Alderson: I just think that the advantages become a bit more debatable. Obviously rapid feedback is very useful, with the web-based test. Having a record of scores and so on in terms of research purposes as well as administrative purposes is very handy. It cuts out the stage of having machine readable answer sheets. So there are advantages but there are not very many construct advantages.

Vongpumivitch: *Beyond improving test administration processes, how can a computer-based or web-based test provide better interpretation of the test scores? How can we have a better measure of language proficiency using computers or the Internet?*

Alderson: I think the incorporation of multimedia is an interesting possibility because you can do all sorts of simulations. Ultimately your test method is almost certainly going to be selective-response type method, which has problems, but you can see how scenarios could be developed for more integrated skills testing, in the receptive skills at least.

Carr: *So you don't see much of a role in the immediate future for automated scoring of open-ended, limited production tasks.*

Alderson: Well as you know, some work on that has been going on at ETS, but I think that's several years down the pike before artificial intelligence progresses far enough to deal with English as a second language. A lot of stuff that they are doing in developing algorithms is English as a first language, and they are having some degree of success.

Carr: *They are doing that for essay tests. But what about specifically for the short-answer type—one word, one sentence?*

Alderson: There are people at ETS, Jill Burnstein is one of them, who believe that they can develop artificial intelligence systems to do that, but I haven't seen that being done successfully.

Vongpumivitch: *But then you only need artificial intelligence in the case of ETS tests, because there are so many people taking their tests. But if your test taker pool is not that big, then the need for artificial intelligence may not be as strong.*

Alderson: Well ETS has the resources to develop artificial intelligence systems, if they would then make them available to the education community that would be great. The size of your test taking population does not matter so much, I don't think, provided you are dealing with the same language and essentially the same construct and test methods.

The other great hope, of course, is corpus-based testing, where you use your corpora to provide criteria for deciding whether something is acceptable or unacceptable. The sorts of corpora I'm thinking of are the sorts that have been developed particularly in Britain—the Bank of English and the British National Corpus, which are most interesting in their written forms, not in their spoken forms. Concordancing is a powerful tool which can be used on corpora. Imagine you have a structural pattern, or frame, in which you want to create an item. One could search in the corpus for that pattern or frame and concordance elements that occur within that frame. The problem is, no matter how big your corpus is, it may not contain the frame that you want to use in your items. So presumably what you then have to do is to take the language from the corpus in order to produce the frame, and then go back to the corpus to get the range of possible responses. With a parsed and tagged corpus, you get information about parts of speech, about the syntax, and you could search the corpus for examples of a particular structural framework.

Carr: *Could you elaborate on the notion of structural frames a little bit?*

Alderson: What's been developed in Lancaster alongside the British National Corpus is the *claws-tagging* which also gives you some information about grammatical function: not just the part of speech, but also the subject, object, verb, that sort of function, the fairly basic aspect of clause structure. That's why it's called *c-l-a-w-s*. You can search for particular syntactic features associated with particular parts of speech. For example you can ask for examples of a pronoun in a syntactic frame, and so on. So you could use that information to construct tests, if that's the sort of test you want to construct. Obviously it will be a test of fairly low level linguistic information. People are working on anaphora marking, and there are two people in Lancaster who have been working on semantic tagging, developing semantic frameworks for identifying meanings.

Vongpumivitch: *But those sound like grammar tests to me.*

Alderson: Of course that may be another reason why you might not be interested, that's right. But there are a lot of things you could do. Any decent corpus will have texts, classified by type, genre, so you could go to your corpus and say, "Give me a text of the following type, on the following topic, with the following structure."

Vongpumivitch: *And you'll get your reading passage right there.*

Alderson: That's right. You then have to go ahead and start constructing items that tests whatever skills you want to test. That's something the computer can't do for you.

Vongpumivitch: *But the answers to those questions, if they happen to be open-ended questions, cannot be checked by the corpus.*

Alderson: Correct, unless we've got a semantic-parsed corpus, and some people have been developing parsing systems that will, for example, suggest synonyms, paraphrases, and so on, of a particular word.

Carr: *At this point you're getting into artificial intelligence.*

Alderson: Yeah, it's getting close.

Carr: *What do you think should be some of the priorities in the development of and research on computer-based, web-based testing?*

Alderson: Whenever I talk about DIALANG, for example, or computer-based testing more generally, people who aren't very keen on information technology ask the question, "What is the added value? What do you get from doing this that you couldn't get using paper-and-pencil?" And obviously some of the answers are practical ones. I think if you can show that you're enhancing the construct, that you're enhancing in some sense the validity, reliability, or the usefulness of the test, then I think that's where research should go, preferably on the validity side. The reliability side is fairly obvious, I think, and usefulness is not always clear. I think the other area to look at is negative impact. What will be the impact of computer-based tests compared to the paper-and-pencil-based test? And nobody has done that kind of research for TOEFL, for example.

Carr: *Do you mean that nobody has looked at how computer use is going to disadvantage some students?*

Alderson: Well, people have looked at the disadvantages, of course. Computer literacy was examined in the initial study for TOEFL CBT. Studies showed some disadvantage in some parts of the population. What they didn't really show was how people with and without computer literacy perform on the new TOEFL, rather than on TOEFL-like tasks. What ETS did is rather different. What ETS did was develop technology familiarity questionnaires that are administered during the regular TOEFL administration. It's not the same thing as looking at the impact of computer literacy on TOEFL CBT. But interestingly, ETS has never done any research into TOEFL washback. There are no TOEFL washback studies, apart from the little thing that I did with Liz Hamp-Lyons (Alderson & Hamp-Lyons, 1996). There should be TOEFL washback studies. There are so many complaints out there about TOEFL. It's unprofessional of people not to have studied its impact.

Vongpumivitch: *But it will be a study of impact on what?*

Alderson: On the students, on how they prepare. On the teachers, on how they teach. The study we did (Alderson & Hamp-Lyons, 1996) was in the United States. We looked at two teachers, teaching their regular classrooms and their TOEFL classes. We looked at impact on the teachers and how each teacher changed what he or she did in the classroom.

Vongpumivitch: *But the majority of people preparing for TOEFL go to test prep schools.*

Alderson: OK, then let's go to the test prep schools and see what they do. They guarantee to increase your scores by 60 points. How do they do that? What is it they're teaching? Why are they teaching that? Why don't they teach something else? It will also be useful. Nobody has looked into it. Why not?

Test Validation and the Quantitative/Qualitative Paradigms

Vongpumivitch: *In your opinion, what's the primary goal of a test validation study? What should be the characteristics of good language test validation research?*

Alderson: What you want a validation study to do is to show that the test you are studying, that the inferences that you've drawn from the test scores, and the uses for the test scores are justified. That's tricky, because what you would like to be able to do is to show that you can make better inferences on one sort of test as opposed to another sort of test. In other words, very often test validation is most useful when it is comparative. Most test validation isn't. Most test validation is also problematic because it tends to be with truncated samples. It tends to look only at those students who had succeeded in some way. Look at predictive validation studies, for example, how do you do a predictive validation study with TOEFL without letting in the students with low scores? The only people you look at are those who got in. So ideally a validation study will be used before the test becomes operational; it will look at the full range of possible consequences of the scores before we use the scores from that test. Most validation studies are much more limited because of the truncated samples and because of the nature of the criteria against which they are comparing the test scores. Indeed they are often limited because of the rather quantitative nature of validation studies. Ideally a validation study is both quantitative and qualitative. For example, I have a student, Jay Banerjee, who's looking at IELTS and its predictive validity. What she's been doing is looking in great detail at how admission officers actually use IELTS scores, how they make decisions by taking into account the IETLS scores, as well as all the other information. Then she's interviewing, again in great depth, the individual students whose scores have been used, looking at the problems they have in their study setting and trying to understand the complexity of their language use prob-

lems. Ideally, a test validation would look in considerable depth at those issues, for students who have got low scores as well as those who have got high scores.

Carr: *Some people criticize language testing research as overemphasizing quantitative or psychometric methodology, at the expense of qualitative methods. But on the other hand, others criticize qualitative methodologists for focusing too much on case studies which are not very generalizable. As students we were told in our research methods classes that the ideal approach involves a complementary use of both paradigms in a study. How realistic a goal do you see this in language testing research?*

Alderson: As always it depends upon the purpose of the validation research or whatever it is you are doing, and it depends upon the resources that you have. To begin with, I don't necessarily see that there is an essential difference between quantitative and qualitative research. It is clear that qualitative researchers have to quantify, because they use words like *some*, or *many*, or *exception*, and so on; that's quantification. And, similarly, quantitative researchers are concerned about the quality of their instruments; that's what validation is. So this is a false dichotomy. I think generally it's accepted now in the social sciences that the most sensible approach is a fusion of both. If we stay with the words for the moment to understand the differences, qualitative research is very resource hungry, not only in terms of gathering the data—you have to interview people, observe them, or whatever you're doing—but also in terms of analysis. Analysis is extremely time consuming with the sort of data you have, even using software. There are good packages out there, but they don't get the grasp of what you have to do in coding. That's why qualitative research is either not very well done, or is only done superficially, or as an afterthought to quantitative analysis, because quantitative results are more amenable to analysis. Of course, quantitative data is relatively easy to gather. People have to take a test anyway You just give them questionnaires as well, and you can get the data quickly. It's hard to get data for qualitative research. Ideally you need both. And a lot of the stuff that I've done has been in fairly small numbers and more toward the qualitative end than the quantitative end, typically because Britain is a smaller place; we have smaller populations to deal with.

Carr: *What do you see is the best way to go about integrating the two paradigms?*

Alderson: I think the value of qualitative research is that it helps you understand the problem, or identify the dimensions of a problem. So you could do the qualitative stuff first as a way of piloting to get inside the complexity of the situation, and then possibly follow up with analysis to identify key variables which might be worth exploring in greater extent or in more depth, and follow that up with quantitative studies. But I think both qualitative and quantitative researchers are worried about generalizability. What is generalizability? Very often, to generalize means

to ignore important variables. We know how important context is, and we know how situation-dependent most of our work is. So I think you have to state the limit, as you understand it, to the generalization you're making.

Carr: *Are there any good examples you can think of in language testing that have succeeded in using both approaches?*

Alderson: I guess the classic one is the paper by Anderson et al. (Anderson, Bachman, Perkins & Cohen, 1991) —the triangulation study. That's a nice paper. It's a nice example of how you can go about examining the same problem from different perspectives. I guess it's true to say that the TOEFL-Cambridge comparability study (Bachman, Davidson, Ryan & Choi, 1995) is not a bad example of an attempt to use content analysis in a quantified way in order to shed light on the quality of the instruments involved. There were still more data they could have gathered, introspective data for example, but I think that it was a good study that was well resourced and took a long time, longer than a graduate student can possibly have done alone.

Vongpumivitch: And it has to be a team effort.

Alderson: Absolutely. See the big problem with a lot of testing research is that it's done by graduate students. You need a team of researchers, and that isn't recognized for the award of a PhD. So we should be having more funded research, teaming people with different skills, taking place over time. Most research that we do is one shot research rather than developmental, and we should be doing more developmental research. But that takes time. We should learn from the sciences. What we should learn is to replicate. Give people PhDs if they replicate adequately because a PhD in our field is an apprenticeship to do research; it qualifies you to be a researcher. So limit what you demand of somebody at the PhD level. Don't expect original, creative research. Do that later. Once you've got your PhD, team people up. That's the way it should be done.

Second Language Reading Research

Vongpumivitch: *You've been in the field of reading for about 30 years now. Reading is a field that is heavily researched. What is it that we know enough about now, and what is it that we still don't know?*

Carr: *What are we sure about?*

Alderson: We're pretty sure it's complex. We're pretty sure people can do it but we don't know how. I often wish I never got into it because I'm not sure what I've learned as a result of 30 years of research.

Vongpumivitch: *Seriously? Even after writing a book about it?*

Alderson: Well, think about the stuff I've done on skills, and again I started on that because of a particular real-world problem that somebody came up with. I was in India in 1986 and somebody was saying, "What we're doing is using Benjamin Bloom's taxonomy to test our students' ability in English," and I said, "You can't do that. Native speakers are different, and so why are you doing that with nonnative speakers?" So the research I set out to do, looking at levels of skills, was to prove them wrong in India. And I proved them wrong, but I'm not sure if I've proved anything beyond that. Everything I've done since then in looking at skills has convinced me of the complexity of the issue, but hasn't reached the solution. As I said in the book (Alderson, 2000a) basically it seems to me that what happens is individuals use strategies and skills in individual and idiosyncratic ways, depending upon purpose and knowledge, etc., and it's very hard to generalize from what individuals do to some deeper understanding of what is involved. Of course we have schema theory but it has got us nowhere. We know background knowledge has an effect and that's hardly a surprise.

Carr: *How compensatory are these varying ways?*

Alderson: I hate to say it, because I was brought up as an applied linguist in the 1970s when we believed that skills and strategies are important, but I think in ESL or French as a foreign language, the language is what you need before you can have adequate transfer. So the threshold hypothesis is still very important. The reason why I say "I hate to say this," is that I fear for the washback of statements like "You need to know the language first," because what people start doing when they teach grammar or vocabulary may not be the best way to learn a language. But I'm certain you need a good linguistic foundation because then you can start all these other things. Now that doesn't help very much. Frankly if you had sat me down 30 years ago I probably would have said the same thing. So what have we learned after 30 years? I don't know.

Vongpumivitch: *Is there any hope for reading researchers at this point? Are there any areas that need to be investigated?*

Alderson: Every area, you name it, can be investigated. Absolutely. I've got a student at this moment investigating reading aloud. I thought that had died out 30 years ago, but she still finds some interesting problems about it. It's all up for grabs, and I'm not sure I want to supervise it anymore. It's very frustrating. *Assessing Reading* is a long book. It caused me a lot of sweat to write that book. I had to read an enormous amount, and I don't think there's a clear message coming out of that book.

Vongpumivitch: *At the end of the book you still had to conclude that this is not a conclusion.*

Alderson: That's right.

Carr: *Is it fair to say that you think the types of research questions that have been addressed over the years haven't really evolved so much, that they have come full-circle? Are we still looking at the same issues as 30 or 40 years ago?*

Alderson: I suspect we're not, but we should be. Obviously we have a better understanding of terms like *strategies*, and even *skills*. We have a better idea of how we can do research, through introspective research, for example. But one of the problems with the field is, what we don't do is, build on each other's research. So what we're not doing is saying, "OK, let's develop a program of research that explores these different angles, and accumulate knowledge." What happens is that people are doing their own things. This goes back to what I was saying about graduate students. Doing research on your own is great, but it's fragmented.

Closing Thoughts

Vongpumivitch: *There's a perception on the part of many people outside the language testing community, that the field of language assessment is strictly quantitative and generally incomprehensible and inaccessible to non-specialists. What is your response to the people who say they want to come into testing, but feel that it is too inaccessible to them?*

Alderson: Applied linguists should know that they can learn enough of the quantification in order to throw light on the problem that they are trying to address. You don't need to do all these fancy things in order to address particular problems. Statistics are there to be used as a tool to help you understand something, and not vice versa.

You could do discourse analysis; look at the studies into OPIs and SOPIs and all the other language testing studies that utilized discourse analysis. Some of the work has been done; Steve Ross and others, for example, have looked at interlocutors' accommodations (Ross & Berwick, 1992). Ross' study is a decent discourse analysis that throws light on how we might improve the training of interlocutors.

Vongpumivitch: *How can a graduate student or a novice researcher start a language testing study?*

Alderson: Start with a problem, and see how testing can help. This was why I got involved in testing. I was told to develop a placement test in my very first teaching

job. What was the problem? To identify weaker students from stronger students and put them into homogeneous groups. So I had to get involved in testing in order to address that problem. That's where interesting testing work comes from—having a real-world problem, then trying to do research.

Carr: *What are some introductory level books that you would recommend for someone who is interested in language testing?*

Alderson: Tim McNamara has a nice new book in the OUP series (McNamara, 2001) It's a nice introduction. Arthur Hughes' book on testing for teachers (Hughs, 1988) is pretty superficial, but it's a good start. Brian Heaton's second edition, *Writing English Language Tests* (Heaton 1988), is a nice easy introduction. Going beyond those, you can get Cyril Weir's book on standardizing and developing tests (Weir, 1993); that's a good book. The book I did with two colleagues (Alderson, Clapham & Wall, 1995) is a bit more technical but people find it readable. If people really got hooked by then, then obviously they should buy all the books in the series by Lyle Bachman and myself, which is learning about constructs. The best way of finding out what vocabulary is in Applied Linguistics is to read John Read's book about vocabulary (Reid, 2000), because he tells you about the construct as well as testing. That's what those books are intended to do, to show the centrality of testing to applied linguistics. It's not peripheral; it's central.

SUMMARY AND CONCLUSION

In this interview, Dr. Alderson shares his opinions and gives suggestions on four key areas of research in language testing: test impact (washback), computer-based/web-based language testing, test validation research, and testing reading comprehension. He emphasizes the need for empirical studies of test impact, especially the investigation of teachers' motivation and the prediction of washback. In terms of computer-based and web-based language testing, Dr. Alderson argues that it is crucial to show that technology enhances the ways language ability can be measured and helps create tests that have better quality and yield better score interpretation. He also discusses the disadvantages of computer-based and web-based language tests, such as the lack of test security and the difficulty of grading open-ended items using complicated artificial intelligence.

Dr. Alderson advocates more cooperations among researchers from different disciplines, arguing that more team research will enrich the quality of language testing studies. He believes that the best test validation research is that which incorporates both quantitative and qualitative research methods, investigating all related issues in great depth. Cooperation is also needed in the area of reading assessment; Dr. Alderson points out that there are still many unanswered questions in all aspects of reading assessment.

Perhaps the most important point in this interview is Dr. Alderson's emphasis that language testing is a central field of study in applied linguistics. Dr. Alderson explains that the perception of language testing as a strictly quantitative area of study is a misunderstanding. Language tests are measurements of language abilities, and language testers need to first have a clear understanding of the nature of language abilities before they can measure them well. Dr. Alderson argues against the notion of a dichotomy between qualitative and quantitative methods since few studies are strictly one or the other. Ultimately the research questions motivated by real-world issues are the language testers' guides, and various kinds of quantitative and qualitative research methods are merely tools for language testers to use to conduct an applied linguistic study.

NOTES

[1] The interview took place at UCLA while Dr. Alderson was at Los Angeles as one of the invited keynote speakers for the Fourth Annual Conference of the Southern California Association for Language Assessment Research (SCALAR 4) at the California State University, Los Angeles, CA, May 11-12, 2001.

[2] Currently there are four books in the Cambridge Language Assessment Series: J.C. Alderson (2000a) Assessing Reading, G. Buck (2001) Assessing Listening, D. Douglas (1999) Assessing Languages for Specific Purposes, and J. Read (2000) Assessing Vocabulary.

[3] DIALANG is a project funded by the European Union for the development of diagnostic language tests in 14 European languages. Tests will be made available on the Internet free of charge. The DIALANG project will offer separate tests for reading, writing, listening, vocabulary, and grammatical structures, covering all levels from beginning to advance. For more information on the DIALANG project, please go to: http://www.dialang.org.

[4] Dr. Alderson taught EFL and applied linguistics in Germany, Algeria, Scotland, and Mexico. He became interested in language testing and applied linguistics when he was an EFL teacher in Germany, which led him to his postgraduate study at the University of Edinburgh. His early works, such as works on reading, cloze tests, and metalinguistic knowledge stemmed from teaching-related problems he encountered as an EFL teacher.

[5] The International English Language Testing System (IELTS) is an academic English as a Second/ Foreign Language test required by Australian, British, Canadian, and New Zealand post-secondary institutions. All non-native English-speaking students wishing to enroll in such institutions have to take this test, which consists of listening, speaking, reading, and writing, from the University of Cambridge Local Examinations Syndicate, UK. Nowadays, some American post-secondary institutions accept the IELTS scores in place of the Test of English as a Foreign Language (TOEFL) scores.

[6] English Language Education in Hungary, Part II: Examining Learners' Achievement in English is a collection of progress reports on the Hungarian English Examination Reform Project co-edited by Alderson.

[7] In language testing, a construct is a definition of the ability that is to be measured by the testing instruments. Such a definition of language ability needs to be appropriate to the particular testing situation, test purposes, test taker population, and types of actual language use in the real world (Bachman & Palmer, 1996, p. 66).

[8] According to Biber, Conrad, and Reppen (1998) and Kennedy (1998), to conduct a corpus linguistic analysis, a corpus user can use software to display all occurrences of a search item, such as a keyword or a syntactic morpheme, in a corpus. A concordance is a formatted display of an exhaustive list of all of the occurrences.

[9] This study on reading comprehension tests investigated the relationship among test taking strategies, content of test items, and the students' test performance by using think-aloud protocols, content analysis of each test item, and statistical analysis of the think-aloud protocol data and the test performance data.

[10] In this study, Bachman et al. investigated the comparability of the TOEFL test and the First Certificate in English (FCE) test, which was created by the University of Cambridge Local Examination Syndicate, by gathering data from ESL students in eight countries around the world and conducting statistical analysis of the test performance data along with expert judges' ratings of the two tests' content.

[11] The Oral Proficiency Interview (OPI) is "a structured, live conversation between a trained interlocutor/rater and a test-taker on a series of topics of varied language difficulty" (Chalhoub-Deville, 2001). The scoring of the interview is based on the ACTFL Proficiency Guidelines. The Simulated Oral Proficiency Interview (SOPI) was developed by the Center for Applied Linguistics and is a tape-mediated speaking test in which the tape prompts several topics of varied language difficulty. The test takers record their responses onto the cassette tapes and their speeches are rated using the ACTFL Proficiency Guidelines.

[12] The Cambridge Language Assessment Series. See Note # 2.

REFERENCES

Alderson, J. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language testing construction and evaluation.* Cambridge: Cambridge University Press.

Alderson, J. C & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13* (3), 280-297.

Alderson, J. C., Nagy, E., & Överges, E. (2000). *English language education in Hungary, Part II: Examining Hungarian learners' achievements in English.* Budapest: The British Council Hungary.

Alderson, J. C. & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14* (2), 115-129.

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*, 41-66.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. C. (1995). *An investigation into the comparability of two tests of English as a Foreign Language.* Cambridge: Cambridge University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Buck, G. (2001). *Assessing listening.* Cambridge: Cambridge University Press.

Chalhoub-Deville, M. (2001). Task-based assessments: Characteristics and validity evidence. In Bygate, M., Skehan, P., & Swain, M. (Eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (Chapter 10, pp. 210-228). Harlow, England: Longman.

Douglas, D. (1999). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Heaton, J. B. (1988). *Writing English language tests.* London: Longman.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.
McNamara, T. (2001). *Language testing*. Oxford: Oxford University Press.
Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.
Ross, S. & Berwick, R. F. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition, 14* (2), 159-176.
Weir, C. J. (1993). *Understanding and developing language tests*. New York: Prentice Hall.